What Works for Children's Social Care

Last updated: February 2021
Version: 1.1

What Works for Children's Social Care
Randomised Controlled Trial Statistical Analysis Guidance

This document outlines the approach to statistical analysis of randomised controlled trials recommended by What Works for Children's Social Care (WWCSC). This guidance is based on the existing methodological literature in related fields, particularly education, which deals with similar cohorts. This guidance has been established by the WWCSC's Research Team, in collaboration with academic advisors. As there are relatively few randomised trials in children's social care at present, we anticipate this guidance changing over time to reflect new learnings and the experiences of researchers and evaluators.

The purpose of the guidance is to allow evaluators to quickly make methodological decisions relating to trials, to ensure that these are comparable across studies, and to ensure that the research outputs from the WWCSC and other researchers are as useful to social workers as possible.

Introduction

Randomised controlled trials, as well as other forms of impact evaluation, should be as comparable as possible. Although different interventions carried out on different cohorts in different locations at different points in time may be challenging to compare, every effort should be taken to maximise the extent to which this comparison is possible. Without comparability between studies, even at a high level, research is not a useful tool for practitioners attempting to make decisions in real time. If nothing else, a sense of value for money is essential for supporting local governments and other agencies who face tight resource constraints.

To facilitate the production of high quality evidence, the WWCSC requires that research protocols are published before trials or other substantive research projects commence, and that these contain detailed analytical plans. These protocols will be published on the WWCSC's website, as well as on the Open Science Framework website (OSF).
Pre-registration of analysis plans prevents researchers or others from selecting analytical strategies, having seen the data, which favour a particular outcome from the research, and hence increase reproducibility, and allow practitioners to have greater faith in the robustness of the research.

Pre-analysis plans should not, however, be viewed as a straitjacket. Additional exploratory analysis can legitimately be conducted that deviates from the analysis plan laid out in the protocol, however, main analysis must be as specified in the protocol; exploratory analysis must be labelled as such, and only presented at the same time as primary analyses. Further exceptions to strict pre-registration of analysis before the trial begins can be found in the case of evaluating complex interventions.

1) Analysis must reflect the design

Analytical strategies chosen for evaluating trials should reflect the design of the trial - whether is individual and parallel randomisation, a waitlist, or clustered trial. Analytical strategies should in general make use of the simplest analytical strategy available for the task, rather than the most complicated or sophisticated, and should in general prefer methodologies which are straightforward to interpret.

The design of an analytical strategy should be based on both the design and the power calculations conducted before the trial begins, and the latter two should also depend on each-other. As a general rule, analytical strategies should be chosen such that they maximise the amount of statistical power available for a given design of trial. Appropriate control variables, based on data availability and either theoretical or empirical grounds, should be used where possible to achieve this goal as long as it does not compromise the design of the trial.

Strategies for randomisation should also reflect the structure of the data and of the trial, with randomisation being preferred at the lowest level possible for a given intervention, and stratification on important, particularly uncommon, variables conducted where appropriate.

2) Use intention to treat analysis

Primary analyses should be undertaken using ‘intention to treat’ analysis, so that everyone who was randomised is included in the analysis as though they had received whichever intervention they were assigned to, regardless of what they ended up receiving. This approach gives the truest account of the effect of the intervention when delivered in real world conditions, without the need for more onerous assumptions.

Where possible, sub-group analyses should be carried out on the full intention to treat sample, with the inclusion of interaction variables for the sub-groups of interest.

Additional analyses, which are described later in this document, can consider treatment effects on the treated, or handle missing outcome data for participants which prevents a true intention to treat from being conducted.

3) Control for prior levels of the outcome measure where appropriate and possible

The best predictor of future behaviour is, in most cases, prior behaviour. As such, the greatest advantage in statistical power is likely to be obtained from the use of prior behaviour as a covariate in analysis.

In education, or in health, this is often a straightforward thing to measure - for example, in education, we might control for grades at an earlier age, or where this is not available, we might conduct a baseline test on children at the beginning of the trial (prior to randomisation). In health, we might be able to control for, for example, prior levels of HbA1C (a measure of blood sugar), when evaluating impacts on current levels either of that variable, or on diabetes risk. In Children's Social Care, lagged values of the outcomes may be harder to come by. For example - if our outcome measure is whether a child is taken into care, children in the trial should probably not be in care at the outset of the trial, and so the baseline level for each child will be 0 - and so of little analytical value.

