Eric W Djimeu
Deo-Gracias Houndolo

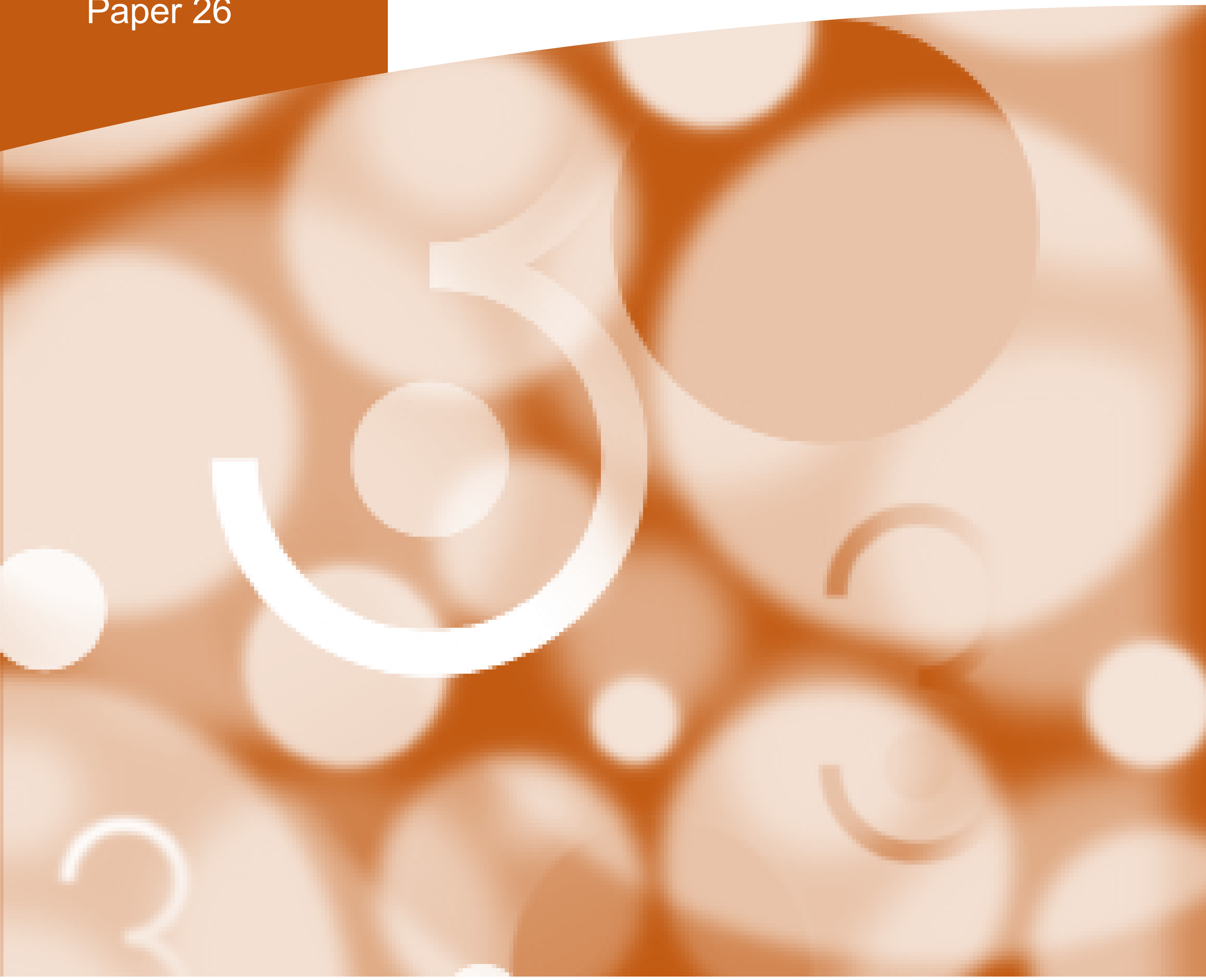
Power calculation for causal inference in social science

Sample size and minimum detectable effect determination

March 2016

Working
Paper 26

3ie manual



International
Initiative for
Impact Evaluation

Power calculation for causal inference in social science: sample size and minimum detectable effect determination

3ie impact evaluation manual

Eric W Djimeu
International Initiative for Impact Evaluation (3ie)

Deo-Gracias Houndolo
International Initiative for Impact Evaluation (3ie)

3ie Working Paper 26
March 2016

International Initiative for Impact Evaluation

About 3ie

The International Initiative for Impact Evaluation (3ie) is an international grant-making non-government organisation promoting evidence-informed development policies and programmes. We are the global leader in funding and producing high-quality evidence of what works, how, why and at what cost. We believe that better and policy-relevant evidence will make development more effective and improve people's lives.

3ie working papers

These papers cover a range of content. They may focus on current issues, debates and enduring challenges facing development policymakers and practitioners and the impact evaluation and systematic review communities. Policy-relevant papers draw on relevant findings from impact evaluations and systematic reviews funded by 3ie, as well as other rigorous evidence to offer insights, new analyses, findings and recommendations. Papers focusing on methods and technical guides also draw on similar sources to help advance understanding, design and use of rigorous and appropriate evaluations and reviews. 3ie also uses this series to publish lessons learned from 3ie grant-making.

About this working paper

This manual was written by 3ie evaluation specialists in response to growing demand for more technical guides on impact evaluation designs and methods. It covers experimental impact evaluations and is designed to be used in conjunction with the online *3ie Sample size and minimum detectable effect calculator*© developed in-house. Any errors and omissions are the sole responsibility of the authors. Any comments or queries should be directed to Eric W Djimeu at edjimeu@3ieimpact.org

Suggested citation: Djimeu, EW and Houndolo, D-G, 2016*. Power calculation for causal inference in social science: sample size and minimum detectable effect determination, 3ie impact evaluation manual, 3ie Working Paper 26*. New Delhi: International Initiative for Impact Evaluation (3ie)

3ie Working Paper Series executive editors: Beryl Leach and Emmanuel Jimenez
Production manager: Deepthy Menon
Assistant production manager: Pradeep Singh
Copy editor: Ruth Pitt
Proofreader: Sarah Chatwin
Cover design: John F McGill
Printer: Via Interactive

Acknowledgments

We thank Howard White for his guidance throughout the preparation of this manual. We thank Annette N Brown, Jyotsna Puri and Beryl Leach for their comments and suggestions regarding the content of this manual. We thank Heather Lanthorn for her detailed comments and suggestions on each part of this manual. We thank Benjamin Wood for his comments and suggestions. We thank the anonymous external reviewer for comments and suggestions shared with us. We also thank Flor Calvo for her excellent research assistance. Finally, we thank the International Initiative for Impact Evaluation (3ie) for funding this work. Any error in this document should be considered solely our own responsibility.

Summary

Experimental and quasi-experimental methods are increasingly used to evaluate the impact of development interventions. However, unless these methods use power calculations to determine sample sizes correctly, researchers are likely to reach incorrect conclusions about whether or not the intervention works.

This manual presents the basic statistical concepts used in power calculations for experimental design. It provides detailed definitions of parameters used to perform power calculations, useful rules of thumb and different approaches that can be used when performing power calculations. The authors draw from real world examples to calculate statistical power for individual and cluster randomised controlled trials. This manual provides formulae for sample size determination and minimum detectable effect associated with a given statistical power. The manual is accompanied by the *3ie Sample size and minimum detectable effect calculator©*, a free online tool that allows users to work directly with the formulae presented section 7 in the manual.

Note to readers

This manual provides straightforward guidance about the process of performing power calculations for experimental impact evaluation research designs. It is written for specialists and non-specialists involved in the design and implementation of impact evaluation studies with a minimal background in impact evaluation and basic knowledge of statistics. The focus is on experimental evaluation design. Having a manual that covers sample size calculations for quasi-experimental designs would also be useful. 3ie intends to develop tools for those designs in the near future. 3ie is also investigating how to further develop the Microsoft Excel®-based *3ie Sample size and minimum detectable effect calculator*© as an online resource.

Contents

List of figures and tables

Abbreviations and acronyms

FFS	farmer field school
H_0	null hypothesis
ICC	intra-cluster correlation
MDE	minimum detectable effect
R^2	coefficient of variation

1. Introduction

Since the 1990s, researchers have increasingly used experimental and quasi-experimental primary studies – collectively known as impact evaluations – to measure the effects of interventions, programmes and policies in low- and middle-income countries. However, we are not always able to learn as much from these studies as we would like. One common problem is when evaluation studies use sample sizes that are inappropriate for detecting whether meaningful effects have occurred or not. To overcome this problem, it is necessary to conduct power analysis during the study design phase to determine the sample size required to detect the effects of interest. Two main concerns support the need to perform power calculations in social science and international development impact evaluations: sample sizes can be too small and sample sizes can be too large.

In the first case, power calculation helps to avoid the consequences of having a sample that is too small to detect the smallest magnitude of interest in the outcome variable. Having a sample size smaller than statistically required increases the likelihood of researchers concluding that the evaluated intervention has no impact when the intervention does, indeed, cause a significant change relative to a counterfactual scenario. Such a finding might wrongly lead policymakers to cancel a development programme, or make counterproductive or even harmful changes in public policies. Given this risk, it is not acceptable to conclude that an intervention has no impact when the sample size used for the research is not sufficient to detect a meaningful difference between the treatment group and the control group.

In the second case, evaluation researchers must be good stewards of resources. Data collection is expensive and any extra unit of observation comes at a cost. Therefore, for cost-efficiency and value-for-money it is important to ensure that an evaluation research design does not use a larger sample size than is required to detect the **minimum detectable effect** (MDE) of interest. Researchers and funders should therefore use power calculations to determine the appropriate budget for an impact evaluation study.

Sample size determination and power calculation can be challenging, even for researchers aware of the problems of small sample sizes and insufficient power. 3ie developed this resource to help researchers with their search for the optimal sample size required to detect an MDE in the interventions they evaluate.

The manual provides straightforward guidance and explains the process of performing power calculations in different situations. To do so, it draws extensively on existing materials to calculate statistical power for individual and cluster randomised controlled trials. More specifically, this manual relies on Hayes and Bennett (1999) for cluster randomised controlled trials and documentation from Optimal Design software version 3.0 for individual randomised controlled trials.

The manual is organised into sections. Following the Introduction, Section 2 presents basic statistics concepts and discusses statistical rationale. Section 3 discusses the concept of power calculation and its applications. Sections 4 and 5 are developed around rules of thumb and common pitfalls in running power and sample size calculations. Section 6 presents formulae to calculate MDEs and sample sizes for experimental impact evaluation designs. The choice of formula depends on the identification strategy used, units of assignment and observation, and the type of outcome variable. Section 7 covers experimental design. Throughout the manual, each formula is followed by an example to make the concepts more tangible and intuitive.

