Innovation Growth Lab

Guide to the analysis of quantitative data from randomised controlled trials

Updated October 2023

Introduction **3**

Key terms used in this guide **4**

Enabling others to follow and replicate your results 4

1 First steps **6**

1.1 Set out clearly what happened in the course of your trial 6

1.2 Revisit the minimum detectable effect size for your trial 7

1.3 Check for balance between the arms of your trial 8

2 Prepare your data for analysis **10**

2.1 Plot your data 10

2.2 Deal with outliers in the data 10

2.3 Deal with missing data 11

2.4 Construct variables required for analysis 14

Sidebar: Understanding causal inference **14**

Sidebar: Sampling-based tests vs randomization inference **17**

3 Estimate the effects of the treatment **18**

3.1 Plot your data 18

3.2 Estimate unadjusted treatment effects 18

3.3 Estimate treatment effects after controlling for covariates 19

3.4 Account for design complications 20

3.5 Account for multiple hypothesis testing 22

3.5 Examine sensitivity of your results to analytical choices 24

4 Further analysis **25**

4.1 If rates of compliance are low, examine the impact among those who took part in the intervention(s) 25

4.2 Examine heterogeneity in the treatment effects 27

5 Report on your findings **28**

5.1 Be clear in communicating the level of uncertainty in the results 28

5.2 Discuss the empirical significance of your findings 29

Further reading, resources & references 30

Acknowledgements

This guide was prepared by Rob Fuller, based on earlier work by Eszter Czibor, Debbie Blair and Seemanti Ghosh. Many thanks to Hugo Cuello, Triin Edovald, Ana Goicoechea, James Phipps and Anna Segura for their review and advice.

Introduction

The Innovation Growth Lab (IGL) is a global initiative that works to increase the impact of innovation and growth policy by ensuring that it is informed by new ideas and robust evidence. We work at the intersection of research and policy, where we help organisations become more experimental, test ideas, and learn from each other.

This document describes IGL's recommended approach to the analysis of quantitative data collected in the course of randomised controlled trials (RCTs) of interventions aimed at promoting entrepreneurship, innovation and growth. These guidelines are recommended for use both by IGL staff and by external partners in analysing the outcomes of RCTs.

We assume in these guidelines that an RCT has been successfully designed and conducted and that outcome data has been collected. (Guidance on the design and implementation of RCTs can be found in IGL's publication, *Running randomised controlled trials in innovation, entrepreneurship and growth: An introductory guide*.) This document provides guidance on how to proceed with the analysis itself. However, many of the decisions about the analytical approach should normally have been made in advance and registered in the trial protocol and/or statistical analysis plan. This guide should therefore also be consulted at the stage of drawing up these documents.

The guidance in this document should be treated as a default approach to follow: most of the content will be appropriate to most of the RCTs managed and/or supported by IGL. However, each trial has its own idiosyncrasies, and we recognise that there is no one-size-fits-all approach. There may be good reasons for diverging from the guidance described here in some cases – but if so, it is important to be aware of the reasons for doing so and the consequences of those decisions. If the trial you are working on is being supported by IGL, analysis approaches that diverge from the guidance here should be discussed with the IGL team in advance.

This guide prioritises transparency of the analysis and ease of interpretation of the results, rather than more complicated approaches that may require more care in interpretation. The guidance is based on frequentist statistics, and does not cover insights from Bayesian statistics.

We welcome any feedback on this guide or suggestions for revisions in the future. Please send any feedback to innovationgrowthlab@nesta.org.uk.

Key terms used in this guide

An **intervention** is a programme, a single activity or a set of activities that are carried out with the aim of achieving some change in the world. **Outcomes** are the various types of change that may take place as a result of the intervention. For example, in the case of a business support programme, the key outcomes of interest may include the knowledge or beliefs of managers, the practices used by the business (such as the adoption of specific

management practices or technologies), and longer-term outcomes such as turnover, profits or productivity. **Outcome measures** are the specific elements on which we can collect data to monitor changes in outcomes. We call the difference that an intervention makes to an outcome (whether positive or negative) an **impact** of that intervention.

In this document, the term **participants** refers to all the individuals or businesses that are included in your trial, whether or not they receive any support or take part in interventions. The participants are the **unit of analysis** for the trial. If the interventions tested in your trial are carried out with individuals and outcomes are assessed primarily at an individual level (such as knowledge, understanding or some individual behaviour), then the participants will be individuals. If the interventions are carried out for a business as a whole and the outcomes are primarily assessed at the business level (such as adoption of technologies or turnover), then the participants will be businesses.

The **arms** of your trial are the different treatment and control groups to which participants are randomly allocated. A simple RCT will have only two arms (a treatment group and a control group), but a more complicated RCT may have multiple treatment groups.

Enabling others to follow and replicate your results

Transparency and reproducibility are key principles of good quality research. It is important to ensure that your work stands up to external scrutiny and that other researchers or evaluators could reproduce your findings if necessary. For example, in reporting the findings of your analysis, you should clearly specify how all primary and secondary outcomes and covariates are constructed and the precise statistical models used to derive your outcome estimates. If possible, the ideal is to make your anonymised data and code publicly available.[1]

Another key step to ensure that results produced in your trial are reproducible is to document in advance what hypotheses will be tested and how the analysis will be carried out. Such pre-commitment to the form of the analysis improves the credibility of the findings of a trial, by demonstrating that the researcher has not engaged (even unconsciously) in 'specification search'.[2] Planning the analysis carefully in advance also enables the evaluator or researcher to carry out the analysis rapidly once the outcome data becomes available, so that the key findings from the trial can be made available in a timely fashion.

The pre-commitment to the analysis approach should be set out in one or both of the following documents:

- The **trial protocol**, a document that should be prepared before launching any RCT. As well as setting out the research question(s) to be tested, the details of the

[1] Anonymising data before publication is a complicated undertaking and requires great care. This subject is beyond the scope of this guide, but useful guidance is available from the UK Data Service.

[2] See, for example, Garret Christensen and Edward Miguel (2018), 'Transparency, Reproducibility, and the Credibility of Economics Research', *Journal of Economic Literature*, 56(3), 920–980, https://doi.org/10.1257/jel.20171350

intervention(s) and the evaluation design, the trial protocol should also describe the outcome measure(s) and give at least an overview of how the analysis will be carried out. The trial protocol should be finalised and uploaded to an online registry before the trial is launched. IGL normally suggests registering trials on the American Economic Association's RCT registry. Alternatives are the Open Science Framework registry or AsPredicted.[3]

- A **statistical analysis plan** (or pre-analysis plan), which gives full details of how the outcome measures and any other variables used in the analysis will be constructed, and the specific models and statistical tests for the analysis. In some cases, this document may be drafted at the same time as the trial protocol. However, it is often desirable to wait until after baseline data has been collected in order to use this data to inform the statistical analysis plan. In any case, the statistical analysis plan should be finalised before any outcome data is collected. The statistical analysis plan should be annexed to the trial protocol and uploaded to the same registry.

IGL has templates that are available to support teams in preparing trial protocols and statistical analysis plans.

Finally, it is also very important that the analysis you carry out is fully documented. Quantitative analysis will normally be carried out using statistical software such as R or Stata. All the steps of the data preparation, cleaning and analysis should be set out in a series of coding files that allow the results to be reproduced from the raw data. Code should be annotated with explanations of what is being done at each step, to assist other researchers. In particular, this will enable you to understand the code yourself if you need to return to it at a later date.[4]

[3] A key advantage of AsPredicted is that it allows a trial registration to be hidden from public view, unless and until the researcher requests that it is published. This can be useful if there is an important reason not to make information about the trial publicly available during the trial's lifetime.

[4] For useful guidance on this, see Jake Bowers and Maarten Voors (2016), 'How to improve your relationship with your future self', *Revista de ciencia política (Santiago)*, 36(3), 829–848, http://doi.org/10.4067/S0718-090X2016000300011

1 First steps

1.1 Set out clearly what happened in the course of your trial

Why do this?