Where this is the case, we should consider controlling for more general, higher level, or aggregate histories of the outcome measure:

General:
For example, although a child may not have been taken into care previously, different young people in the sample may have different prior involvement in children's social care, for example having been on a child protection plan, been classed as a child in need, or subject to a more temporary care order (police protection of section 20). These prior experiences, although not strictly a lagged value of the outcome measure, are likely to hold considerable explanatory power.

Higher level:
For example, if young people themselves have not got experience of the care system, it is possible that their family does, and information on the previous care system involvement of the family and other children within it is likely to hold some explanatory value. Where this is not possible at a family level, the level of a social worker, their team, or even in large trials their local authority could be considered as explanatory variables.

Aggregate:
Where outcome measures are not available for an individual, but are available historically for an institution of which they are a part, aggregate historical levels of the outcome measure can be used. This could include, for example, school or class levels of the outcome measure for a previous year.

Baseline data, wherever possible, should be included in its most intuitive form, which will typically be its raw form, without transformation. As described above, other covariates, where there is a strong theoretical or empirical grounding, should be included in the analysis is well, and pre-specified. Proportion of variance explained by these covariates should be factored into power calculations.

4) Take account of clustering

When trials are randomised at the level of a cluster, be that a team, local authority, social worker, or other level higher than the individual case, it is necessary to reflect this in the design of the analytical strategy.

For consistency, the preferred analytical solution of the WWCSC is to cluster standard errors at the level at which randomisation occurs. If errors are not clustered in this way, the estimates of the effect of an intervention will be unbiased, but our confidence in them will be spuriously high. Alternatives, and in particular the use of multi-level models, are not preferred here, as per the previous description of parsimonious analytical strategies being preferred.

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

Effect sizes should be reported first in absolute terms - an X% and Y% point increase in a particular outcome for example, with confidence intervals around those estimates. To standardise our understanding of impacts across multiple studies, effect sizes should also be expressed as a proportion of the standard deviation in the *control group*, also known as Glass's Delta. This measure allows for comparison between studies more readily than, for example cohen's D, as it is not influenced by any impacts that the intervention has on the variability of outcomes as well as on the mean.

For trials of different types - for example clustered trials, or when additional covariates are used, we continue to recommend using Glass's Delta, calculated using *unconditional* (that is, unadjusted) standard deviations, as this allows for greater comparability of effect sizes, combined with greater flexibility of initial analytical strategy compared to the alternatives.

Where outcomes are binary (a child being taken into care, for example), risk ratios and percentage point changes should be used to express effect sizes, as these are the most intuitive ways to present results. Odds ratios should not be used as they are confusing and prone to misinterpretation. For comparability, where possible these effect sizes should also be presented in terms of what they represent in terms of moving along a national distribution from the mean (i.e. a 5% fall in CP cases escalating to children being taken into care is the equivalent of moving from average local authority to a local authority in the 40th percentile).

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 Intracluster Correlation Coefficients

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 the level of randomisation, but can report more if appropriate (e.g., form, when only clustering at school level was assumed).

8) Multi-Site Trials

A large number of trials conducted in children's social care will, by necessity, be multi-site trials. Many of these will involve site-level randomisation, and so the level of heterogeneity between (as well as within) sites should be considered carefully when developing a randomisation strategy.

For 'true' multi-site trials, multiple sites are involved in the trial and randomisation occurs within each site. Although there exists guidance for the design of these studies from the perspective of statistical power, we do not consider this as part of this guidance as many of the trade-offs faced in the design of multi-site trials (for example the decision of how to sample of individuals within studies), cannot be made by evaluators in children's social care.

Analytically, evaluators should control for participants' baseline characteristics as normal, and for site-level variations using fixed effects. Evaluators should estimate site level treatment effects, but should report the variances in these only, as well as, where pre-specified through contextual features, the correlates of site level treatment effects.

9) Handling Trade-offs

Evaluators will often face trade-offs between priorities in the design of a study. For example, maintaining the comparability of analysis across studies, or the simplicity of a single study may come at the loss of statistical power. Similarly, the selection of primary outcome measures might often involve choosing between the most 'important' outcome measure, and one which has a narrower variance or is thought to be more malleable, effectively forcing a trade-off between statistical power and the usefulness of the design.

In these circumstances, evaluators should specify the trade-offs they face explicitly, which they are choosing to prioritise, and why. In the case of trade-offs in statistical power, these should be explicitly modelled prior to protocolisation so that an assessment can be made as to the value of these trade-offs.