Section 7 is designed to be used with an accompanying Microsoft Excel®-based customised worksheet, the *3ie Sample size and minimum detectable effect calculator©,* so that readers can run their own power analyses.

2. Basic statistics concepts: statistical logic

This section provides a background explanation of the basis for determining an appropriate sample size for evaluating the effectiveness of an experiment, intervention or treatment. It covers the following concepts: hypothesis testing, null and alternative hypotheses, type I error, type II error and p-value. This section may not be useful for readers already familiar with statistics and data analysis, but we encourage most users to read it.

As Prajapati, Dunne and Armstrong (2010) indicate, anyone who wants to calculate sample size and determine statistical power must first understand the notion of **hypothesis testing**. Here is a typical example that we can use to show how to draw a hypothesis and test it:

> Let’s assume that a developing country government is planning to launch an innovative youth employment scheme policy to increase employment rate from 40 per cent to 60 per cent in 3 years among young people aged between 23 and 32 years old. As the new policy has never been implemented in the country, the change expected (a predicted increase in employment rate) is considered as a hypothesis in statistical terms. In this setting, the government needs to find out whether the hypothesis is likely to be met in reality. Hence the government requests an evaluation design that would allow the probability of finding a true positive – when there really is an effect of a given size – to be at least 80 per cent (power) and the probability of making a false positive to be 5 per cent (significance level).

We would start by admitting the hypothesis that any intervention implemented would change something in that environment, and therefore it would yield *some* effect. This rationale would be applied to the youth employment policy. The core question is whether the hypothesised change is merely due to chance or if it is specifically *caused by* the intervention. *It is also essential to indicate the direction and magnitude of the expected or desired change. This step is crucial because it contributes to assessing the relevance of the change for policy and programming decisions.*

By convention, statisticians start from the premise that any effect observed from an intervention is caused by chance. In the case of impact evaluation, this premise would be that the treatment group is not distinguishable from the control group. Statisticians refer to this premise as the **null**

hypothesis. Obviously, the null hypothesis (noted H_0) may not be true. The policy being evaluated (for example, a youth employment scheme) may have caused the observed effect, and therefore a change in the outcome of interest (employment rate) would not be due to chance alone. The contrast to the null hypothesis is the **alternative hypothesis**. It states that the effect observed from an intervention did not occur just by chance, but is likely to be a real effect attributable to the intervention. This means that, given the assumptions of the identification strategy,[1] any change observed in employment rate would be the result of the government's new policy on youth employment.

If the difference in employment rates recorded before and after the implementation of the policy is not due to chance alone, then H_0 is rejected and the alternative hypothesis is favoured or accepted. Another way of reporting this change is that the effect of the intervention was statistically significant, meaning that the alternative hypothesis is accepted and the null hypothesis rejected. Similarly, a report of insignificant effects means that the null hypothesis cannot be rejected. In failing to reject the null hypothesis, we cannot conclude that the observed effects were due to anything but chance.

Deciding whether to accept or reject H_0 is based on a criterion or threshold chosen by the researcher. Different disciplines have different norms for setting this criterion, known as a **significance level**, from which equivalent **confidence level** is derived (and vice versa).

The significance level gives the probability of a **false positive** result – a result that indicates that a given condition is present when it is not, or that a treatment has an effect when it does not. In other words, the significance level is the probability of detecting an effect that is not present. The significance level is known as **α (alpha)**, and the confidence level is defined as $(1 - \alpha)$. The confidence level is the probability that we do not find a statistically significant effect if the treatment effect is zero. In social science, three significance levels (values of α) are commonly used: $\alpha = 10$ per cent, $\alpha = 5$ per cent and $\alpha = 1$ per cent. Therefore, three confidence levels are commonly used: 90 per cent, 95 per cent and 99 per cent. These values would be considered far too large in a field such as genetics or aeronautics, but are suitable for most social science research.

Returning to the case of the youth employment programme, where the youth employment rate is the outcome of interest, a confidence level of 95 per cent would indicate that, if the new employment policy is repeated in similar settings 100 times, then 95 times out of 100 it would produce an effect that would be less than or equal to 1.96 times the **standard deviation** of the youth employment rate before the intervention. Hence, in 95 per cent of cases, the null hypothesis would be accepted, and we would conclude that the change observed occurred by chance and was not specifically caused by the youth employment programme.

A confidence level of 95 per cent also implies that if the new employment policy is repeated in similar settings 100 times, in ninety five cases it *would produce* an effect that would be either less than or equal to 1.96 times the standard deviation of the outcome of interest before the

[1] An identification strategy in impact evaluations is the strategy designed to identify the effects caused by an intervention or policy separately from any other factors.

intervention. That is to say, in a maximum of five cases the new employment policy would produce an effect that would be larger than 1.96 times the standard deviation of the outcome of interest before the intervention. In conclusion, if the confidence level is set as large as 95 per cent, the null hypothesis would be rejected in a maximum of five cases out of 100, which is a 5 per cent significance level.

Another important term in statistics is the **p-value**. As indicated by Goodman (2008), the p-value is a measure of statistical evidence. It is defined as the probability of the observed result, or a more extreme result, if the null hypothesis were true. In algebraic notation, it is expressed as Prob ($X \leq x \mid H_0$), where X is a random variable corresponding to some way of summarising data (such as a mean or proportion), and x is the observed value of that summary in the empirical data. This is shown graphically in Figures 1 and 2. The curve on Figure 1 represents the probability of every observed outcome under the null hypothesis. The p-value is the probability of the observed outcome (x) plus all ‘more extreme’ outcomes, represented by the shaded ‘tail area’ (Goodman 2008).

Figure 1: Graphical depiction of the definition of a one-sided p-value

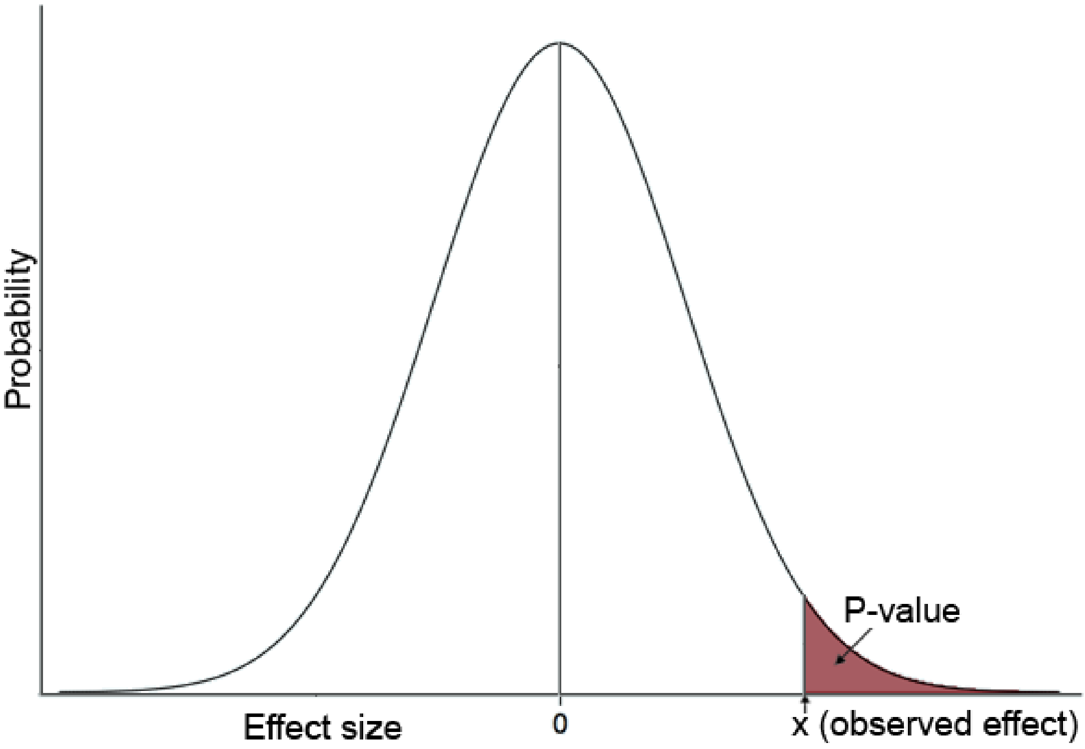
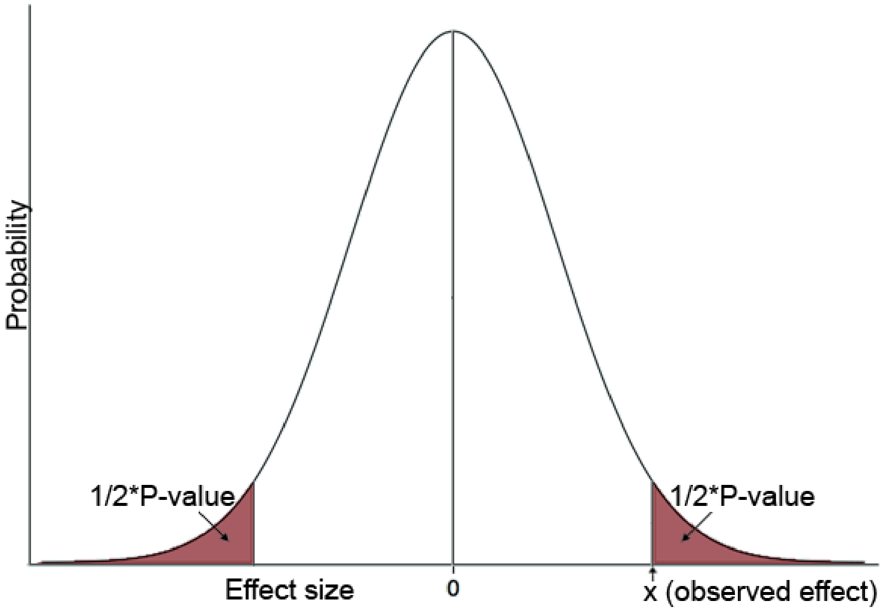


Figure 2: Graphical depiction of the definition of a two-sided p-value (adapted from Goodman 2008)



As discussed previously, the significance level (α) is set as the basis for whether the null hypothesis will be rejected or not. The p-value allows us to make the same decision but it is *calculated,* not set, on the basis of actual data collected from the study.

The nuance between the significance level and the p-value resides in the fact that the significance level is set or decided by researchers based on reasoning and the desired confidence interval. But the p-value is calculated based on actual data collected from the study and gives the actual confidence interval of the findings.

Hence the decision to accept or reject the null hypothesis is based on comparing the p-value with the significance level. When the p-value is smaller than the significance level, the null hypothesis is rejected. When the p-value is larger than the significance level, the null hypothesis is accepted.