Analysing and interpreting the results of your trial relies on having a good understanding of the rates (and, as far as possible, the causes) of:

- **Attrition**: The attrition rate refers to the proportion of participants who withdraw from the trial before the end. If participants leave the trial, you may not be able to collect data on final outcomes for them, and/or they may withdraw their consent for their data to be analysed.
- **Response** to survey(s): High non-response rates may mean that you do not have data on important outcomes for many of the participants.
- **Compliance** with the intervention(s): Compliance refers to whether the participants took part in the interventions that were intended for them. If there is significant non-compliance in the treatment group, the statistical power to detect impacts of the treatment is likely to be affected. In some cases, it is also possible that some of those allocated to the control group ended up participating in the treatment interventions, another factor that will reduce the power to detect differences in outcomes between the treatment and control groups.

Setting this information out clearly will enable you to assess whether any of these problems have affected your trial, and will help to convince a reader of your report that your conclusions are valid.

How to do this?

Make sure you report how many participants:

- Were recruited (i.e. agreed to participate in the trial), if applicable
- Were included in baseline data collection (before randomisation), if applicable
- Were randomly allocated between arms of the trial
- Participated in the interventions, as appropriate to each arm of the trial
- Were included in follow-up data collection (e.g. through responding to surveys), broken down by trial arm
- Were included in the analysis of primary and secondary outcomes, broken down by trial arm

We recommend including a CONSORT flowchart[5] or a similar diagram in your report, to map out the progress of all participants through the trial process – from recruitment and random allocation through to participation in the intervention(s) and then follow-up data collection and analysis.

In many trials, the number of participants included in the final analysis is lower than the number that were recruited at the start. If that is the case, it is important to explain why and how participants were lost at each stage, and to discuss whether this may result in either or both of:

- Bias between the treatment arms
- A change in the types of participants that the findings apply to.

1.2 Revisit the minimum detectable effect size for your trial

Why do this?

At the trial design stage you will have calculated the minimum detectable effect size (MDES) – that is, the minimum size of impact that your trial can be expected to detect, given the expected sample size and what is known about the outcome measure(s). However, the sample size you end up with for the analysis may be less than that used in the original MDES calculation. There are two reasons for updating your assessment of the MDES before carrying out any analysis:

- If your analysis has any null findings (i.e. is not able to reject the null hypothesis that the treatment did not have an effect), it will be useful to know how large the effect *would have needed to have been* for you to be confident that the trial would have detected it.
- If you find a treatment effect that is smaller than the MDES, it is likely that the result is exaggerated, even if it has a small p-value.[6]

How to do this?

Once the final dataset is available but before you carry out any analysis, revise the power calculations set out in the trial protocol, using the figures that are now available for:

- The sample size available for analysis

[5] Kenneth F Schulz, Douglas G Altman and Daniel Moher (2010), 'CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials', *BMJ*, 340:c332, https://doi.org/10.1136/bmj.c332. Code to produce CONSORT-style flowcharts in R and Stata is available at https://cran.r-project.org/web/packages/visR/vignettes/Consort_flow_diagram.html and https://github.com/IsaacDodd/flowchart/.

[6] For an explanation of this point, see Andrew Gelman and John Carlin (2014), 'Beyond power calculations: assessing type S (sign) and type M (magnitude) errors', *Perspectives on Psychological Science*, 9(6), 641–651, https://doi.org/10.1177/1745691614551642

- Estimates of the standard deviation of the primary outcome measure(s). These can be taken either from the baseline data or from the follow-up data among the control group only.
- Estimates of the proportion of the variance in the primary outcome measure(s) that is explained by covariates that will be included in your regression models. This can be estimated either from the baseline data or from the follow-up data among the control group only.
- *(for clustered RCTs only)* Estimates of the intra-cluster correlation of the primary outcome measure(s). These can be taken either from the baseline data or from the follow-up data in the control group only.

Note that it is not valid to use the estimated effect size from the trial to assess the power of estimating an effect of that size.[7]

1.3 Check for balance between the arms of your trial

Why do this?

Balance tests will help to assess whether there are systematic differences in the observable baseline characteristics (i.e. those characteristics for which you have baseline data) between the participants in the different arms of your trial. If there were no attrition in your trial, then the randomisation process should result in comparable groups in each of the arms (provided that the sample size was reasonably large); in this case, a balance check will help to confirm that there were no problems with the randomisation process. If there is significant attrition in your trial, then the balance check will help to assess how comparable the groups are at the analysis stage.

How to do this?

Calculate the average (mean) figures for each of the baseline characteristics for which you collected data, between each arm of the trial.

Do this both on the original sample of participants as randomised, and on the sample that you have available for analysis.

We do not recommend checking for the statistical significance of differences between the arms for each of the baseline characteristics separately: if there are a moderate number of covariates, then it will be likely that some differences will show as statistically significant by chance. Instead:

- Examine the **size of the differences**, rather than their statistical significance. In

[7] See John M Hoenig and Dennis M Heisey (2001), 'The abuse of power: The pervasive fallacy of power calculations for data analysis', *The American Statistician*, 55(1), 19–24, https://doi.org/10.1198/000313001300339897 or David McKenzie and Owen Ozier (2016), 'Why ex-post power using estimated effect sizes is bad, but an ex-post MDE is not', World Bank, https://blogs.worldbank.org/impactevaluations/why-ex-post-power-using-estimated-effect-sizes-bad-ex-post-mde-not

particular, focus on balance among those characteristics that you expect to be **strongly correlated** with the outcome or that could help/prevent your intervention from working. In order to compare the size of the differences between covariates, it is helpful to *normalise* or *standardise* them, by dividing the difference in the means between the trial arms by the pooled standard deviation. Imbalances of larger than 0.1 could be considered problematic, particularly if they are in variables that are highly predictive of outcomes.[8]

- Use a **joint test**: Test whether the observable characteristics jointly predict treatment assignment, using the same specifications that you will use to test for differences in outcomes. This can be done as long as the number of participants in the dataset is considerably larger than the number of characteristics to be tested. The joint test can be carried out by constructing a regression model of the form

$$T_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_n X_{n,i} + \epsilon_i$$

where T_i is an indicator of the trial arm (normally $T_i = 1$ for units in the treatment group and $T_i = 0$ for units in the control group), $X_{n,i}$ are the *n* different baseline characteristics, and ϵ_i is the error term. Then test the hypothesis that the covariates do not predict the trial arm – that is, test for

$$\beta_1 = \beta_2 = \ldots = \beta_n = 0$$

using an F-test (if using a linear regression model) or a χ^2-test (if using a logit or probit regression model). If there are more than two trial arms, either carry out this procedure in pairs or use a multinomial logit or probit model.[9]

If you find large differences in baseline characteristics between the trial arms, it is important to assess why this has happened. Some potential explanations are:

- Error or other disruption in the random assignment process: In this case, you will need to investigate exactly what occurred, and assess how much this compromises the comparison of outcomes between the trial arms.
- Differential attrition or survey response between the trial arms: It may be clear that this is a problem if you find large differences in attrition rates or response rates between the arms. However, it is possible that there is unobservable attrition bias or non-response bias even if the overall *rates* of attrition or response are similar between the arms. For example, it is possible that a treatment affects the type of participants who are most likely to respond to a survey. In either case, differential attrition or non-response has the potential to bias any comparison of outcomes between the trial arms. Some suggestions for dealing with this are discussed in Section 2.3.

[8] See, for example, Peter C. Austin (2009), 'Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples', *Statistics in Medicine* 28(25), 3083–3107, https://doi.org/10.1002/sim.3697

[9] David McKenzie (2015), 'Tools of the Trade: a joint test of orthogonality when testing for balance', World Bank, https://blogs.worldbank.org/impactevaluations/tools-trade-joint-test-orthogonality-when-testing-balance

2 Prepare your data for analysis

2.1 Plot your data

Why do this?

Plotting your data before carrying out any analysis enables you to understand the distribution of your outcome and control variables and to identify errors or anything else of concern (such as outliers or floor or ceiling effects).

How to do this?

Generate a simple histogram or scatter plot for each of the variables you are using in your analysis. Scatter plots can be particularly useful to examine the relationship between two variables.