In general, but not always, the extent to which we are able to learn from a study should be maximised, meaning that statistical power and 'usefulness' should be prioritised, while comparability and simplicity may be less important for any given trial.

PRIMARY OUTCOME ANALYSIS

Different forms of outcomes

Analyses specified in trial protocols should be divided into three categories - primary, secondary, and robustness checks. Analysis conducted that are not specified in the trial protocol are to be classed as exploratory. Any reporting of these results, including in executive summaries, presentations and press releases, should stipulate which type of analysis a finding relates to.

Primary Analysis:

Primary analysis should be limited to a small number of outcome measures, and usually analysis conducted on your entire sample + perhaps including interactions with one key sub-group of interest. The ideal number of primary analyses is 1 - the primary outcome measure evaluated for the whole sample. Under normal circumstances, primary analyses should be limited to four - two primary outcome measures, evaluated for two subgroups. Subgroup analysis should be conducted using interaction effects (see table below).

	Whole Sample	Young people with prior social care involvement
Taken into care	Basic Analysis	Interaction
Educational Attainment	Basic Analysis	Interaction

Secondary Analysis

Secondary analysis should be used to answer research questions that are important to the understanding of the trial and its effects on young people's lives, but which are not primary concerns. This could include outcomes that are relevant but not the direct focus of the intervention (for example, criminal justice outcomes in a trial focusing on education), or subgroups for which the outcomes may be particularly interesting but for which there is no practical or theoretical reason to expect a differential effect.

Compliance

Even where Intention To Treat is our primary means of analysis, we may still be interested in the effects of an intervention on those who complied with the treatment. For many interventions, which are only provided with the consent of a family, it is possible that an intervention will only actually be taken up by some, perhaps a minority of, participants. Where this is the case, if the benefits for those who engage fully with the intervention are substantial, and/or if the cost and effort of delivering the intervention is also concentrated on those who actually ultimately receive it (as might be the case, for example, with Family Group Conferences), an intervention might be recommended on the strength of a strong effect on compliers alone.

In order to estimate the impact of an intervention on compliers, evaluators should in general pre-specify that they will conduct Complier Average Causal Effect Analysis, as it is likely that for most interventions in children's social care there are few 'always takers' (who will receive an intervention regardless of their assignment), and relatively more never takers (who will *not* receive the intervention, regardless of their assignment). Following Ye at al (2014), CACE offers the lowest level of bias in the presence of substantial non-compliance where this is the case.

Robustness

Robustness checks should be used to confirm primary analyses, but not considered to be primary results themselves. They should primarily be used to identify or to address limitations in your analytical strategy. For example, if your definition of 'closeness to care' in

one trial is subjective and requires selecting an arbitrary cutoff for how deprived a family is, then robustness checks might try to identify how sensitive the results of the trial are to the specific cutoff chosen. In the case of more elaborate trials, such as a stepped wedge, robustness checks might aim to evaluate how sensitive the results are to particular analytical decisions taken. For example, if time in a stepped wedge trial (which is a confound in this case, correlated with assignment) is controlled for linearly, robustness checks might consider more flexible functional forms.
Where outcomes are similarly arbitrary - for example placement stability, the results should be checked for their sensitivity to the definition of the outcome chosen

Exploratory

Any analysis not pre-specified, and not covered by the above categories of analysis, should be considered exploratory. Exploratory analysis can be used to uncover nuance and patterns within results, and to suggest hypotheses for future research, or to allow for the making of more tailored recommendations, but should be viewed as a weaker source of evidence than other forms.

Analysis of Harms

Research conducted in children's social care may (and arguably should) have positive outcomes as its primary focus - for example reducing the likelihood that a child enters care, and hence maximising the chance they can continue to live with their family.

However, the risk of harm arising, either as a consequence of an intervention directly, or as a 'side-effect' - for example exposing children to more risk by having them remain at home. Evaluators should consider and specify the likely risks emerging for a young person (and other participants) in their research, and say how these will be measured and evaluated over the course of the research.

Service users, as well as social workers, should be consulted as part of the logic model development of research, and give their impressions both of the likely risks of harm and an appropriate level of risk tolerance given the nature of the intervention to be evaluated. Analysis of harms should be presented alongside analysis of main effects in research reports.