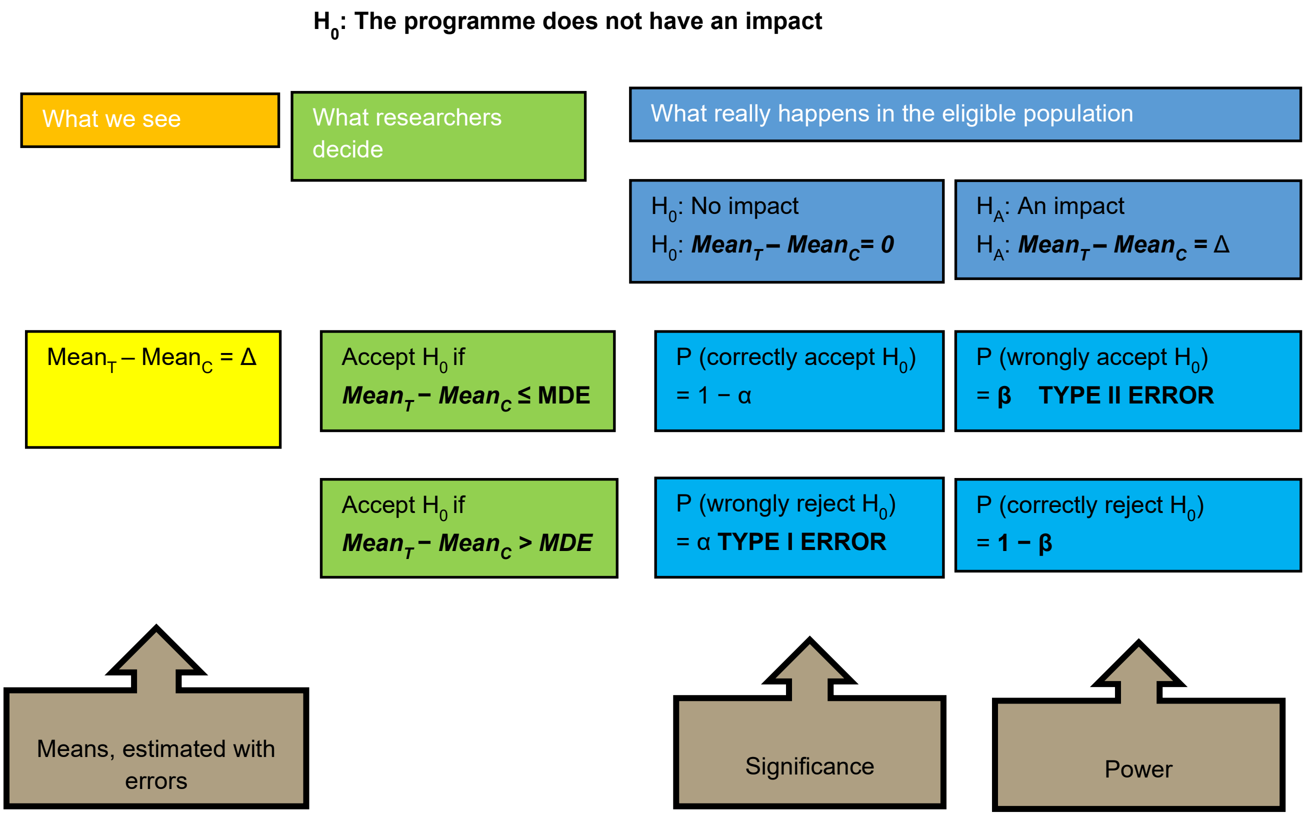
A researcher can make two types of error when deciding whether to accept or reject the null hypothesis: either H_0 is wrongly rejected (**type I error**) or it is wrongly accepted (**type II error**).

By making a type I error, the researcher states a false positive, concluding that an effect or relationship does exist and does not occur just by chance when, in reality, the observed effect took place only by chance.

On the other hand, by making a type II error, the researcher states a false negative, concluding that any effect observed is due to chance and therefore there is no true effect of the intervention, but in reality the intervention does cause an effect that cannot be attributed to chance.

With respect to symbols, note that α, indicates the significance level, and it denotes the probability of making a type I error; the probability of making a type II error is denoted by beta (β). In the same line, the probability of correctly rejecting H_0 is denoted (1 − β) and is called **power***.* These concepts are summarised in Figure 3.

Figure 3: Power: from what we know to what we decide and what happens in reality



H_0: The programme does not have an impact
What we see
What researchers decide
What really happens in the eligible population
H_0: No impact
H_0: $Mean_T - Mean_C = 0$
H_A: An impact
H_A: $Mean_T - Mean_C = \Delta$
$Mean_T - Mean_C = \Delta$
Accept H_0 if
$Mean_T - Mean_C \leq$ MDE
P (correctly accept H_0)
= 1 − α
P (wrongly accept H_0)
= β TYPE II ERROR
Accept H_0 if
$Mean_T - Mean_C > MDE$
P (wrongly reject H_0)
= α TYPE I ERROR
P (correctly reject H_0)
= 1 − β
Means, estimated with errors
Significance
Power

3. Power calculation: concept and applications

Power analysis and power calculation are not exactly the same, even though the terms are used interchangeably in some circumstances.

Power calculation is the determination of the minimum sample size required to detect (that is, to find statistically significant) a minimum effect, which is set *ex ante*, given the specific parameters set for the rest of the study, such as power, significance level and sampling approach. Power calculation is a straightforward process – based on mathematical formulae – when all parameters are known.

Power analysis is the decision-making process about sample size, given real-world constraints, including budget, time, accessibility of samples of interest, distance, surveyor safety concerns, politics and the policy-relevant magnitudes of effects in policy-relevant time frames. Power analysis tries to optimise sample size and power against these constraints. Though power calculation is an *ex ante* process, there are instances when *ex post* power calculations can be performed to assess whether the lack of impact is due to the sample size, inherent to the intervention itself, or due to some other reason such as implementation failure or contamination.

Based on Gleason (2010), power analysis addresses three main questions:

1. How likely is it that a particular design will detect an impact or effect size that the intervention being examined is likely to produce in the study time frame?
2. Given the research design, how large does the true impact of the intervention need to be in order to detect it? In other words, what is the MDE for a given design?
3. Given the true impact that an intervention is likely to produce, how large does a sample have to be in order to detect it?

Altogether, power *analysis* addresses the following broad question: accounting for the true size of the effect an intervention is likely to produce in a given time frame, and given all of the constraints, what is the most efficient sample size in which a meaningful effect can be detected?

3.1 Parameters required to run power calculations

Power calculations depend on a number of measurable parameters that are important to clarify during the research design phase. Some of these parameters are discretionary (chosen by the evaluator and under their control) and others are inherent (chosen by the evaluator but not under their control). Although the choice of some parameters relies on rules of thumb, it is advisable to discuss these with a statistician and justify the choice of each parameter. This is especially important for inherent parameters (see also Section 3.1.3).

3.1.1 Definition of power calculation parameters

This subsection starts with a brief review of hypothesis testing, followed by the definition of discretionary and inherent parameters.

Hypothesis testing is the use of statistical methods to determine the probability that a specific hypothesis is true or not. The first step of hypothesis testing consists of defining the null hypothesis and the alternative hypothesis.

In impact evaluation, the null hypothesis states that there is no difference between mean of the outcome variable of the treatment group and the mean of the same outcome variable in the control group that should be attributed to anything other than chance. That is, there is no causal effect of the programme. The alternative hypothesis states that there is a difference between the two groups. As it is not possible to observe the effect of an intervention before it takes place, it is therefore not possible to test the alternative hypothesis. Hence researchers test the null hypothesis by default (including for impact evaluation studies).

3.1.2 Discretionary parameters

a. ***MDE:*** Defined by Bloom (1995) as the smallest effect that, if true, has an X per cent chance of producing an impact estimate that is statistically significant at the Y level; where X is the statistical power of the experiment and Y is the level of statistical significance.

i. *Significance level,* Y above, but generally denoted α, is the probability of concluding that there is an impact of the intervention when actually there is no impact. This is known as the probability of making a **type I error**. It is commonly set between 90 percent and 95 percent in social sciences.

ii. *Statistical power*, X above, but formally expressed as (1 − β), is the probability of correctly concluding that an intervention has no statistically significant effect. In other words, it is the probability of not committing a type II error. It is commonly set as 80% or 90% in social sciences.

Note that MDE should not be an *ad hoc* choice. Rather, it should be a choice informed by cost-efficiency considerations, policy and political considerations, time constraints, context, existing theory, models and empirical evidence.

It is essential to realise that an MDE is itself a function of time, as the effect of an intervention would generally vary (increase, decrease or plateau) over time. Therefore, in the process of defining or setting an MDE, we need to have a good idea of the effect trajectory as a function of time: how would the effect vary throughout time? What is the trajectory of the effect measured? Is it a short-term, medium-term or long-term effect? After how much time would the effect be expected to take place? These crucial questions inform the decision about when to conduct endline data collection to measure the MDE that the intervention is expected to have.

Type I error arises when we mistakenly reject (fail to accept) the null hypothesis. That is, we conclude that an effect or relationship exists and does not occur just by chance, while in reality, such an effect does not exist. In practical terms, this would be deciding that an ineffective treatment actually works, thereby wasting resources.

b. *Type II error* arises when we mistakenly accept (fail to reject) the null hypothesis. That is, it is the probability of wrongly concluding that there is no impact of an intervention beyond what would occur due to randomness. The probability of making a type II error is often denoted β and typically set as 10% or 20% in social science. In practical terms, this would be deciding that a treatment does not have a causal impact when it actually does and therefore potentially discontinuing a useful treatment.

c. ***Statistical tests*** can be performed as one-sided or two-sided: specifying which will be used is important for power calculations. One-sided tests require smaller sample sizes compared with two-sided tests. Therefore, the use of the former should always be justified based on prior knowledge about the expected effect, and not merely for the sake of reducing the sample size required to detect the expected effect.

i. ***One-sided statistical tests*** are used when the alternative hypothesis is expected to be unidirectional. That is, the researcher can convincingly argue that the intervention is expected to either raise *or* lower the value or occurrence of the outcome of interest. A unidirectional hypothesis test is chosen when the researchers can predict in which way (positive or negative) the intervention is expected to have an effect.

ii. ***Two-sided statistical tests*** are used when the alternative hypothesis is non-directional. In other words, such tests are used when researchers cannot predict whether the intervention will have a positive or negative impact, but they expect that there will be an impact.

d. The proportion of the study sample that is randomly assigned to the treatment group: Though optimisation of statistical power suggests 50/50 treatment/control group allocation (Bloom 1995), there are conflicted cases in which there is a trade-off between optimising statistical power and following a policymaker's suggestion to keep the number of participants assigned to the control group to a minimum. In such a trade-off, the resulting proportion of the sample assigned to treatment and control may be unbalanced (60/40; 70/30; 80/20) at the expense of statistical power but for the sake of budget and political constraints.