Figure 1: Example of a histogram, showing the distribution of a baseline variable

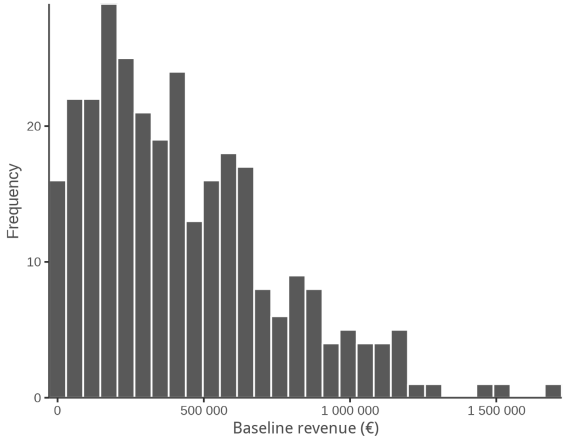
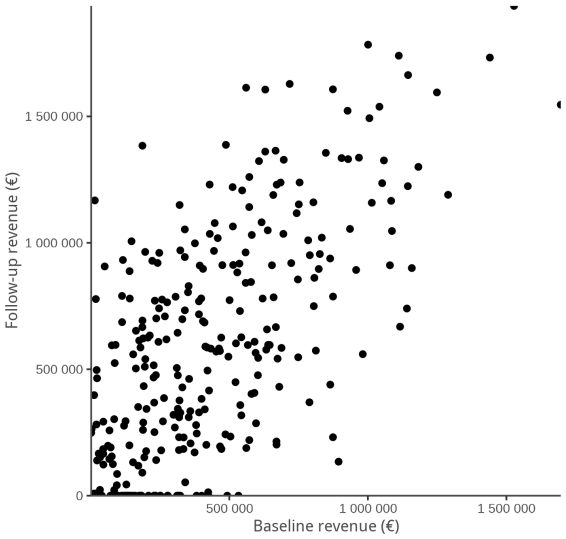


Figure 2: Example of a scatter plot, showing the distribution of the outcome variable at baseline and follow-up



2.2 Deal with outliers in the data

Why do this?

Outliers are values in your dataset that are notably different from other data points, and they can cause problems in statistical procedures. They can be identified from visual inspection of the data, either by using a histogram or a boxplot or by examining the maximum and minimum values for a variable.

Outliers can arise for two reasons:

- Measurement error: e.g. a user may accidentally add a zero at the end of a number when completing a survey. Common sense will play a role in identifying these cases. For example, if a microbusiness is recorded with turnover in the billions of euros, this is likely to be an error. It is important to have robust quality assurance procedures in place while data is being collected, to minimise the number of errors in the dataset.[10]

- True outliers: Some units really will have values that are substantially higher than typical in the dataset. For example, there may be a few businesses in the dataset with revenue that is much higher than average.

In some cases it may not be possible to identify whether an outlier is the result of an error or is a true figure that is higher than many others.

How to do this?

If outliers appear to be the result of errors, then it is normally best to clean them at this stage. This should be done by using judgement to correct for what appears to be the cause of the error.

If it is possible that there are true outliers in the data, it may still be necessary to take steps to deal with them, since they may skew later analysis. We recommend one of the following approaches:

- Include the outliers in the main analysis, but also show how the results change when outliers are excluded. This allows you to assess the robustness of the findings to the presence of these extreme cases.

- Use a weighted least squares regression, which offers a compromise between excluding these points entirely from the analysis and treating all of the data points equally in a regression model. The idea is to weigh the observations differently on the basis of how closely an observation fits the regression line. Values with large deviations from the line of best fit (i.e. those with large residuals) are given less weight in the regression analysis; as the absolute value of the residual decreases, the weight of the observation is increased.

2.3 Deal with missing data

Why do this?

The datasets we use to analyse the results of trials – particularly those based on surveys – often have missing data. In some cases individuals may decline to participate in a survey, resulting in all data points for that individual being missing for that round of data collection. In other cases, individuals may decide not to respond to or mistakenly omit

[10] For example, following Innovation for Poverty Action's procedures on high-frequency checking will help to minimise errors when collecting survey data: https://github.com/PovertyAction/high-frequency-checks/wiki

specific survey questions, meaning that data is missing for specific variables (this is often referred to as 'item non-response').

Missing data has three important consequences for the analysis of outcomes from a trial:

- If units are dropped from the analysis because data points are missing, this reduces the statistical power available.
- If the reasons for data being missing depend on respondent characteristics (for example, if businesses that are busier are less likely to respond to a survey), this affects the population to which the results apply. This must be considered when interpreting the results of the trial.
- If the reasons for data being missing differ between arms of the trial (e.g. if the control group are less satisfied with their experience in the trial and more likely to decline to respond to a survey than the treatment group), then dropping these participants from the analysis may result in the estimates of outcomes being biased.

Again, the ideal is to avoid having significant missing data in your dataset, through following robust quality assurance procedures at the data collection stage.[11]

How to do this?

Two factors are important when dealing with missing data: the **extent** of the missingness and the **patterns** of missingness. You should assess and discuss the following in the report:

- How many units are complete cases (i.e. those without any missing data).
- How many units have missing data for whole rounds/waves of data collection. (This should be documented in the CONSORT flowchart – see section 1.1.)
- The extent of item non-response for outcome variables and control variables, for units that did participate in the relevant rounds of data collection.
- The reasons for non-response or the mechanisms that led to data not being available, either for whole rounds of data collection or for specific data points.

The approach to dealing with missing data depends on whether this is among the outcome variable or (in rare cases) the treatment indicator, or among the control variables.[12]

Outcome variable or treatment indicator missing:

If data is missing completely at random, then carrying out the analysis only with units for which the data is complete will not bias the results. This may happen, for example, if a

[11] See the previous footnote for a useful resource on error checking during the collection of survey data.

[12] For further discussion on both cases, see Tara Slough (not dated), '10 things to know about missing data', Evidence in Government and Politics, https://egap.org/resource/10-things-to-know-about-missing-data/

technical fault with the data-collection system resulted in some potential survey respondents not being able to submit their responses, and that the problem was equally likely to affect all respondents. However, it is very unusual for data to be missing completely at random. If the problem with the data-collection system occurred towards the start of the data-collection period, and if participants who had a more positive experience in the trial were more likely to have tried to respond to the survey during that initial period than those in the control group, the pattern of missing data would not be random.[13]

Even if the reasons for data being missing are not random, it may still be possible to estimate the treatment effects among the subgroup of those for whom data is not missing, as long as the extent and pattern of missingness is similar between the trial arms. A first test of this condition is to examine whether the proportions with missing data are similar between each of the trial arms. If those proportions are similar, you can go on to examine whether any participant characteristics for which you have data (e.g. from a baseline survey) are associated with missingness. This can be done by regressing an independent variable designating whether there is missing data for a particular unit (e.g. defined as $M_i = 1$ for units with missing data and $M_i = 0$ for units without missing data) on any and all participant characteristics that are thought to be potential predictors of missingness. (This analysis can be carried out either for a whole round of data collection or for a particular variable for which there is missing data.) If no such associations are found, this may provide some confidence in restricting the estimation of treatment effects to those with non-missing data. However, it is still possible that there are correlations between missingness and participant characteristics in unobserved characteristics (those for which you do not have data), which could bias the estimation of treatment effects.

It is not possible to know how much bias missing data introduces into the estimates of treatment effects. However, it is possible to put **bounds** on these estimates – that is, to define the range of values over which the treatment effect could vary as a result of bias from the missing data.[14] **Manski bounds** can be derived by replacing missing values in the outcome measures with the theoretical lowest possible and the theoretical highest possible values of those measures. However, if the number of missing observations is high, then these bounds will be wide.[15] An alternative is to use **Lee bounds**: these are narrower but rely on an additional assumption (that of 'monotonicity': treatment assignment can lead either to otherwise-missing data to become non-missing or to otherwise-non-missing data to become missing, but not both) and restrict the effect estimates to the units for which data would be non-missing which trial arm they were assigned to.[16]

---

[13] Technically, even if the process that led to some data being missing is not random, resulting estimates will not be biased if missingness is independent of potential outcomes. However, there are few situations in which this could be argued to be the case.

[14] For further discussion, see Berk Özler (2017), 'Dealing with attrition in field experiments', World Bank, https://blogs.worldbank.org/impactevaluations/dealing-attrition-field-experiments. For guidance on implementing these approaches in Stata, see Razan Amine (2022), 'Coder's corner: Manski and Lee bounds', University of Oxford, https://csae.web.ox.ac.uk/sites/default/files/csae/documents/media/coderscorner_ht22week5fm.pdf.