Evaluators, when considering harms to participants, may wish to draw on Lorenc and Oliver's (2014) taxonomy of harms, which considers;

- Direct Harms
- Psychological Harms
- Equity Harms
- Group and Social Harms

- Opportunity Cost Harms

Focuses for outcome measures

Outcome measures for research conducted by or funded by What Works for Children's Social Care should in general focus on one of four areas;

- Outcomes and decisions in the children's social care system;
- Outcomes in the education system (school attendance, grades, exclusion and suspension, progression, application to HE).
- Outcomes in the health system (illness, non accidental injury, A&E attendance)
- Outcomes on the criminal justice system (caution, arrest, charge, prosecution, outcomes from prosecution, recidivism).

Sub-Group Analyses

Correction for multiple comparisons

Problems of multiple comparisons occur due to the nature of tests of statistical significance. By accepting a 5% probability of a false positive (type one error), it is the case that the more tests that we conduct, the more likely a false positive becomes. We can think about this in terms of rolling a die - the probability of getting a six on one throw is ⅙. The Probability of getting at least one six on two consecutive rolls of a fair die is (⅙ x ⅚) + + (⅚ x ⅙) +(⅙ x ⅙) = 11/36 - which is substantially higher than the probability of rolling a 6 on one dice (⅙ or 6/36). Similarly with two independent tests each with a 5% false positive rate, the probability that both results a false positive is low (1/400), but the probability that either of them is a false positive is higher than 5% (9.75%). Hence, by running large numbers of tests, we increase the probability that we find a spuriously significant result.

There are a number of different options for correcting for multiple comparisons, but the premise of these is largely the same. When conducting multiple comparison tests, the burden of proof is raised based on the number of tests that are being conducted, such that your burden of proof across the tests remains broadly constant. The simplest example of this is a Bonferroni correction, which mechanically decrease the type 1 error tolerance for each additional test. For example, if the analysis conducted contains 2 tests, the type 1 error tolerance would be 0.05/2 = 0.025. If five tests were conducted, the tolerance falls to 0.05/5 = 0.01. The effect of this on sample size requirements for studies with more tests is not so straightforward, but can be calculated easily.

The Bonferroni correction, described above, applies a high premium to multiple comparisons. WWCSC instead makes use of Hochberg's step-up procedure, which preserves more statistical power while still ensuring an acceptable false-discovery rate.

How to use the Hochberg step-up procedure:

Suppose you are running *k* hypothesis tests (I will take *k*=5 in this example). Rank the *p*-values from smallest to largest and compare them with a sequence increasing uniformly from 0.05/*k* to 0.05 for the 5% significance level, from 0.01/k to 0.01 for the 1% significance level and 0.1/k to 0.1 for the 10% significance level, respectively.

Example: you run 5 hypothesis tests, which produce *p*-values of

H1: 0.04
H2: 0.06
H3: 0.2
H4: 0.015
H5: 0.005

We rank these in increasing order:

H5:0.005
H4:0.015
H1:0.04
H2:0.06
H3:0.2

And compare these, if we are looking at a 5% significance level, to a sequence uniformly increasing between 0.05/5 and 0.5:

H5 must be < 0.01 to be accepted (which it is)
H4 must be < 0.02 to be accepted (which it is)
H1 must be < 0.03 to be accepted (which it isn't)

When to correct for multiple comparisons

We advise correcting for multiple comparisons within a category of outcome (that is, primary or secondary outcomes), if there is an especially high number of comparisons within that category. The number of comparisons depends on both the number of arms, and the number of outcomes. Where:
Comparisons = (arms - 1)*number of outcomes

The below box indicates in what instances we recommend using multiple comparison adjustments.

Shaded boxes indicate when to use multiple comparison adjustments					
		Number of outcomes within category			
		1	2	3	4+
Number of arms	2				
	3				
	4				
	5				
	6+				

Thus, if a trial had 2 arms, and 3 primary outcomes and 5 secondary outcomes, you would not need to conduct multiple comparisons on the primary analysis, but you would need to conduct it on the secondary analysis.

Attrition

In any long term study, attrition by participants is likely to occur. There are two main forms of attrition with which we might be concerned - attrition from analysis (participant does not provide sufficient information for them to be included in the final analysis for the study), and attrition from intervention (participant starts taking part in the intervention and then stops). Participants in the treatment group can, of course, do both.

Lack of completion of a course of treatment will be considered here under “Compliance”, below.