3.1.3 Inherent parameters

a. Mean*:* In this case, the arithmetic mean of a sample, $x_1, x_2, \ldots, x_n$, is the sum of the sampled values divided by the number of items in the sample.

b. Standard deviation of the mean (standard error) is used for power calculations, calculated for the main outcome variable in the absence of the intervention.

c. Intra-cluster correlation *(ICC)* is a measure used when sampling is based on clusters to capture the relatedness of data collected within a cluster. This is done by comparing the variance *within* clusters with the variance *between* clusters.

d. Coefficient of variation (R^2) is the proportion of the variance in outcome that is explained by the explanatory variables included in the prediction model.

3.1.4 Definition of other factors affecting power calculation: attrition, compliance, take-up

a. **Attrition** refers to a reduction of the initial sample size involved in a study, which means that outcome data cannot be collected for all units of observation. Attrition may occur as a result of migration of some participants from the study area, their refusal to continue participating in the research, their death, etc. Attrition affects the magnitude of statistical power as it decreases the sample size. It may also affect internal validity, if attrition is not randomly distributed across the treatment and control groups. It is always better to anticipate the potential attrition rate during the evaluation preparation phase and account for it through prudent oversampling at the beginning of the study. It is also advisable to account for attrition by oversampling in any particular subgroup (whether treatment or control) according to that subgroup's likelihood of attrition.

For example, a maternal and neonatal nutritional programme in Benin provided multivitamin tablets and education about infant nutrition to pregnant women, to improve the nutritional health status of newborns. Eligible women were required to visit health centres to access the benefits of the programme. The Ministry of Health commissioned an impact evaluation that required 75 treatment villages and 75 control villages. During baseline data collection, 750 women took part in the survey. However, only 680 women took part in the final data collection. In this case, the initial sample of 750 women shrunk to 680 women, because 70 women could not be surveyed as the result of attrition.

b. **Compliance** refers to the units of observation in a study sample obeying by the treatment (or no treatment) status assigned to them. Some people in the treatment group will not actually use the treatment, while some in the control group will try to access the treatment. For example, a maternal and neonatal nutrition programme in Benin provided multivitamin tablets and education about infant nutrition to pregnant women, to improve the nutritional status of newborns. Eligible women were required to visit health centres to access the benefits of the programme. The Ministry of Health commissioned an impact evaluation that required 75 treatment villages and 75 control villages. However, in the course of the programme implementation, it appeared that not all the women with children aged under five in the treatment villages visited the health centres. Therefore, only some women received the intervention. In consequence, there is only partial compliance, which has an implication for take-up of the programme in the treatment villages (which will be discussed below).

A critical consequence of non-compliance to an assignment group is that estimates of the true treatment effect may be biased. In the example above, if some women in the treatment group did not access the programme, while others in the control villages did receive the intervention, outcome measures in both groups would not depict the true effect of actual assignment status (treatment or control). Instead, it would show a confounded effect, which would yield biased estimates with low power because the study would be less likely to conclude correctly that the intervention had no statistically significant effect. The power decreases because the real sample size of the study reduces, and the standard error of the estimation of the nutrition status of new-borns increases.

c. **Take-up** is a measure of compliance in the treatment group, but in relative terms. In general, take-up is expressed as the percentage of those who use or access the treatment among people who were offered the treatment. Take-up is a critical parameter in power calculations, as a low take-up decreases the sample that is actually treated. If this low take-up is not accounted for when determining the sample size, it can reduce the expected power magnitude because the standard error of the outcome measured would increase. Let's consider a programme to access healthcare that was offered to 750 women and only 630 claimed the benefits of the programme. In that case, take-up was 84 per cent [(630 / 750) * 100].

It may seem that compliance and take-up refer to the same notion but they are different. Compliance is when each unit of observation follows and does what is required according to his or her assignment status. Take-up is used to indicate the use of the treatment among beneficiaries.

3.2 Statistical power and sample size determination

As previously indicated, statistical power is defined as the probability of correctly rejecting the null hypothesis. More specifically, power is calculated based on the primary outcome variables (first-order outcome variables) and not necessarily on secondary outcome variables.

Considering the example of the youth employment scheme, let's assume that the evaluation team set 80 per cent power for the evaluation findings, which implies that they have an 80 per cent chance of correctly concluding that an intervention has no statistically different effect from what would have happened in the absence of the scheme. If the research team concludes that the new youth employment policy has significantly increased the employment rate by 20 per cent, which is the MDE, it implies that in 80 per cent of similar but independent contexts where the government might implement the policy, it is likely that the change in employment rate would be at least the MDE.

Technically, power calculations entail estimating and evaluating the minimum effect that is detectable at a designated level of statistical significance and power. As expressed in Bloom (1995), most of the definitions of power calculations refer to the minimum sample size required to detect the smallest effect that, if true, has an X per cent chance of producing an impact estimate that is statistically significant at the Y; where X is the statistical power of the experiment and Y is the level of statistical significance.

Three approaches may be considered in deciding the sample size required for a study: MDE approach, power determination approach and sample size determination approach.

The **MDE approach** consists of knowing the desired level of power, the sample size that the study can afford, and the minimum effect that can be detected for a given sample size. In most cases, the sample size is adjusted to reach the expected MDE.

Based on the **power determination approach**, researchers determine the power that a study would have considering an expected MDE, and assuming a hypothetical sample size. Using this approach, the sample size would be adjusted so that the expected power is reached.

The **sample size determination approach** requires clearly setting the power beforehand, as well as the MDE that the intervention is expected to have. This approach is a straightforward way to determine the minimum sample size required to meet power level and MDE set for the study.

In this paper, we use the MDE approach and the sample size determination approach, as they seem to be the focus of most researchers and they are more easily understood by policymakers in general.

Power calculations using the MDE approach use the following the formulae:

$$MDE = \left(t_{1-{\alpha}/_{2}} + t_{1-\beta}\right) e$$

MDE Minimum detectable effect

${\alpha}/_{2}$ Rate of type I errors (false positives)
(Typically in social science, ${\alpha}/_{2}$ = 2.5%)

β Rate of type II errors (false negatives)
(Typically in social science, β = 10 to 20% which translates to power = 80 to 90%)

e Standard error of the estimated effect

Looking at the formula, it can be seen that whatever design is used in a study, MDE is a function of t-values and standard error of the estimated effect. These t-values are straightforward and quite easy to find in most statistics or econometrics books. Table 1 provides values of different levels of statistical power and statistical significance. However, the most critical, sensitive and sometimes complex data to obtain, in order to run a power calculation, are standard errors of estimated effect of the intervention on the outcome. Standard error is determined by sample size, sampling approach (simple random sampling or multiple stage random sampling) and variance of the estimated effect.

Table 1: MDE as a multiple of the standard error of the impact estimate for different levels of statistical power and statistical significance

	Significance level		
Statistical power	*.10*	*.05*	*.01*
One-sided hypothesis test			
90%	2.56	2.93	3.61
80%	2.12	2.49	3.17
70%	1.80	2.17	2.85
Two-sided hypothesis test			
90%	2.93	3.24	3.86
80%	2.49	2.80	3.42
70%	2.17	2.48	3.10

Source: Bloom (1995)

Considering a simple random sampling, the standard error is estimated as illustrated here:

$$e = \sqrt{\frac{2\sigma^2}{n}}$$

e	Standard error of the outcome
n	Sample size of each group
σ^2	Variance of the outcome – for a prevalence, σ^2 = P(1−P)

This assumes treatment and control groups of the same size, the same variance and selected using a simple random sampling approach.

However, when the sampling approach is, for example, a two-stage random sampling, it is critical to account for design effect, due to ICC that increases the standard error of outcomes of interest and would consequently increase the minimum effect that the design could detect. Therefore, the larger the standard error, the lower the statistical power. This is illustrated in Figures 4, 5 and 6 which were adapted from Muñoz (2013).The green area depicts the statistical power and becomes smaller and smaller (from Figure 4 to 5), as the standard error becomes larger and larger. Figure 7 combines all key parameters (α, type I error, type II error) to depict the concept of statistical power.

Figure 4: Depiction of power (in green) with large sample size

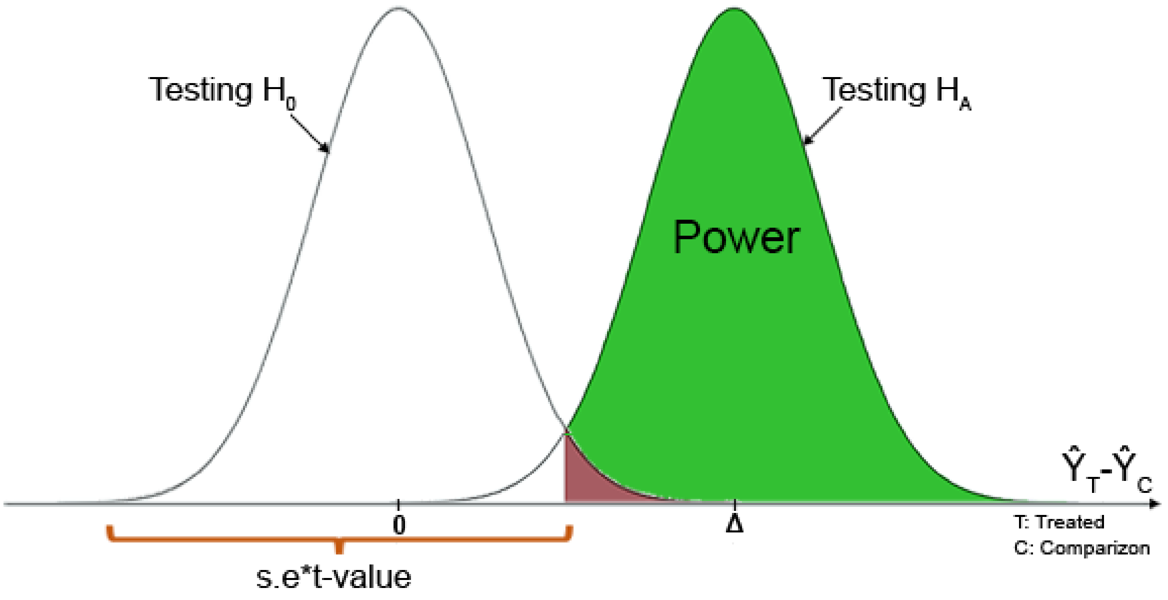


Figure 5: Depiction of power (in green) with reduced sample size vis-à-vis Figure 4