[15] Joel L. Horowitz and Charles F. Manski (2000), 'Nonparametric analysis of randomized experiments with missing covariate and outcome data', *Journal of the American Statistical Association*, 95(449), 77–84, https://doi.org/10.2307/2669526

[16] David S. Lee (2009), 'Training, wages, and sample selection: Estimating sharp bounds on treatment effects', *The Review of Economic Studies*, 76(3), 1071–1102, https://doi.org/10.1111/j.1467-937X.2009.00536.x

In summary:

- If missingness is truly random or only a small proportion of data is missing, carry out the analysis on the sample with non-missing values.
- If missingness is not at random or affects a substantial number of cases, carry out the analysis on the sample with non-missing values but calculate bounds for the treatment effect. Remember when interpreting the results that the findings apply only to the subpopulation with non-missing data, not to the whole population from which the trial participants were originally sampled.

Control variables missing:

Include observations in the analysis by 'imputing' (filling in) missing values:[17]

- If less than 10% of units have missing data, replace the missing value with the unconditional mean of the variable in the non-missing observations.
- If more than 10% of units have missing data, create an additional binary variable indicating whether data is missing for a particular observation, and replace the missing values by zero. Include the indicator of missingness in the analysis as a control variable, alongside the variable itself.

An alternate route is multiple imputation. This method specifies multiple (*N*, where $N > 1$) imputation models, rather than just a single imputation model. As such, *N* complete data sets are obtained by imputing the missing values *N* times. Using each of the imputed data sets, the analysis of interest is carried out *N* times with the *N* estimates being combined into a single result.

2.4 Construct variables required for analysis

Why do this?

Analysis of outcomes is not always based directly on variables recorded in the raw data. Often the outcome variables and/or covariates need to be constructed from one or more of the variables recorded in the raw data. For example, a business's productivity may be calculated from turnover, cost and employment figures reported in a survey.

How to do this?

Construct the variables from the raw data, using the same procedure for all units in the dataset. The procedure to be used for each variable should normally be specified in your

[17] This guidance follows Winston Lin, Donald P. Green, and Alexander Coppock (2016), 'Standard operating procedures for Don Green's lab at Columbia', https://alexandercoppock.com/Green-Lab-SOP/Green_Lab_SOP.html. Justification for this approach is given by Anqi Zhao and Peng Ding (2021), 'To adjust or not to adjust? Estimating the average treatment effect in randomized experiments with missing covariates', *arXiv*:2108.00152, https://doi.org/10.48550/arXiv.2108.00152 , summarised by Berk Özler (2023), 'Missing values of baseline covariates in RCTs: an old favorite gets the nod…', World Bank, https://blogs.worldbank.org/impactevaluations/missing-values-baseline-covariates-rcts-old-favorite-gets-nod.

statistical analysis plan. It may occasionally be necessary to diverge from the plan (for example, if a problem during data collection meant that one or more variables were not collected for some or all units), but any such divergences should be clearly described and justified in the report. In these cases, it is up to you to convince a future reader that any divergences from the statistical analysis plan were necessary and reasonable – and in particular were not driven by a desire to produce more favourable results from the analysis.

Once you have created all the outcome variables and covariates required for your analysis, you should also plot them and examine any outliers, as described in Sections 2.1 and 2.2.

Sidebar: Understanding causal inference

The potential-outcome or counterfactual-based model of causal inference

In the potential outcomes framework, a causal effect is understood as a comparison between outcomes in two states of the world: the *actual* state and the *counterfactual* state.

Let's consider the effect of a single binary intervention. T_i is an indicator variable that takes on a value of 1 if a particular unit *i* receives the treatment, and 0 otherwise. The potential outcomes for this unit are Y_i^1 if *i* receives the intervention, and Y_i^0 if it does not. In the real world only one of the two potential outcomes, Y_i^1 or Y_i^0 actually occurs. We call this the *realised* outcome.

We can use this framework to define the causal effect of the intervention on unit *i* – that is, the **unit-specific treatment effect** – as the difference between two potential outcomes: $\delta_i = Y_i^1 - Y_i^0$.

> *A classic example is the causal effect of aspirin at reducing headache severity. Here the actual state of the world is the person who took aspirin and reports the severity of their headache; the counterfactual state is the severity of that headache had the person not taken the aspirin. The difference between these two outcomes would be considered the causal effect of taking aspirin of the severity of this person's headache.*

For any one unit *i*, we can only ever observe one state of the world, in which *i* received the treatment or *i* did not receive the treatment. The counterfactual state – that which did not occur – is by definition unobservable. As a consequence, the unit-specific treatment effect δ_i is also unobservable. This inability to observe the effect of a

treatment on a specific individual is known as the **fundamental problem of causal inference**.

However, it is possible to aggregate unit-specific treatment effects. The effect of a treatment overall can be understood as the average of the effects on individual units. This **average treatment effect (ATE)** is therefore defined as:

$$\begin{aligned} ATE &= \delta^{*} = E[\delta_i] \\ &= E[Y_i^1 - Y_i^0] \\ &= E[Y_i^1] - E[Y_i^0] \end{aligned}$$

Calculating the ATE precisely would still require us to observe both potential outcomes, which is inherently impossible. However, it is possible to *estimate* the average treatment effect.

Estimating the average treatment effect

When analysing data from RCTs, researchers usually estimate the ATE as the difference between the average outcomes of those who were randomly allocated to receive the treatment and those who were not:

$$\delta = E[Y_i^1 \mid T_i = 1] - E[Y_i^0 \mid T_i = 0]$$

This approach to estimating the ATE depends critically on the following assumptions:

- **Conditional independence:** A unit's treatment assignment is independent of its potential outcomes, conditional on observable covariates. This implies that a unit's assignment to the treatment or control group must be unrelated to its potential gains from receiving the treatment. Random assignment to treatment automatically results in the conditional independence assumption being fulfilled. However, attrition or non-response to surveys may result in treatment assignment being correlated with potential outcomes among the subset of participants whose data is available for analysis – this is the reason for being careful over the handling of missing data, discussed in section 2.3.
- **Stable unit treatment value assumption (SUTVA):** This refers to the following two conditions:
    - **No interference:** The potential outcome of an individual is unrelated to the treatment status of any other individual. This assumption is violated if there are spillovers between the treatment and control groups, or if there are general equilibrium effects. Such effects may be positive (for example, if the treatment group share information about training they have received under the intervention with members of the control group)

or negative (for example, if an intervention enables the treatment group to compete for customers at the expense of the control group). The potential for spillovers or general equilibrium effects should be considered at the outset, and the RCT designed to minimise the potential for these (for example, by using cluster randomisation). There is little potential for using analytical approaches to deal with any such problems once the trial has been implemented, though data from the trial may be helpful for assessing the size of any spillovers. For example, the implementation and process evaluation can examine whether the control group participated in any of the interventions intended for the treatment group.

- No difference in dosage: Each participant receives the same versions/dosage of treatment, ruling out different potential outcomes between units arising from different levels of exposure.

Intention-to-treat (ITT) estimator

Typically in a RCT, the average effect we are seeking to estimate (i.e. the 'estimand') is the **intention-to-treat (ITT)** effect. The ITT compares the average outcomes of the participants who were randomly assigned to treatment and those who were randomly assigned to control, *regardless of whether they actually participated in or received the corresponding intervention(s)*. This approach is adopted in order to preserve the benefit of randomisation. If the analysis were restricted only to those who chose to participate in the interventions offered to them, then it is likely that this would bias the treatment/control comparison. In addition, the ITT is often the quantity that is of most interest for understanding the impact of an intervention, taking account of individuals' decisions about whether to participate.

Further reading

Susan Athey and Guido M. Imbens (2017), 'The econometrics of randomized experiments', in *Handbook of economic field experiments*, https://doi.org/10.1016/bs.hefe.2016.10.003 (open-access version: https://arxiv.org/pdf/1607.00698.pdf)

Sidebar: Sampling-based tests vs randomization inference

When analysing data from randomised experiments, we can choose between two approaches to calculating *p*-values for our hypothesis tests regarding the size of the average treatment effect:

- The conventional, sampling-based approach, where we assume that treatment assignment is fixed, outcomes are random, and subjects are drawn from a larger

population. Our inference relies on assumptions regarding the sample size and error structure.