Avoiding Attrition

Attrition from data collection effectively creates a missing data problem, which can be handled statistically, following the “Missing Data” approaches detailed below on our data. This section will therefore describe an approach to minimising and reporting attrition.

Attrition should be minimised wherever possible through the use of administrative data that allows young people and their families to be tracked through time even if they move between local authorities. If this national level administrative data is not available, then local level

administrative data should be preferred to bespoke data collection which is only collected for your research project.

Where administrative data is not available, every step should be taken to avoid attrition from data collection, and, to the extent that attrition cannot be avoided, to ensure that attrition is balanced across treatment and control groups. In practice, this means;

Divorcing Data Collection and Intervention: the process of data collection should be separated from anything that has to do with treatment. If data collection is tied to treatment, than attrition is unlikely to be even across treatment and control groups, introducing bias, and in particular bias towards those with a high degree of compliance with the treatment.

Minimising the burden of data collection: Where survey instruments need to be collected from individuals, minimising the level of complexity of data collections, through the use of age and education appropriate scales, piloted with the cohort of interest, and through judicious use of scales. Implementation plans for data collection should factor in the need for repeated visits, and be timed to be convenient for those whose data you are collecting.

Reporting attrition

Attrition should be reported following the conventions laid out by Dumville et al (2006). This means that the full sample baseline characteristics, the baseline characteristics of those lost to follow up, and the baseline of those analysed should be reported side by side. Note that this deviates from CONSORT guidelines, which do not require the reporting of baseline characteristics.

Missing Data

When not using administrative data, some amount of missing data is almost inevitable. If data are missing, we need to consider both why the data are missing, and the extent to which they are missing.

If data are substantively missing (more than 5% of outcome data), it is likely that missingness is impacting on the results of the trial. There are a number of different ways that data can be missing in a randomised controlled trial, described in the table below, along with implications.

Type of missingness	Description	Implication
Missing completely at random	Data are missing in a way that is not correlated with either observable or unobservable variables	Complete case analysis will yield unbiased estimates of all coefficients including treatment variables. Imputation methods are acceptable
Missing at random	Data are missing in a way that is correlated only with observable variables	Complete case analysis will yield biased estimates of coefficients on missing variables, as will naive imputation. Regression, stochastic and multiple imputation will produce unbiased estimates of coefficients, and multiple imputation will yield unbiased estimates of errors.
Missing not at random	Data are missing in a way that is correlated with unobservable variables (and observable variables)	Complete case analysis will yield biased estimates of coefficients. Imputation techniques will typically yield biased estimates of coefficients.
Missing experimentally at random (Assignment)	Data may be correlated with observable or unobservable variables, but not with treatment assignment	Complete case analysis will yield unbiased treatment effect estimates for participants who are observed
Missing experimentally at random (Impact)	Data may be correlated with observable or unobservable variables, but not with the magnitude of their latent treatment effect	Complete case analysis will yield unbiased treatment effect estimates for the full sample

For the purposes of a randomised controlled trial, the latter two forms of missingness are of most substantial concern. Where missingness is not correlated with treatment assignment, we are still usually able to extract a good estimate of the treatment effect on the observed, and if neither relevant covariates nor assignment are correlated with missingness, this effect on the observed participants is likely to be a good estimate of the effect on the treated within your sample.

The case for imputation is strongest where data are missing experimentally not at random.

Covariates

In general covariates which are missing but which were pre-specified in the analytical strategy should either be imputed using multiple imputation or null imputation. In general, multiple imputation is more power preserving, and more likely to preserve sample-unbiased estimates of the covariate coefficients where data are not MNAR, but will create biased estimates of coefficients and downward biased variances if data are MNAR. Null imputation preserves less power, but does not bias the estimate of the coefficient for the covariate for participants whose observations are not missing in the presence of non random missingness. In either case, the coefficient on the treatment variable should be unbiased.

Where aggregation can be used to preserve power, in general it should be. An example other this might be with missing baseline data including a covariate that takes the value of the baseline measure for those for whom it is observed and the cluster mean value for those where it is not.

Outcomes

Where there is no correlation between participants' missingness for their outcome measures and their treatment (MEARA), complete case analysis for outcomes should be preferred. Where data are missing experimentally not at random, imputation should be attempted where possible within treatment condition, and sensitivity to imputation assumptions used as a robustness check. In particular, where multiple imputation is used as a primary solution, last observation carried forward (LOCF), and/or control drifted observation carried forward (CSOCF) should be used as robustness checks.