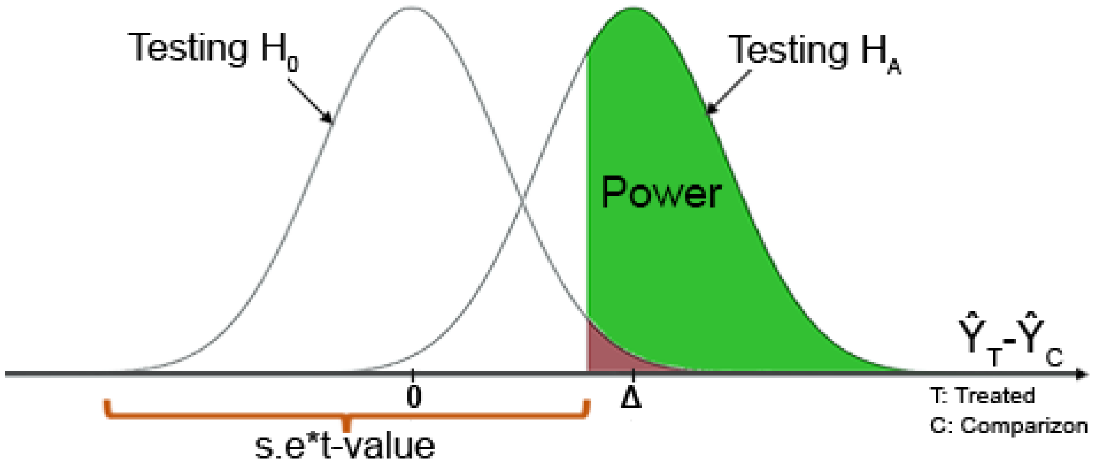


Figure 6: Depiction of power (in green) with much smaller sample size vis-à-vis Figures 4 and 5

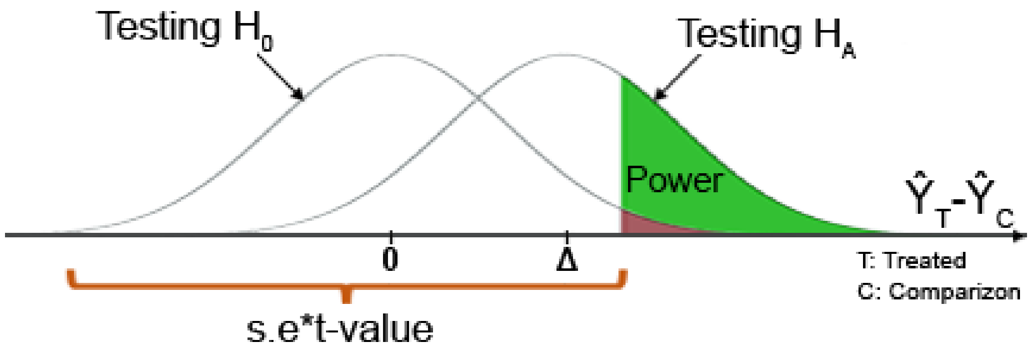
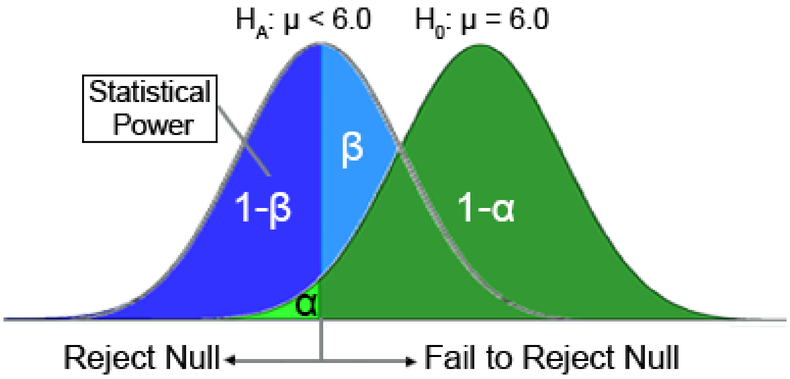


Figure 7: Graphical visualisation of critical statistical power and other key parameters



In mathematical terms, standard errors increase if, instead of taking a simple random sample of *n* households, we take a two-stage sample, with *k* primary sampling units and *m* households per primary sampling unit ($n = k * m$), as depicted in the formula:

$$e_{TSS}^2 = e_{SRS}^2[\overbrace{1 + \underset{\substack{\uparrow \\ \boxed{\text{Intra-cluster Correlation}}}}{\rho}(m - 1)}^{\text{Cluster effect}}]$$

TSS: Two-stage sample
SRS: Simple Random Sample

3.3 How to run power calculation: single treatment or multiple treatments?

A power calculation involves determining the sample size required to test a null hypothesis (H_0) with sufficiently large power (minimum of 80 per cent in social science). That means a researcher should always first list all the primary null hypotheses (based on primary outcomes) that a study plans to test, and then determine the sample size required for each. The largest sample size from this list should be considered for the study (that is, it is the preferred sample size without taking into account other factors such as budget and timing). The logic behind power calculation is the same whether a study has one, two, or multiple treatment arms, but the calculation does need to account for the different treatment arms.

For example, in a multiple treatments evaluation design, a research team is planning to evaluate the most effective approach to increase girls' enrolment, attendance and educational performance in secondary school. Three approaches are selected by the government: supplying a free midday meal to schools, granting a bursary to all schoolgirls, and building toilets and sanitation infrastructure in schools for girls. There are four treatment groups in the study:

schools with no treatment (status quo), schools that supply a free midday meal, schools that grant a bursary to girls, and schools equipped with toilets and sanitation infrastructure. The sample size determination will consist of listing the primary null hypotheses, calculating the required sample size in each case and selecting the largest sample size calculated to conduct the study.

With respect to **midday meal supply**, the following are examples of primary null hypotheses that can be tested:

H_0 a: Supplying midday meals to schools has no effect on school enrolment for girls compared with the effect of no treatment.
H_0 b: Supplying midday meals to schools has no effect on girls' attendance rates compared with the effect of no treatment.
H_0 c: Supplying midday meals to schools has no effect on girls' educational performance compared with the effect of no treatment.

With respect to **granting bursaries to girls**, the following primary null hypotheses can be tested:

H_0 d: Granting a bursary to school girls has no effect on school enrolment for girls compared with the effect of no treatment.
H_0 e: Granting a bursary to school girls has no effect on girls' attendance rates compared with the effect of no treatment.
H_0 f: Granting a bursary to school girls has no effect on girls' educational performance compared with the effect of no treatment.

With respect to **building toilets and sanitation infrastructure**, the following primary null hypotheses can be tested:

H_0 g: Building toilets and sanitation infrastructure in schools has no effect on school enrolment for girls compared with the effect of no treatment.
H_0 h: Building toilets and sanitation infrastructure in schools has no effect on girls' attendance rates compared with the effect of no treatment.
H_0 i: Building toilets and sanitation infrastructure in schools has no effect on girls' educational performance compared with the effect of no treatment.

With respect to the comparison between the **effectiveness of midday meal supply and bursary policy**, the following null hypotheses can be tested:

H_0 j: Granting a bursary to school girls has no effect on girls' school enrolment compared with the effect of supplying midday meals to schools.
H_0 k: Granting a bursary to school girls has no effect on girls' attendance rates compared with the effect of supplying midday meals to schools.
H_0 l: Granting a bursary to school girls has no effect on girls' educational performance compared with the effect of supplying midday meals to schools.

4. Rules of thumb for power calculation

Even though power calculation is a technical task, it is also true that there are **a number of rules of thumb** that are applied and can always serve as guidance:

- When power increases, the probability of finding a true impact of the intervention (if it exists) increases. In social science, researchers aim to have at least 80 per cent power which means allowing 20 per cent chance of committing a type II error.
- The larger the sample size, the smaller the standard error, and therefore the higher the power.
- The smaller the MDE, the larger the required sample size.
- For any given number of clusters, the larger the ICC, the lower the power.
- For any given number of units of observation per cluster, the larger the number of clusters, the higher the power.
- Increasing the units of observation per clusters will generally not improve power as much as increasing the number of clusters (unless ICC is zero).
- ICC increases when observations within clusters are increasingly identical relative to other clusters, which lowers the number of independent observations and, effectively, the sample size.
- Baseline covariates are used in model specification to increase the statistical power of the study because they reduce the standard error of outcome, and therefore increase the likelihood of reducing the minimum effect that the design can detect.

5. Common pitfalls in power calculation

This section indicates pitfalls that are commonly reported or pointed to with respect to power calculation in social science. Formulae presented in this manual are mainly limited to simple random sampling and two-level random sampling. For more complex designs, such as three-level cluster design, formulae can be found in the literature and in the documentation of specialised software such as Optimal Design, G*Power or Stata. These software packages are useful for running power calculations; however, it is critical to read carefully the assumptions and other considerations behind each formula or command used in these packages. The researcher will need to tell the software whether you are working with binary or continuous outcomes and whether you are using simple random, multistage or stratified/blocked sampling. For example, the commands 'sampsi' and 'sampclus' are not interchangeable in Stata and do not yield the same results. To continue with common pitfalls, it may be useful to be aware of the following pitfalls:

1. Sample size should be determined for all main outcome variables before a final decision on study sample size is made. For instance, in the case of the secondary school intervention mentioned in the section 3.3, it is not appropriate to run a power calculation only for school attendance when learning outcomes are also a main outcome of interest.

2. The MDE of an intervention is a function of the impact trajectory of the intervention over time. Therefore, it is essential to take into account the expected timeline of the intervention before deciding the magnitude of the MDE.

3. Power calculations must account for ICC in the case of cluster sampling because ICC varies with sampling methods and therefore it influences the standard error that in return affects power. In other words, not all samples of the same size have equal power.

4. If a study does not detect a statistically significant effect of an intervention, it does not necessarily mean that the study is underpowered. It may be because the intervention fails to deliver according to plan (implementation failure), or it is not the right intervention for the problem at hand. Do not blame lack of power for all statistically insignificant results.

5. Attrition is a major threat to evaluation, because it decreases the size of the sample for which there is full information, and therefore reduces power. There is no genuine way to rectify sample size after attrition occurs. To minimise attrition, it is necessary to collect enough data to be able to track participants. To avoid the effects of expected attrition, it is advisable to oversample or take all necessary measures (without compromising the intervention) to avoid or limit attrition.

6. Spillover and contamination are other ‘ghosts’ that bias estimates and therefore affect power. Spillover is when the control group is affected by the intervention through a different mechanism. Contamination is when the treatment or control groups are affected by a similar intervention during the study, which will bias the attributable effect estimates for the intervention studied. Study design and programme implementation should guard against spillover and contamination.