- Randomisation-based inference, which treats subjects' potential outcomes as fixed and considers their assignment to treatment as random. Many experts view randomisation inference as strongly preferable for data from randomised experiments. Note, however, that this approach provides an exact test of a sharp null: rather than testing whether the ATE is zero, this approach tests the null hypothesis that the treatment had no effect on any participant at all.[18]

In what follows, unless otherwise specified, we will be assuming that the analysis follows a sampling-based approach. When the randomisation procedure is complicated or in case of multiple comparisons, we do recommend that a randomisation inference-based approach is considered.

[18] For more information on randomisation-based inference, see Donald Green (not dated), '10 things to know about randomization inference', Evidence in Governance and Politics, https://egap.org/resource/10-things-to-know-about-randomization-inference/

3 Estimate the effects of the treatment

3.1 Plot your data

Why do this?

A simple graphical comparison of the main outcome measure in each arm of your trial helps to demonstrate the size of the difference between the outcome measures in each treatment group.

How to do this?

Plot the outcome variable among each treatment arm, on the same pair or axes. The two most straightforward plots are:

- Bar chart of the mean of the outcome variable, with confidence intervals shown as error bars. If baseline data for the outcome variable is available, this can also be shown in the bar chart, alongside the follow-up data.
- Histogram or distribution function of the outcome variable, to see how the distribution differs between the treatment arms.

Figure 3: Example of a bar chart, comparing the treatment and control groups in terms of the outcome variable at both baseline and follow-up

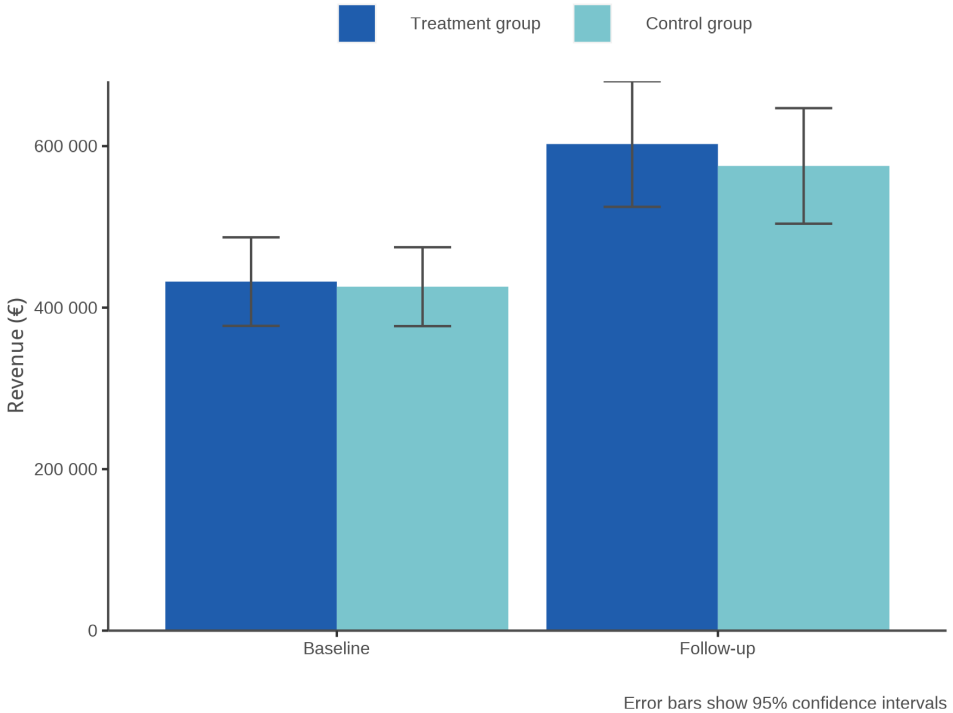
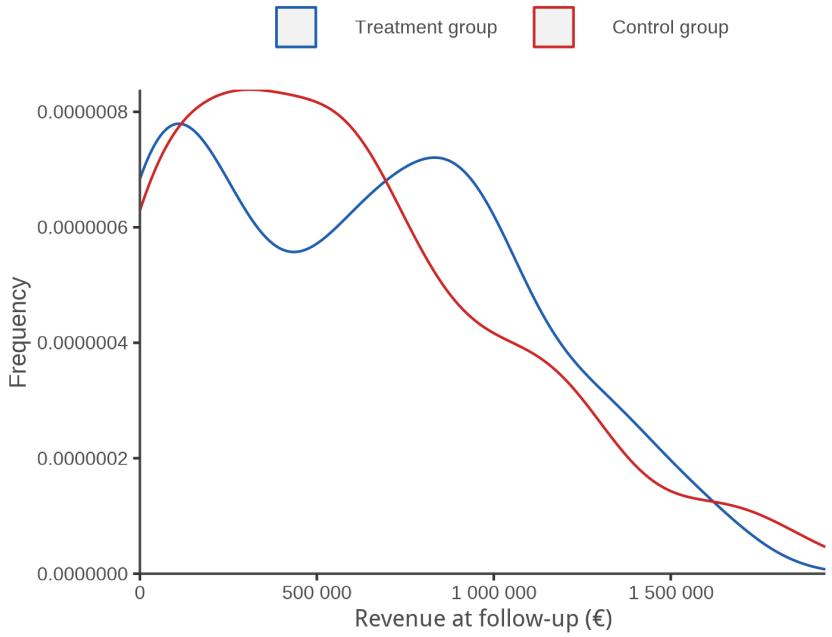


Figure 4: Example of a plot showing the distribution of the outcome variable in the treatment and control groups



3.2 Estimate unadjusted treatment effects

Why do this?

The unadjusted treatment effects provide an initial indication of the impact of the treatment(s) on the outcome measures and the associated level of uncertainty.

How to do this?

Begin by reporting the mean value of the outcome variable in the different treatment arms.

Estimate the size of the difference between the treatment arms and the level of uncertainty in those estimates, using a simple regression model of the form

$$Y_i = \alpha + \beta T_i + \varepsilon_i$$

where Y_i is the outcome variable, T_i is the treatment indicator, and ε_i is the error term.

If the outcome measure Y_i is continuous, we recommend using an ordinary least squares (OLS) regression model.

If the outcome measure Y_i is binary, we usually recommend using an OLS regression model (i.e. a linear probability model, LPM) anyway. Coefficients from the LPM are easier to interpret (especially when estimating a model with interaction terms) and LPMs tend to fit

the data well, unless the probabilities are very close to 0 or 1.[19] If you are concerned about unboundedness – that is, about predicted probabilities lying outside the 0 to 1 interval – a practical approach is to estimate the LPM and check how often this occurs in reality. If there is a good reason for preferring to use a logit or probit model to estimate the effect on a binary variable, we recommend that you report the average *marginal effects* rather than estimated coefficients.[20] We also recommend that you check that the estimates from the logit or probit model are similar to those derived from an LPM.

Note that, if there are no complications in your analysis (in particular, no clustering that requires adjustment at the analysis stage), then estimation using OLS is equivalent to carrying out a *t*-test for the difference in means.

Report the confidence intervals for the difference between groups as estimated from those tests. (*p*-values can also be reported, though these are more difficult to interpret for the typical reader than confidence intervals. Report exact *p*-values: do not summarise the results with statements such as 'less than 0.05' or 'significant at the 5% level'. This is discussed more in Section 5.1.)

3.3 Estimate treatment effects after controlling for covariates

Why do this?

Adding relevant covariates (control variables) to a regression model can improve the precision of the estimates. If there are baseline imbalances in key observable characteristics between the treatment arms (see Section 1.3), then controlling for these imbalances also acts as a test of the robustness of the findings. Including covariates should only change the precision of the treatment effect estimates, not the size of those estimates.

How to do this?

Add relevant covariates to the regression model used in Section 3.2, to produce a model of the form

$$Y_i = \alpha + \beta T_i + \gamma X_i + \varepsilon_i$$

where X_i is a vector of covariates. As discussed in Section 3.2, we normally recommend estimating this model through OLS, even if the outcome variable is binary.