7. Power calculation is run to decide on the sample size required for an evaluation study. It is an *ex ante* activity and not an *ex post* decision. When such calculations are run *ex post*, they can be used to check actual power but the purpose is completely different from that of power calculation. In an *ex post* power check, the objective is to determine the power of the study, given the actual sample size used for analysis but using the same values for all other parameters used while running *ex ante* power calculation.

8. Using a randomised controlled trial as the identification strategy does not alone guarantee that power will be sufficient.

9. Power calculation formulae or programming are not the same for continuous versus binary outcome measures. It is a mistake to use the same formula in each case. Even when using software packages, it is critical to specify the nature (continuous or binary) of the outcome variable of interest.

6. Power calculations in the presence of multiple outcome variables

When testing the effect of an intervention on multiple outcome variables, it is probable that the researcher will find purely by chance a statistically significant effect of the intervention on a few outcomes. In fact, a significance level of 5 per cent means that the chance of erroneously finding a statistically significant impact is 5 per cent. For example, if you test the effect of your intervention on 20 different outcomes, and for all of them the null hypothesis is actually true, you'd expect about one of the tests to be significant at the $p<0.05$ level, just due to chance. Another way to present this is to calculate the probability of observing at least one significant result by chance when you have 20 hypotheses to test, and a significance level of 0.05. In this configuration:

$$\begin{aligned} P\text{ (at least one significant result)} &= 1 - P\text{ (no significant result)} \\ &= 1 - (1 - 0.05)^{20} \\ &\approx 0.64 \end{aligned}$$

Consequently, with 20 hypotheses to test, we have a 64 per cent chance of observing at least one significant result, even when the intervention has no significant impact on any of 20 outcomes.

Thus, when assessing the impact of an intervention on multiple outcomes, the idea of testing the impact of an intervention on each outcome using the standard value of significance level (0.05) may lead to erroneously finding a statistically significant impact when in fact there is no impact.

Methods for assessing the impact of an intervention on multiple outcomes call for adjusting the value of the significance level, so that the probability of observing at least one significant result by pure chance remains low or below the desired significance level.

The most popular method to estimate the value of the significance level when testing the impact of an intervention on multiple outcome variables is the Bonferroni correction. The Bonferroni correction sets the significance level cut-off at α/n, where α is the standard significance level (0.05) and n is the number of outcomes. For instance, in the example above, to assess the impact of the intervention on 20 outcomes, and with the standard significance level of 0.05, the null hypothesis will be rejected only if the p-value is less than 0.0025 (0.05/20). Thus, for a study testing the impact of an intervention on 20 outcomes, the significance level that should be used is 0.0025.

Given that the significance level is among the parameters used to perform power calculations, a change of the significance level will affect statistical power if other parameters are kept constant. The ultimate implication is that power calculations in the presence of multiple outcome variables will be affected and will be different from power calculations when assessing the impact of the intervention on one outcome variable.

With the example above, and with a significance level equal to 0.0025 and keeping other parameters used to perform power calculations constant, the study will have a reduced statistical power. In addition to the Bonferroni correction, there are other methods for adjusting the significance level to take into account that the study will evaluate the impact of interventions with multiple outcomes. In general, these adjustments result in a study with reduced statistical power if other parameters are kept constant.[2]

We recommend that researchers adjust for the significance level when performing power calculations in the presence of multiple outcome variables. Researchers can use the Bonferroni correction to calculate the adjusted significance level to be used for power calculations because it is quite a straightforward method. However, the Bonferroni correction yields the most conservative value of the significance level among the correction methods (Schochet 2008). Consequently, the Bonferroni method leads to the most reduced statistical power and reduces the probability of rejecting the null hypothesis when it is false. Given the drawback of the Bonferroni method, we recommend that researchers consider their main and most critical outcome variables when they intend to use the Bonferroni correction method.

7. Experimental design

A Microsoft Excel® worksheet has been developed as an online supplement for this section. Users can enter values for different parameters to determine either the minimum sample size required or the MDE, depending on the design used. Results of examples used in this section are calculated using this online tool, the *3ie Sample size and minimum detectable effect calculator*©.

7.1 Individual-level randomisation

In this section, we present formulae to calculate expected MDE as a function of sample size (among other parameters) for study designs where the intervention is randomly assigned to individual units of observation. In such a case, the unit of observation is the same as the unit of assignment. In addition to accounting for assignment approach, we account for the characteristic of the outcome of interest (continuous or binary) as this determines which power calculation formula to use.

7.1.1 Single-level trials with a continuous outcome variable

For a continuous outcome variable when the intervention is assigned to individual units, the following formulae are used.

[2] Schochet (2008) provides a very good description of different methods for adjusting the significance level when performing multiple testing and limitations of each of them.

MDE:

$$\delta = (t_{1+}t_2)\sigma_y\sqrt{\frac{1}{P(1-P)n}}$$

Sample size:

$$n = \left\{\frac{1}{P\delta^2}\sigma_y^2\frac{(t_1+t_2)^2}{-P+1}\right\}$$

Table 2: Parameters required for single-level trials with a continuous outcome variable

δ	Minimum detectable effect
α	Desired significance level
β	Desired power of the design
t_1	t-value corresponding to the desired significance level of the test*
t_2	t-value corresponding to the desired power of the design*
σ_y	Standard deviation of the outcome variable
P	Proportion of the study that is randomly assigned to the treatment group
n	Sample size

Note: * t-values are a function of specific sample size. Taking into account the central limit theorem and the law of large numbers, a minimum sample size assumption needs to be made to calculate precise t-values. Otherwise, users may purposively decide the value for t_1 and t_2.

(1) t_1 is the student's t-critical value of the desired significance level of the test
(2) t_2 is the student's t-critical value of the desired power of the design
(3) $t_{1-\alpha/2}$ is the student's t-critical value of 1 − α/2
(4) $t_{1-\beta}$ is the student's t-critical value of 1 − β

Example

A research team is planning to do an experiment to determine if an apprenticeship programme increases annual earnings of youth in Kisumu, Kenya. Due to budget constraints, the study can only afford 1,000 participants. Fifty per cent will be assigned to the treatment group and the rest will be assigned to the control group. The research team is uncertain of the direction of the impact, so they opt to use a two-sided hypothesis test. They set the significance level at 0.05 and decide to use 80 per cent statistical power.

This example has the following parameter values: t_1 = 1.96, t_2 = 0.84, σ_y = 2,400 Kenyan shillings, p = 0.5 and n = 1,000, which yields MDE = 425.7 Kenyan shillings.

7.1.2 Single-level trials (continuous outcomes) with covariates

In this scenario, the same parameters hold but baseline covariates (explanatory variables) are also used, which can increase the statistical power of the study.

MDE:

$$\delta = (t_{1+}t_2)\sigma_y\sqrt{\left(\left(\frac{1}{P(1-P)n}\right)(1-R^2)\right)}$$

Sample size:

$$n = \left\{\frac{1}{P\delta^2}\sigma_y^2\frac{(t_1+t_2)^2}{-P+1}\{-R^2+1\}\right\}$$

Table 3: Parameters required for single-level trials (continuous outcomes) with covariates

δ	Minimum detectable effect
α	Desired significance level
β	Desired power of the design
t_1	t-value corresponding to the desired significance level of the test*
t_2	t-value corresponding to the desired power of the design*
σ_y	Standard deviation of the outcome variable
P	Proportion of individuals randomly assigned to the treatment group
n	Sample size
R^2	Proportion of outcome variance explained by level 1 covariate(s)[3]

Note: * t-values are a function of specific sample size. Taking into account central limit theorem and the law of large numbers, a minimum sample size assumption needs to be made to calculate t-values. Otherwise, users may purposively decide the value for t_1 and t_2.

Example

A research team is planning to do an experiment to determine if an apprenticeship programme increases annual earnings of youth in Kisumu, Kenya. Due to budget constraints, the study can only afford 1,000 participants. Fifty per cent are assigned to the treatment and the rest will be assigned to the control group. The research team is planning to do the same experiment described above, but this time they have auxiliary data, in the form of data from a similar study in Kenya. In the previous study, the endline annual earnings regressed on baseline value of education gave an R^2 equal to 0.5 – that is, baseline education accounts for half of the variance in the outcome. Using a two-sided hypothesis test at 0.05 significance level, the research team will calculate the MDE for a study aiming at 80 per cent statistical power.

In this example, $t_1 = 1.96$, $t_2 = 0.84$, σ_y = 2,400 Kenyan shillings, $p = 0.5$, n = 1,000 and $R^2 = 0.5$, which yields MDE = 301 Kenyan shillings.

[3] In this manual, the proportion of outcome variance by the level 1 covariate(s) baseline is the squared correlation coefficient between the baseline measure of covariate(s) and the post-implementation measure of the outcome, which is also known as the coefficient of determination.

7.1.3 Single-level trials with binary outcomes

The following formulae are used when the outcome variable is a binary variable.

MDE:

$$\delta = (\mathrm{t}_{1+}\mathrm{t}_2)\sqrt{\frac{\mathrm{P}(1-\mathrm{P})}{\mathrm{T}(1-\mathrm{T})\mathrm{n}}}$$

Sample size:

$$n = \left\{\frac{P}{T\delta^2}\frac{-P+1}{-T+1}(-t_1 - t_2)^2\right\}$$

Table 4: Parameters required for single-level trials with binary outcomes

δ	Minimum detectable effect
α	Desired significance level
β	Desired power of the design
t_1	t-value corresponding to the desired significance level of the test*
t_2	t-value corresponding to the desired power of the design*
P	Proportion of the study population that have a value of 1 for the outcome in the absence of the programme
T	Proportion of individuals randomly assigned to the treatment group
n	Sample size

Note: * t-values are a function of specific sample size. Taking into account central limit theorem and the law of large numbers, a minimum sample size assumption needs to be made to calculate t-values precisely. Otherwise, users may purposively decide the value for t_1 and t_2.

Example

In order to assess whether compensating for transportation costs and loss in wages increases the uptake of male circumcision in Zambia, a research team is planning to conduct a randomised controlled trial in Makululu District. The study plans to provide food and transportation vouchers as an incentive to the uncircumcised male population aged 15–49 years. The research team can only afford a sample size of 1,000 individuals in total. Fifty per cent will be randomly assigned to the treatment group and the other 50 per cent to the control group. Following a pilot phase, the team found that 3 per cent of individuals in the target population were already circumcised before the intervention. Using a one-side hypothesis test at the 0.05 significance level, the research team will calculate the MDE for a study with 80 per cent statistical power.

In this example, $t_1 = 1.65$, $t_2 = 0.84$, $P = 0.03$, $T = 0.5$, $n = 1,000$, which yields MDE = 0.027.

7.1.4 Single-level trials (binary outcomes) with covariates

The following formulae are used when the outcome variable is a binary variable and baseline covariates are used to increase the statistical power of the study.