---

[19] For discussion of this point, see Paul von Hippel (2015), 'Linear vs. logistic probability models: Which is better, and when?', Statistical Horizons, https://statisticalhorizons.com/linear-vs-logistic/ , Jed Friedman (2012), 'Whether to probit or to probe it: In defense of the Linear Probability Model', World Bank, https://blogs.worldbank.org/impactevaluations/whether-to-probit-or-to-probe-it-in-defense-of-the-linear-probability-model, and Robin Gomila (2021), 'Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis', *Journal of Experimental Psychology: General*, 150(4), 700–709, https://doi.org/10.1037/xge0000920 (open-access version: https://www.robingomila.com/files/publications_pdfs/Gomila_2020_Logistic_vs_Linear.pdf)

[20] For guidance on doing this in Stata, see Richard Williams (2012), 'Using the margins command to estimate and interpret adjusted predictions and marginal effects', *The Stata Journal*, 12(2), 308–331, https://www.stata-journal.com/article.html?article=st0260

The following variables should be used as covariates:

- Any variables that were used for stratification
- Any variables that were used to determine the probability of being assigned to treatment
- Other variables that are strong predictors of the outcome variable, based either on theory or prior literature, or by testing for correlations in the baseline data. We do not recommend including covariates simply because there were baseline imbalances – there should be an empirical or theoretical reason for including them.

Including covariates that are not strong predictors of the outcome variable may reduce the precision of the estimates, by using up degrees of freedom. This is a particularly important consideration when working with a small sample size.

If data on the baseline value of the outcome variable is available, this is likely to be a strong predictor of the endline value, so should be included as a covariate in the regression. The regression model will then be of the form

$$Y_{i,1} = \alpha + \beta T_i + \gamma X_i + \delta Y_{i,0} + \varepsilon_i$$

where $Y_{i,0}$ and $Y_{i,1}$ are the values of the outcome variable measured at baseline and endline respectively. Note that this model is usually preferred to a difference-in-difference model, because it (a) typically has higher precision,[21] and (b) allows for the details of how the outcome variable has been measured to change between baseline and endline.[22]

If it is not clear which covariates to include, alternative sets of covariates can be used, as tests of the robustness of the results. However, the statistical analysis plan should clearly specify which covariates will be included in the primary analysis – i.e. in the model that will be treated as the main one for reporting purposes.

Covariates should have been measured prior to randomisation, unless they are characteristics that could not possibly be affected by the treatment. Characteristics that could have been affected by the treatment should never be included as covariates.

3.4 Account for trial design features

Why do this?

Some decisions made at the design stage of the trial can have important consequences that should be taken into account in the analysis. Ignoring these complications would

[21] Berk Özler (2015), 'Why is difference-in-difference estimation still so popular in experimental analysis?', https://blogs.worldbank.org/impactevaluations/why-difference-difference-estimation-still-so-popular-experimental-analysis

[22] David McKenzie (2015), 'Another reason to prefer Ancova: dealing with changes in measurement between baseline and follow-up', https://blogs.worldbank.org/impactevaluations/another-reason-prefer-ancova-dealing-changes-measurement-between-baseline-and-follow

result in misleading results from the trial. Here we discuss two of the most common types of complication that arise in trials support by IGL:

- Different units having different probabilities of being assigned to treatment
- Use of cluster randomisation

How to do this?

Unequal probabilities of treatment assignment

When units have differing probabilities of being assigned to different treatment arms, analysis based on the mean of the units will not be valid. For example, if the treatment probability is different in different blocks or strata, then treatment assignment will be correlated with background characteristics on which you blocked/stratified. There are two ways of dealing with this problem.[23]

1. Estimate the average treatment effect block by block and then the average of these effects across blocks, weighting by the size of the block relative to the entire sample.

2. Estimate the treatment effect with a single regression model as normal, but with the units weighted using inverse probability weighting (IPW). In IPW, weights are defined as $\frac{1}{p}$ for treated units and $\frac{1}{1-p}$ for control units, where p refers to the probability of assignment to treatment.

Clustered design

In some RCT designs, randomisation may be carried out at a higher level than the unit of analysis. For example, in an evaluation of a training scheme for employees, outcomes may be measured for individual employees, but entire businesses (each with multiple employees) are randomised to receive or not receive the intervention. This is known as a cluster RCT, with each business included in the trial forming a cluster.

If the number of clusters is reasonably large, we recommend running the analysis as described above, but to calculate cluster-robust standard errors.[24] However, when the number of clusters is small, cluster-robust standard errors tend to be biased downwards; in this case, using a randomisation-inference approach or wild bootstrap is preferable.[25]

---

[23] See Lindsay Dolan (not dated), '10 things to know about randomization', Evidence in Government and Politics, https://egap.org/resource/10-things-to-know-about-randomization/, section 6

[24] Note to Stata users: this should be done using the `vce(hc3)` option, not the `robust()` option. See Uri Simonsohn (2021), 'Hyping Fisher: The most cited 2019 QJE paper relied on an outdated Stata default to conclude regression p-values are inadequate', Data Colada, http://datacolada.org/99.

[25] See James G. MacKinnon and Matthew D. Webb (2018), 'The wild bootstrap for few (treated) clusters', *The Econometrics Journal*, 21(2), 114–135, https://doi.org/10.1111/ectj.12107 and James G. MacKinnon, Morten Ørregaard Nielsen and Matthew D. Webb (2022), 'Cluster-robust inference: A guide to empirical practice', *Journal of Econometrics*, https://doi.org/10.1016/j.jeconom.2022.04.001

The threshold for what counts as a large enough number of clusters depends on the specifics of the situation, but around 50 is a good guideline.[26,27]

An alternative option is to use multilevel modelling (also known as hierarchical linear modelling). This approach typically requires additional assumptions, which may or may not be justified in the specific case. However, this approach may be recommended when there is little data in some clusters.[28]

3.5 Account for multiple hypothesis testing

Why do this?

Multiple hypothesis testing (MHT), or the multiple comparisons problem, refers to the practice of simultaneously considering multiple statistical inferences. Most trials will be testing multiple hypotheses, as a consequence of having one or more of:

- Multiple outcome variables
- More than two treatment arms
- Subgroup analysis

Not accounting and correcting for MHT would increase the likelihood of obtaining false positive results in your analysis.[29] As an example, consider a study in which a researcher jointly tests *N* mutually independent hypotheses. The treatment in reality has no effect and therefore all of the null hypotheses are true and therefore should be accepted. Fixing the type I error rate for a single comparison at a level α, the probability of at least one false rejection among all comparisons in this case is $1 - (1 - \alpha)^N$. Setting α to the conventional level of 0.05, if there are just three hypotheses being tested, the probability of observing at least one false positive is more than 14%. If testing 14 or more hypotheses, the probability of obtaining at least one false positive exceeds 50%.

How to do this?

We recommend some combination of the following approaches:

- **Limit**: avoid too many comparisons:[30]

[26] See further discussion in A. Colin Cameron and Douglas L. Miller (2015), 'A practitioner's guide to cluster-robust inference', *Journal of Human Resources* 50(2), 317–372, https://www.jstor.org/stable/24735989.

[27] It is also possible to use a cluster-aggregated approach, in which treatment effects are calculated within each cluster and then aggregated. However, this is generally not statistically efficient.

[28] For an introduction to this topic, see Andrew Gelman, Jennifer Hill and Masanao Yajima (2012), 'Why we (usually) don't have to worry about multiple comparisons', *Journal of Research on Educational Effectiveness* 5(2), 189–211, https://doi.org/10.1080/19345747.2011.618213 (open-access version: https://stat.columbia.edu/~gelman/research/published/multiple2f.pdf)

[29] John A. List, Azeem M. Shaikh and Yang Xu (2019), 'Multiple hypothesis testing in experimental economics', *Experimental Economics* 22, 773–793, https://doi.org/10.1007/s10683-018-09597-5

[30] As recommended in IGL's 'Running randomised controlled trials in innovation, entrepreneurship and growth: An introductory guide', we recommend identifying between one and three primary research questions at the design stage of a trial.

    - Do not make all possible comparisons between treatment arms unless your theory suggests there is a good reason to.
    - Create (pre-specified) summary indices pooling multiple related outcomes into a single measure. (However, note that such a composite variable may be hard to interpret.)
    - Only conduct a few, pre-specified and theoretically motivated subgroup analyses – or use machine learning techniques to study heterogeneity in treatment response.[31]

- **Adjust** the *p*-values from your trial to reduce the probability of type I error:[32]
    - Control the family-wise error rate (the probability of obtaining one or more false positive results). The simplest approach to this is to use a Bonferroni correction: divide your critical alpha level by the number of comparisons. However, the Bonferroni correction tends to be overly conservative – i.e. it is likely to result in more type II errors. Other less-conservative approaches are also available, such as the Romano–Wolf step-down procedure.[33]
    - Control the false discovery rate (the proportion of statistically significant results that should be expected to be type I errors), using the approaches of Benjamini and Hocherg or Benjamini, Krieger, and Yekutieli.[34]
- **Acknowledge** the problem:
    - Report the number of comparisons conducted.
    - Label the results as suggestive if they do not hold up after a correction for MHT.
    - Describe additional analyses as exploratory research. These may be considered as indicative results, but would need to be examined in future trials before being considered as robust findings.

[31] See, for example, Susan Athey and Guido Imbens (2016), 'Recursive partitioning for heterogeneous causal effects', *Proceedings of the National Academy of Sciences*, 113(27), 7353–7360, https://doi.org/10.1073/pnas.1510489113

[32] For more guidance on approaches to adjustment, see Alexander Coppock (not dated), '10 things to know about multiple comparisons', Evidence in Government and Politics, https://egap.org/resource/10-things-to-know-about-multiple-comparisons/

[33] Code to implement these approaches in Stata is reviewed in David McKenzie (2021), 'An updated overview of multiple hypothesis testing commands in Stata', World Bank, https://blogs.worldbank.org/impactevaluations/updated-overview-multiple-hypothesis-testing-commands-stata

[34] See the summary in Michael L. Anderson (2008), 'Multiple inference and gender differences in the effects of early intervention: a reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects', *Journal of the American Statistical Association* 103(484), 1481–1495, https://doi.org/10.1198/016214508000000841

3.5 Examine sensitivity of your results to analytical choices

Why do this?

The previous sections of this guide have highlighted that there are many decisions to be made in analysing the results of a trial – such as the way the outcome variables are constructed from the raw data, or how missing values are handled. It is possible your findings may be sensitive to the specific decisions made, so it is valuable to test alternative specifications to check that you obtain similar results. While it is important to pre-commit to the core analysis that will be carried out and reported (so as to avoid '*p*-hacking' or specification search), it is valid to examine alternative specifications as a check on the robustness of the core results.

How to do this?

Repeat the analysis, using different analytical decisions. For example:

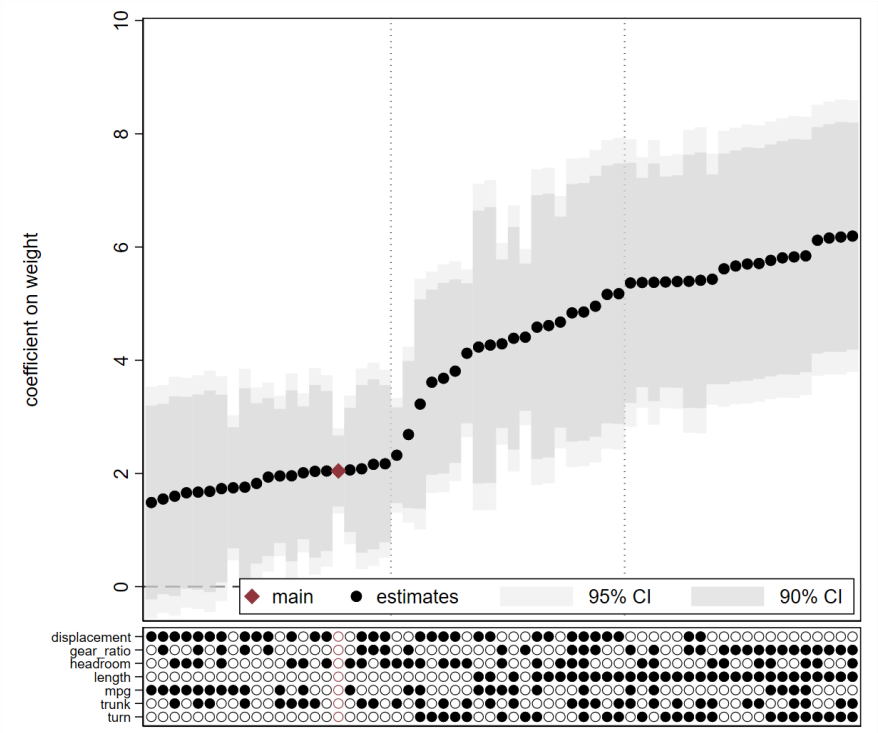
- Handling outliers in a different way (see section 2.2)
- Handling missing values in a different way (see section 2.3)
- Constructing outcome variables differently – e.g. using alternative thresholds for binary variables, or doing logarithmic transformation of continuous variables (see section 2.4)
- Using alternative estimation methods – e.g. using both a linear probability model or a logit or probit model for estimating results for binary outcome measures (see section 3.2)
- Including different sets of covariates in the regression model (see section 3.3)

It is not necessary to report the detailed result of all these robustness checks in a final report, but it is useful at least to mention which types of checks you have carried out. If there are notable differences in the results obtained, these should be discussed in the report – including an assessment of whether and to what extent this calls into question the findings of the core analysis.

If you are carrying out many robustness checks and wish to show visually how the analytical decisions affect the size of our outcome estimates, this can be done using a **specification curve**.[35]

[35] See Uri Simonsohn, Joseph P. Simmons & Leif D. Nelson (2020), 'Specification curve analysis', *Nature Human Behaviour*, 4, 1208–1214, https://doi.org/10.1038/s41562-020-0912-z (open-access version: https://urisohn.com/sohn_files/wp/wordpress/wp-content/uploads/specification-curve-published-hand-corrected.pdf). Packages for implementing this in R, Stata or Python are available at https://urisohn.com/specification-curve/.

Figure 5: Example of a specification curve, from Martin Andresen



Source: https://github.com/martin-andresen/speccurve/blob/master/

4 Further analysis

4.1 If rates of compliance are low, examine the impact among those who took part in the intervention(s)

Why do this?

'Compliance' refers to whether participants in the trial took part in the interventions that were intended for them. It is frequently the case that a proportion of those who register to take part in a trial and are randomly allocated to a treatment or control group do not take up whatever interventions are subsequently offered to them. In some trials it is also possible that some of those in the control group end up taking part in the intervention intended for the treatment group(s), whether deliberately or by accident.[36]

Even if rates of compliance are low, the intention-to-treat (ITT) estimates based on the original randomisation (discussed earlier) are unbiased, and should be reported. In such a case, the ITT estimate captures the impact of being assigned to the treatment group rather than the control group – i.e. the impact of being *offered* the treatment. However, this may not always be the effect that is of most interest. For example, if the treatment intervention has a large positive impact but few of those allocated to the treatment group actually participate, then the ITT estimate may turn out to be close to (and perhaps not statistically

[36] Occasionally the implementer of a trial may decide to make the treatment available to some members of the control group. This could happen, for example, if some of those who were allocated to the treatment group drop out but the implementing organisation needs to meet a target to deliver the intervention to a fixed number of participants.

distinguishable from) zero. In such a case, policymakers may also be interested in the causal impact of *taking part* in the treatment intervention(s).

How to do this?

If the ITT does not capture the effect of interest, you can supplement this by estimating the **local average treatment effect (LATE)**, also known as the **complier average causal effect (CACE)**. The LATE estimates the average causal impact of the treatment among the 'compliers': that is, those who choose to take up the treatment if and only if they are assigned to the treatment.

If it is likely that LATE analysis will be carried out in your trial, the specific definition of 'compliance' and how it will be measured should be agreed and recorded in the trial protocol, before the trial begins. For example, for the purposes of an evaluation of a business training programme, compliance may be defined as attendance at two thirds of the training sessions. If the treatment involves several different components, then 'compliance' could be defined as participation in all or a certain number of these components.[37]

The LATE should be estimated using an instrumental variable (IV) approach, in which initial random assignment to the treatment or control group is the instrument for compliance with the treatment.[38] This relies on two assumptions:

- Monotonicity: Being randomly assigned to the treatment group does not make one *less* likely to participate in the treatment than being randomly assigned to the control group.
- Exclusion restriction: It is the intervention itself that has an effect on outcomes, not the random assignment. This implies that the outcome is the same for those who would not have taken up the treatment, regardless of whether they are randomly assigned to treatment or control.