MDE:

$$\delta = (\mathrm{t}_{1+}\mathrm{t}_2)\sqrt{\left(\frac{\mathrm{P}(1-\mathrm{P})}{\mathrm{T}(1-\mathrm{T})\mathrm{n}}(1-\mathrm{R}^2)\right)}$$

Sample size:

$$n = \left\{\frac{P}{T\delta^2}\frac{-P+1}{-T+1}(-t_1-t_2)^2(-R^2+1)\right\}$$

Table 5: Parameters required for single-level trials (binary outcomes) with covariates

δ	Minimum detectable effect
α	Desired significance level
β	Desired power of the design
t_1	t-value corresponding to the desired significance level of the test*
t_2	t-value corresponding to the desired power of the design*
P	Proportion of the study population that have a value of 1 for the outcome in the absence of the programme
T	Proportion of individuals randomly assigned to the treatment group
n	Sample size
R^2	Proportion of outcome variance explained by level 1 covariate(s)

Note: * t-values are a function of specific sample size. Taking into account central limit theorem and the law of large numbers, a minimum sample size assumption needs to be made to calculate precise t-values. Otherwise, users may purposively decide the value for t_1 and t_2.

Example

In order to assess whether compensating for transportation costs and loss in wages increases the uptake of male circumcision in Zambia, a research team is planning to conduct a randomised controlled trial in Makululu District. The study plans to provide food and transportation vouchers as an incentive to the uncircumcised male population aged 15–49 years. The research team can only afford a sample size of 1,000 individuals in total. Fifty per cent will be randomly assigned to the treatment group and the other 50 per cent to the control group. Following a pilot phase, the team found that 3 per cent of individuals in the target population were already circumcised before the intervention. Using data from a similar study in Kenya, the research team noted that a regression of the endline uptake on the ethnic Kikuyu gives an R^2 equal to 0.6. Using a one-sided hypothesis test at the 0.05 significance level, the research team will calculate the MDE with 80 per cent statistical power.

In this example, $t_1 = 1.65$, $t_2 = 0.84$, $P = 0.03$, $T = 0.5$, $n = 991$, $R = 0.6$, which yields MDE = 0.017.

7.1.5 Single-level trials (rates)

The following formulae are used for cases using incidence rates with a person-years denominator, such as the mortality rate or the incidence rate of severe disease.

MDE:

$$\mu = \left\{\frac{-1}{R}\left(-\frac{1}{2}a - R\mu_0 + \frac{1}{2}\sqrt{8Ra\mu_0 + a^2}\right)\right\}$$

Person-year per group:

$$R = \frac{(z_{1+}z_2)^2\,(\mu_{0+}\mu_1)}{(\mu_{0-}\mu_1)^2}$$

Table 6: Parameters required for single-level trials (rates)

R	Person-year in each group
α	Desired significance level
β	Desired power of the design
z_1	z-value corresponding to the desired significance level of the test
z_2	z-value corresponding to the desired power of the design
μ_1	True (population) rate in the presence of the intervention
μ_0	True (population) rate in the absence of the intervention

Example

A research team from Cheikh Anta Diop University has developed a new and promising vaccine against malaria. In order to test whether this vaccine reduces child mortality, the research team is planning to conduct a randomised controlled trial in Touba village in Senegal. Specifically, 50 per cent of the study sample will be randomly assigned to the treatment group (vaccine) and the rest will be assigned to the control group (no vaccine). The trial is expected to last two years. Mortality data for two years prior to the study indicates that there were a total of 1,667 deaths over 23,141 person-years of observation, giving an overall mortality rate of μ_0 = 1,667/23,141 = 0.07203 (72 per 1,000 per year). The research team assumes that the mortality rate in the control group remains constant and the vaccine will reduce mortality by 40 per cent. Using a two-sided hypothesis test at the 0.01 significance level, the research team will calculate person-years of observation in each group for a study with 90 per cent statistical power.

In this example, z_1 = 2.58, z_2 = 1.28, μ_0 = 0.072, μ_1 = 0.0432, which yields 2,067 person-years in each group.

7.2 Cluster-level randomisation

7.2.1 Two-level cluster randomised controlled trials with individual-level outcomes (continuous outcome)

MDE:

$$\delta = \frac{t_1 + t_2}{\sqrt{P(1-P)J}}\sigma_y\sqrt{\rho + \frac{1-\rho}{n}}$$

Number of clusters:

$$J = \left\{\frac{1}{P\delta^2}\sigma_y^2\frac{(t_1+t_2)^2}{-\rho+1}\left(\rho+\frac{1}{n}(-\rho+1)\right)\right\}$$

Table 7: Parameters required for two-level cluster randomised controlled trials with individual-level outcomes (continuous outcome)

δ	Minimum detectable effect
α	Desired significance level
β	Desired power of the design
t_1	t-value corresponding to the desired significance level of the test*
t_2	t-value corresponding to the desired power of the design*
σ_y	Standard deviation of outcome variable
J	Number of clusters
ρ	Intra-cluster correlation coefficient
P	Proportion of individuals assigned to the treatment group
n	Number of individuals per cluster

Note: * t-values are a function of specific sample size. Taking into account central limit theorem and the law of large numbers, a minimum sample size assumption needs to be made to calculate precise t-values. Otherwise, users may purposively decide the value for t_1 and t_2.

Example

The South African Ministry of Water and Environmental Affairs is seeking effective measures to reduce deforestation and land degradation. The Ministry plans to introduce farmer field schools (FFS) for land conservation. Before scaling up FFS at the national level, the Ministry has decided to work with a research team from the University of the Witwatersrand to evaluate whether FFS reduces land degradation. Farmers will be taught how to reduce deforestation and land degradation. In order to avoid spillovers or contamination from neighbouring farmers who might change their behaviour relating to deforestation and land degradation after observing and talking to farmers who received the intervention, the research team decides to use a cluster randomised design to assess the impact of FFS on land degradation.

Half of 240 villages that participate in this study will be assigned to the treatment group and the other half will be assigned to the control group. In each village, 20 farmers will be sampled. In the treatment group, the 20 farmers sampled will participate in FFS. A formative research work conducted by the research team in 10 villages close to the selected area of study shows that the ICC of land degradation is 0.037. The mean and standard deviation of degraded land are 1.26 hectares and 0.47 hectares respectively. Using a two-sided hypothesis test at the 0.01 significance level, we will calculate the MDE for a study with 90 per cent statistical power.

On the basis of this example, t_1 = 2.58, t_2 = 1.28, P = 0.5, J = 240, σ_y = 0.47, ρ = 0.037and n = 20, which yields MDE = 0.0683 hectares.

7.2.2 Two-level cluster randomised controlled trials with individual-level outcomes (continuous outcome) with covariates

MDE:

$$\delta = \frac{t_1 + t_2}{\sqrt{P(1-P)J}}\sigma_y\sqrt{\left[\rho + \frac{1-\rho}{n}\right](1-R^2)}$$

Number of clusters:

$$J = \left\{\frac{1}{P\delta^2}\sigma_y^2\frac{(t_1+t_2)^2}{-\rho+1}(-R^2+1)\left(\rho + \frac{1}{n}(-\rho+1)\right)\right\}$$

Table 8: Parameters required for two-level cluster randomised controlled trials with individual-level outcomes (continuous outcome) with covariates

δ	Minimum detectable effect
α	Desired significance level
β	Desired power of the design
t_1	t-value corresponding to the desired significance level of the test
t_2	t-value corresponding to the desired power of the design
σ_y	Standard deviation of the outcome variable
J	Number of clusters
ρ	Intra-cluster correlation coefficient
P	Proportion of individuals assigned to the treatment group
n	Number of individuals per cluster
R^2	Proportion of outcome variance explained by level 1 covariate(s)

Note: * t-values are a function of specific sample size. Taking into account central limit theorem and the law of large numbers, a minimum sample size assumption needs to be made to calculate precise t-values. Otherwise, users may purposively decide the value for t_1 and t_2.

Example

The South African Ministry of Water and Environmental Affairs is seeking effective measures to reduce deforestation and land degradation in South Africa. The Ministry plans to introduce FFS for land conservation, but before scaling up FFS at the national level, the Ministry has decided to work with a research team from the University of the Witwatersrand to evaluate whether FFS reduces land degradation. Farmers will be taught how to reduce deforestation and land degradation. In order to avoid spillovers or contamination from neighbouring farmers who might change their behaviour relating to deforestation and land degradation after observing and talking to farmers who received the intervention, the research team decides to use a cluster randomised design to assess the impact of FFS on land degradation.

Half of 240 villages that participate in this study will be assigned to the treatment group i group and the other half will be assigned to the control group. In each village, 20 farmers will be sampled. In the treatment group, the 20 farmers sampled will participate in FFS. A formative research work conducted by the research team in 10 villages close to the selected area of study shows that the ICC of land degradation is 0.037. The mean and standard deviation of degraded land is 1.26 hectares and 0.47 hectares respectively. In addition, this formative research reveals that degraded land regressed on income gives an R^2 equal to 0.4. Using a two-sided hypothesis test at the 0.01 significance level, we will calculate the MDE for a study with 90 per cent statistical power.

On the basis of this example, t_1 = 2.58, t_2 = 1.28, P = 0.5, J = 240, σ_y = 0.47, ρ = 0.037, R^2 = 0.4 and n = 20, which yields MDE = 0.053 hectares.

7.2.3 Two-level cluster randomised controlled trials with individual-level outcomes (binary outcome)[4]

MDE:

$$\mu_1 = \left\{ \frac{1}{a(K^2 n - 1) + n - Jn} \left(\frac{\mu_0 (n - Jn) - \frac{1}{2} a}{-\frac{1}{2}\sqrt{a\big(a + (-4)\mu_0 (a - 2n + 2Jn - aK^2 n)(\mu_0 - K^2 n \mu_0 - 1)\big)}} \right) \right\}$$

Number of clusters:

$$J = 1 + \frac{(z_1 + z_2)^2 \left[\frac{\mu_0(1-\mu_0)}{\mathrm{n}} + \frac{\mu_1(1-\mu_1)}{\mathrm{n}} + \mathrm{k}^2(\mu_0^2 + \mu_1^2) \right]}{(\mu_0 - \mu_1)^2}$$

Table 9: Parameters required for two-level cluster randomised controlled trials with individual-level outcomes (binary outcome)

α	Desired significance level
β	Desired power of the design
z_1	z-value corresponding to the desired significance level of the test
z_2	z-value corresponding to the desired power of the design
μ_1	True (population) proportion in the presence of the intervention
μ_0	True (population) proportion in the absence of the intervention
n	Number of individuals in each cluster
J	Number of clusters in each group
k	The coefficient of variation of true proportions between clusters within each group[5]

Example

In North Cameroon, one of the leading causes of morbidity among children under 5 years is vitamin A deficiency. In order to address this problem, a not-for-profit organisation works to educate mothers on the importance of vitamin A. Specifically, organisation staff visit houses where there are children under 5 years old, and if a vitamin A supplement is not being administered to a child, the mother or caregiver is advised to visit the closest health facility to receive a supplement. In order to evaluate whether this approach reduces morbidity, the organisation is working with a research team from the University of Maroua. The research team

[4] Formulae under section 1 and 2 apply for unmatched studies, for pair-matched studies, formulae can be used with two modifications. Firstly, the addition of 2 rather than 1 to the required numbers of clusters. Secondly, k is replaced by Km, the coefficient of variation in true proportions (rates) between clusters within the matched pairs in the absence of intervention (Hayes and Bennett, 1999).