The analysis is normally carried out using two stage least squares (2SLS) regression analysis. Results for the first stage should be reported, as well as the correlation between the instrument and the endogenous variable; and an *F*-test.

4.2 Examine heterogeneity in the treatment effects

Why do this?

The approaches discussed in section 3 are aimed at assessing the impact of an intervention, averaged across all those that signed up to participate. However, we are often also interested in going beyond the average, to understand how the impacts are

[37] It is also possible to define multiple compliance thresholds (for example, minimal and optimal compliance), in order to estimate bounds for the treatment effects. Refer to Alan S. Gerber and Donald P. Green (2012), *Field experiments: Design, analysis, and interpretation*, W. W. Norton & Company, p. 165 for more information.

[38] Guido W. Imbens and Joshua D. Angrist (1994), 'Identification and estimation of local average treatment effects', *Econometrica*, 62(2), 467–475, https://doi.org/10.2307/2951620, open-access version: https://www.nber.org/papers/t0118

distributed across the participants. For example, we may be interested in questions such as:

- What is the impact on particular subgroups (e.g. on women-owned businesses, on micro, small or medium-sized businesses, or on rural v. urban businesses)?
- For which groups are the impacts small or large?
- Does the intervention have adverse effects for certain groups?

How to do this?

We discuss two approaches to examining heterogeneity in impacts: distribution impact analysis and subgroup analysis.

Distributional impact analysis

Distributional impact analysis takes two main forms.[39]

Firstly, we can examine impact on the outcome distributions, by directly comparing the distributions of outcomes under each of the trial arms. It is often useful to examine plots of this kind as the first step in analysis, as discussed in Section 3.1. However, this approach is not informative about how the programme's impact varies across individuals, because different individuals may lie in different parts of the distribution under each of the arms.

Secondly, distribution impact analysis can also be used to examine particular characteristics of the distribution of treatment impacts, such as what fraction of the population experiences negative impact from the programme (even if the average effect is positive). However, this second form requires strong assumptions about how participants fare in a counterfactual state.

Subgroup analysis

Subgroup analysis involves estimating the **conditional average treatment effect (CATE)**, that is the average treatment effect specific to a subgroup defined by participant characteristics (e.g., women-owned businesses or microbusinesses) or attributes of the context in which the experiment occurs (e.g., participants located in a specific region in a multi-regional experiment). Note that subgroups can only be defined by pre-intervention (i.e. baseline) characteristics.

It is important to specify in the trial protocol or the statistical analysis plan which groups you are planning to carry out subgroup analysis for. Note that conducting subgroup analysis is likely to lead to concerns about multiple comparisons (see Section 3.4).

In addition to estimating the CATEs separately, it may also be of interest to estimate the size of the difference between two CATEs – for example, to test whether an intervention had greater impact among women or men. This can be done by adding a term to the

[39] Guadalupe Bedoya, Luca Bittarello, Jonathan Davis and Nikolas Mittag (2018), 'Distributional impact analysis: Toolkit and illustrations of impacts beyond the average treatment effect', IZA Discussion Paper No. 11863, https://doi.org/10.2139/ssrn.3261720

regression model, accounting for the interaction between treatment status and the characteristic of interest, that is, estimating:

$$Y_i = \alpha + \beta T_i + \gamma X_i + \delta Z_i + \eta T_i Z_i + \varepsilon_i$$

Using the same notation as in sections 3.2 and 3.3, but where Z_i is an indicator variable for the specific characteristic of interest (for example, $Z_i = 1$ for women and $Z_i = 0$ for men. The coefficient η then provides an estimate of the interaction between the treatment and the characteristic (e.g. in our example, it estimates the additional impact among women, on top of the estimated impact among men that is given by the coefficient β). However, this analysis provides only a *descriptive* measure of the interaction between treatment and the characteristic: it cannot be taken as a causal relationship unless the characteristic of interest has been randomly assigned.[40]

5 Report on your findings

5.1 Be clear in communicating the level of uncertainty in the results

Why do this?

No matter how well designed and implemented your trial is there is always some uncertainty in the results. There are usually at least two reasons for this:

- When participants are selected to take part in a study, random sampling leads to sampling uncertainty. Even if the participants are sampled at random from the population, the characteristics of the sample will never be perfectly representative of the characteristics of the population as a whole.
- When participants are then randomly allocated to the treatment or control groups, random assignment leads to allocation uncertainty. Although we expect the randomisation to lead to the groups having similar characteristics on average, the individual participants in the treatment and control groups are different. The estimated effect size therefore depends to some extent on the random allocation of the individuals between these groups.

It is important that the reader is aware of the level of uncertainty when interpreting the results of the trial. We should communicate the range of values over which the estimated effect size should be expected to vary, if the same experiment were to be repeated in the same conditions. Conventionally this is done by discussing the confidence interval or compatibility interval of an estimated effect.[41]

[40] For more detail on this topic, see Albert Fang (not dated), '10 things to know about heterogeneous treatment effects', Evidence in Governance and Politics, https://egap.org/resource/10-things-to-know-about-heterogeneous-treatment-effects/

[41] For further guidance, see Education Endowment Foundation (2020), 'Statement on statistical significance and uncertainty of impact estimates for EEF evaluations', https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Research_Report/Statement_on_statistical_significance_and_uncertainty_of_impact_estimates_for_EEF_evaluations.pdf and Guido W. Imbens (2021), 'Statistical

How to do this?

When discussing the estimates of effect sizes, or of any tests of statistical significance, be sure to report the 95% confidence or compatibility intervals, alongside the point estimate. When summarising the results for a non-technical audience, statements such as this can be used: 'The best estimate of impact from the treatment is [point estimate], but the results are also compatible with an impact ranging from *[lower bound of confidence interval]* to *[upper bound of confidence interval]*.'

Do not make dichotomous statements about whether a particular estimate is or is not 'statistically significant'. If you are reporting *p*-values, then report the value itself, instead of a range like '$p < 0.05$' or '$p < 0.01$'.

5.2 Discuss the empirical significance of your findings

Why do this?

Even in a well designed, internally valid trial with statistically significant positive effects, we still need to think carefully about the empirical significance of the results. Is the effect we see 'meaningful' in the real world? This is not always easy to do, especially if the outcome is opaque. For instance, if the outcome is constructed using a factor analysis of many inputs it can be hard to interpret what any increase actually means. We should however, always attempt to assess the 'economic significance' of the results, or their relevance for policy and practice.

How to do this?

Some key questions to ask yourself when trying to answer this

- How does it compare to other programmes?
- How does it fit with the relevant literature?
- Is it cost-effective?[42]

Always include a discussion about the real world magnitude of effects, anchoring them in as meaningful a measure as possible.

Significance, p-Values, and the Reporting of Uncertainty', *Journal of Economic Perspectives*, 35(3), 157–74, https://doi.org/10.1257/jep.35.3.157

[42] This would involve comparing the relevant costs and benefits of an intervention to determine whether it represents value for money, perhaps also comparing it to estimates from alternative approaches that were not examined within this trial. How best to undertake cost benefit analysis can be very complex and is a topic we intend to expand on in future iterations of this guide.

Further reading, resources & references

Key references on specific topics are highlighted in footnotes in the text of the guide. Some useful references for overall guidance on the analysis process are:

Methods guides from Evidence on Governance and Politics (EGAP)

World Bank Development Impact blog, particularly this list of posts on technical topics

World Bank Development Impact (DIME) Wiki, particularly the sections on data cleaning, data analysis and reproducible research

Abdul Latif Jameel Poverty Action Lab (J-PAL) research resources, particularly section 6 on processing and analysis

Innovations for Poverty Action and Global Poverty Research Lab guide to data cleaning, including Stata code

Esther Duflo, Rachel Glennerster and Michael Kremer, 'Using randomization in development economics research: A toolkit' (open-access version here)

Macartan Humphreys, 'I saw your RCT and I have some worries! FAQs'

What Works Clearinghouse, 'Procedures handbook', version 4.0

Winston Lin, Donald P. Green, and Alexander Coppock, 'Standard operating procedures for Don Green's lab at Columbia'