[5] For binary outcomes, the relationship between the ICC and K can be found in Pagel et al. (2011).

will evaluate the impact of this intervention (door-to-door visits for promoting vitamin A supplements) on coverage of vitamin A supplements (proportion of children under 5 years with vitamin A supplementation). To avoid contamination, the research team decides to conduct a cluster randomised controlled trial where a cluster is defined as a health facility catchment area constituted of a few villages. Half of the health facilities will be assigned to the intervention group and the other half will be assigned to the control group. In each village, 50 children will be sampled in the catchment area served by the health facility. A survey conducted by the organisation a few years previously indicates that the proportion of children already receiving a vitamin A supplement is 0.25 and that *k* is equal to 0.25. The aim of the organisation is to increase this coverage to 0.65. Using a two-sided hypothesis test at the 0.01 significance level, the research team will calculate the number of health facilities required in each group for a study with 80 per cent statistical power.

On the basis of this example, z_1 = 2.58, z_2 = 0.84, μ_0 = 0.25, μ_1 = 0.65, k = 0.25 and n = 50, which yields four health facilities (clusters) in each group.

7.2.4 Two-level cluster randomised controlled trials with individual-level outcomes (rates)

MDE:

$$\mu_1 = \left\{\frac{1}{n(ak^2 - J + 1)}\left(-\frac{1}{2}\right)\left(a - 2n\mu_0 + 2Jn\mu_0 - \sqrt{a\big(a + (-4)\mu_0 n(aK^2 - 2J + 2)(K^2 n\mu_0 + 1)\big)}\right)\right\}$$

Number of clusters:

$$J = 1 + \frac{(z_1 + z_2)^2\left[\frac{(\mu_0+\mu_1)}{n} + k^2(\mu_0^2 + \mu_1^2)\right]}{(\mu_0 - \mu_1)^2}$$

Table 10: Parameters required for two-level cluster randomised controlled trials with individual-level outcomes (rates)

α	Desired significance level
β	Desired power of the design
z_1	z-value corresponding to the desired significance level of the test
z_2	z-value corresponding to the desired power of the design
μ_1	True (population) rate in the presence of the intervention
μ_0	True (population) rate in the absence of the intervention
n	Number of individuals in each cluster
J	Number of clusters in each group
k	The coefficient of variation of true proportions between clusters within each group

Example
In North Cameroon, one of the leading causes of morbidity among children under 5 years is vitamin A deficiency. In order to address this problem, a not-for-profit organisation works to educate mothers on the importance of vitamin A. Specifically, organisation staff visit houses where there are children under 5 years old, and if a vitamin A supplement is not being administered to a child, the mother or caregiver is advised to visit the closest health facility to receive a supplement. In order to evaluate whether this approach reduces morbidity, the organisation is working with a research team from the University of Maroua. The research team will evaluate the impact of this intervention (door-to-door visits for promoting vitamin A supplements) on the morbidity of children under 5 years over a 3-year period. To avoid contamination, the research team decides to conduct a cluster randomized controlled trial where a cluster is defined as a health facility catchment area constituted of a few villages. Half of the health facilities will be assigned to the intervention group and the other half will be assigned to the control group. In each village, 50 children will be sampled in the catchment area served by the health facility. Mortality data for 3 years prior to the study indicates that there were a total of 1,872 deaths over 12,561 person-years of observation in 31 health facilities in the extreme north region with data. The research team assumes that the mortality rate in the control group remains constant and the intervention will reduce mortality by 50 per cent. The estimated value of *k* is equal to 0.25. Using a two-sided hypothesis test at the 0.01 significance level, the research team will calculate the number of health facilities required in each treatment group for a study with 80 per cent statistical power.

On the basis of this example, $z_1 = 2.57$, $z_2 = 0.84$, $\mu_0 = 0.05$, $\mu_1 = 0.025$, $k = 0.25$ and $n = 50$, which yields 33 health facilities (clusters) in each group.

References

Bloom, HS, 1995. Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review*, 19, pp.547–556.

Gleason, P, 2010. *The analysis of statistical power for estimating program impacts*. Washington DC: MATHEMATICA Policy Research.

Goodman, S, 2008. A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, 45, pp.135–140.

Hayes, RJ and Bennett, S, 1999. Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology*, 28, pp.319–326.

Hoenig, JM and Heisey, DM, 2001. The abuse of power. *The American Statistician*, 55, pp.19–24.

Killip, S, Mahfoud, Z and Pearce, K, 2004. What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *Annals of Family Medicine*, 2, pp.204–208.

Muñoz, J. (2013). Statistical power and impact evaluation.
Paper presented at the REGIONAL IMPACT EVALUATION AND SURVEY METHODS WORKSHOP Evaluating the Impact of Development Programs: Turning Promises into Evidence, New Delhi, India.

Pagel et al.: Intracluster correlation coefficients and coefficients of variation for perinatal outcomes from five cluster randomised controlled trials in low and middle-income countries: results and methodological implications. Trials 2011 12:151

Prajapati, B, Dunne, M and Armstrong, R, 2010. Sample size estimation and statistical power analyses. *Optometry Today*, 16, pp.10–18.

Raudenbush, SW, Liu, X-F, Congdon, R and Martinez, A, 2011. *Optimal Design for Longitudinal and Multilevel Research: Documentation for the Optimal Design Software Version 3.0.*

Schochet, PZ, 2008. *Technical methods report: Guidelines for multiple testing in impact evaluations (NCEE 2008-4018).* Washington, DC: National Center for Education Evaluation and Regional Assistance: Institute of Education Sciences, U.S. Department of Education.

Publications in the 3ie Working Paper Series

The following papers are available from http://www.3ieimpact.org/en/publications/working-papers/

Evaluations with impact: decision-focused impact evaluation as a practical policymaking tool, 3ie Working Paper 25. Shah, NB, Wang, P, Fraker, A and Gastfriend, D (2015)

Impact evaluation and policy decisions: where are we? A Latin American think-tank perspective, 3ie Working Paper 24. Baanante, MJ and Valdivia, LA (2015)

What methods may be used in impact evaluations of humanitarian assistance? 3ie Working Paper 22. Puri, J, Aladysheva, A, Iversen, V, Ghorpade, Y and Brück, T (2014)

Impact evaluation of development programmes: experiences from Viet Nam, 3ie Working Paper 21. Nguyen Viet Cuong (2014)

Quality education for all children? What works in education in developing countries, 3ie Working Paper 20. Krishnaratne, S, White, H and Carpenter, E (2013)

Promoting commitment to evaluate, 3ie Working Paper 19. Székely, M (2013)

Building on what works: commitment to evaluation (c2e) indicator, 3ie Working Paper 18. Levine, CJ and Chapoy, C (2013)

From impact evaluations to paradigm shift: A case study of the Buenos Aires Ciudadanía Porteña conditional cash transfer programme, *3ie Working Paper 17*. Agosto, G, Nuñez, E, Citarroni, H, Briasco, I and Garcette, N (2013)

Validating one of the world's largest conditional cash transfer programmes: A case study on how an impact evaluation of Brazil's Bolsa Família Programme helped silence its critics and improve policy, *3ie Working Paper 16*. Langou, GD and Forteza, P (2012)

Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework, 3ie Working Paper 15. White, H and Phillips, D (2012)

Behind the scenes: managing and conducting large scale impact evaluations in Colombia, *3ie Working Paper 14.* Briceño, B, Cuesta, L and Attanasio, O (2011)

Can we obtain the required rigour without randomisation? 3ie Working Paper 13. Hughes, K and Hutchings, C (2011)

Sound expectations: from impact evaluations to policy change, 3ie Working Paper 12. Weyrauch, V and Langou, GD (2011)

A can of worms? Implications of rigorous impact evaluations for development agencies, 3ie Working Paper 11. Roetman, E (2011)

Conducting influential impact evaluations in China: the experience of the Rural Education Action Project, *3ie Working Paper 10*. Boswell, M, Rozelle, S, Zhang, L, Liu, C, Luo, R and Shi, Y (2011)

An introduction to the use of randomised control trials to evaluate development interventions, *3ie Working Paper 9*. White, H (2011)

Institutionalisation of government evaluation: balancing trade-offs, *3ie Working Paper 8.* Gaarder, M and Briceño, B (2010)

Impact evaluation and interventions to address climate change: a scoping study, *3ie Working Paper 7*. Snilstveit, B and Prowse, M (2010)

A checklist for the reporting of randomised control trials of social and economic policy interventions in developing countries, *3ie Working Paper 6.* Bose, R (2010)

Impact evaluation in the post-disaster setting, 3ie Working Paper 5. Buttenheim, A (2009)

Designing impact evaluations: different perspectives, contributions, 3ie Working Paper 4. Chambers, R, Karlan, D, Ravallion, M and Rogers, P (2009) [Also available in Spanish, French and Chinese]

Theory-based impact evaluation, *3ie Working Paper 3*. White, H (2009) [Also available in French and Chinese]

Better evidence for a better world, *3ie Working Paper 2.* Lipsey, MW (ed.) and Noonan, E (2009)

Some reflections on current debates in impact evaluation, *3ie Working Paper 1*. White, H (2009)

Experimental and quasi-experimental impact evaluation designs are increasingly used to measure the impact of development interventions. This manual presents the basic statistical concepts used in power calculations for experimental design. It provides detailed definitions of parameters used to perform power calculations, useful rules of thumb and different approaches that can be used when performing power calculations. The authors draw from real world examples to calculate statistical power for individual and cluster randomised controlled trials. This manual provides formulae for sample size determination and minimum detectable effect associated with a given statistical power. The manual is ccompanied by the *Sample size and minimum detectable effect calculator,* © a free online tool developed by the authors. It allows users to work directly with the formulae presented in the manual.

Working Paper Series

International Initiative for Impact Evaluation
202-203, Rectangle One
D-4, Saket District Centre
New Delhi – 110017
India

3ie@3ieimpact.org
Tel: +91 11 4989 4444

3ie www.3ieimpact.org