e-valuate

sample size calculator

User guide

Version 1: July 2017

Contents

1. Overview

E-valuate provides three simple calculation tools for anyone seeking to do an evaluation. It is designed for researchers or evaluators, or those seeking to commission an evaluation and need a quick, easy and reliable means of estimating the key variables for their study.

In an impact evaluation, a test is undertaken to see whether a change measured in an intervention group is greater or less than that in a comparison group, and whether the change is statistically significant. The intervention group undertakes a 'treatment' of some kind, which the researcher would like to test in terms of its effectiveness in changing a variable of interest. The comparison group does not receive the treatment and therefore serves as a control for whether the treatment has an impact. The selection of individuals into treatment and comparison groups could be random as with a randomised control trial (RCT), but they could be matched through other means, often known as quasi-experimental methods. The same calculations apply for these different sampling methods in terms of sample size, power and calculating the minimum detectable effect size.

The bigger a change is measured in the variable of interest, the more likely it is to be statistically significant for a given sample size. Other variables are also vital, such as the power of the evaluation, where a larger sample size would be needed for 90% statistical power than for 80% for example. In addition, researchers and evaluators need to consider whether *clustering* needs to be applied. E-valuate provides all the key variables required to make these calculations.

The key variables involved and statistical terms associated are explained in the instructions below. Follow the instructions alongside your calculation to ensure you determine the most robust and accurate results for your evaluation.

The three calculators and what they do

E-valuate has three calculators depending on which variables are known or set. Click on the links below to find the instructions for each calculator.

The **Sample Size Calculator** allows you to find out the sample size required for your evaluation for a given level of statistical significance and power, as well as the minimum detectable effect size you would like to measure.

The **Power Calculator** provides a measure of the statistical significance of your evaluation for a given sample size and minimum detectable effect size.

The **Effect Size Calculator** allows you to measure what your minimum detectable effect size will be for a given sample size and measure of statistical significance.

2. Sample Size Calculator

The E-valuate Sample Size Calculator allows you to find out the sample size required for your evaluation for a given level of statistical significance and power, as well as the minimum detectable effect size you would like to measure.

The Sample Size Calculator provides all the key variables you need to complete the calculation. Each of the key variables is discussed below.

Type of Test - binary or continuous variable

The variable of interest is the outcome or impact you want to measure a change in. This is the variable for which you would like to test whether the intervention or treatment causes the change. There are two main types of variable that you are likely to encounter and each requires a different statistical test.

A **binary variable** is one that can take two values – e.g. 1 or 0. This could be, for example, formal school enrolment or the take-up of a vaccine among a population. A child may be formally enrolled (1) or not formally enrolled (0); a vaccine can either be applied (1) or not applied (0). A binary test between a treatment and comparison group would then seek to test whether enrolment or vaccine-use is increased by the intervention. In a binary test the change will be in the *'proportion'* of children enrolled or individuals using vaccines. The effect size will therefore be measured as a change between two proportions. For the Sample Size Calculator, you will need to estimate an initial proportion (Proportion 1) and a final proportion (Proportion 2), where the difference in between is the minimum detectable effect size you will be able to measure. The bigger the difference between the two proportions and therefore the effect size, the smaller the sample size will be.

A **continuous variable** is one that can take many numerical values and these may range widely. This could be, for example, the distribution of scores on a school test. The effect size for an evaluation using a continuous variable is usually measured in terms of the number of standard deviations, a measure of change relative to the existing distribution. For example, if you wanted to estimate whether an improvement in school test scores in an intervention group compared to a comparison group of 5 was significant, and the standard deviation of test scores was 10, the effect size in this case would be 0.5 standard deviations (5 divided by 10). The bigger the effect size, the smaller the sample size will need to be.

Type of test - one-sided or two-sided

Statistical tests can be performed as one-sided or two-sided tests. Specifying which will be used is important for power calculations as one-sided tests require smaller sample sizes compared with two-sided tests. The choice between the two should be made

based on an understanding if the type of test being conducted and the existing evidence of similar evaluations conducted.

One-sided tests are used when the direction of change is expected to be in one direction, i.e. the researcher can argue that the intervention is expected to either raise or lower the value of the outcome variable of interest. Ideally, a strong body of evidence should be available, leading the researcher to predict that the new intervention will improve outcomes when compared to a comparison group. This normally justifies using a one-sided test to interpret the statistical results.

A **two-sided statistical test** should be used when the direction of change could be in either direction, i.e. when the researcher cannot predict whether the intervention will have a positive or negative impact. This can be justified when there is a lack of pre-existing evidence on the type of intervention and the level and direction of impact it might have.

Ratio between treatment and comparison samples

In many evaluations the sample size for the treatment group and comparison group will be the same, and statistically it is better if they are the same. However, for practical reasons, it may sometimes not be possible to have the same sample size. The ratio between the two sample sizes is then important for the statistical calculation. For the E-valuate Sample Size Calculator, the default value is 1 (i.e. the treatment sample size and comparison sample size are the same, a 1:1 ratio). However, you can set a different value. For example, for a treatment group sample size that is double the comparison group sample size, enter "2" (representing a 2:1 ratio).

Power & Significance

There are two measures of statistical significance required in calculating a sample size, which measure the probability of different errors taking place with the evaluation.

The **level of significance** (or 'alpha') gives the probability of detecting an effect that is not present, i.e. a *false positive result*, also known as a 'Type 1 error'. This would be a result inferring that the intervention lead to a change when in reality the measured change occurred by chance. The significance level is known as α (alpha); its converse, the *confidence level,* is defined as (1 – α). The confidence level is the probability that you do not find a statistically significant effect if the treatment effect is zero. The default value of the level of significance in the E-valuate Sample Size Calculator is 5% (0.05), equivalent to a confidence level of 95%. However, some researchers may require less rigorous experimental results and therefore set it at 10% (0.10), while others may look for a more rigorous result and set it at 2.5% (0.025) or 1% (0.01). This decision should be made based on the severity of consequences and costs of producing a false positive result, which might mean concluding a treatment is useful and spending resources on

expanding it, when in fact it is not an effective treatment.

The **statistical power** is the probability of correctly concluding that an intervention does not have any statistically significant effect. Statistically it is formally expressed as (1 – β), with β as the probability of concluding an observed effect is due to chance when in reality a genuine effect was present. This would be a *false negative result,* known as a 'Type 2 error'. Power is usually set at 80% (0.80), which is the default value for the E-valuate Sample Size Calculator, implying the probability of making a type 2 error is 20%. This is generally seen as the minimum value for a good evaluation, but some researchers may prefer to have a higher value such as 90% (0.90) or higher. In practical terms, the value should be considered based on the severity of consequences and costs of producing a false negative result, i.e. inferring the treatment does not have a causal impact when it actually does, and therefore potentially discontinuing a useful treatment.

Effect size if continuous variable (Minimum Detectable Effect)

For a continuous variable, the effect size should be entered in terms of the number of standard deviations of impact expected. The calculator will provide a sample size to at least capture this 'minimum detectable effect' (MDE). Note that MDE should not be an ad hoc choice. It should be based first of all on existing theory, models and empirical evidence about the scale of impact that is anticipated for the given intervention. In different contexts the implications of the scale of impact in standard deviation terms may vary. For example, in an education evaluation, a 0.2 standard deviations effect on test scores may be seen as a good or moderate effect, while 0.5 standard deviations is likely to be seen as a large effect; while for a medical trial of drugs to reduce blood pressure, a 0.5 standard deviations reduction may be seen as a small effect relative to other drugs available.

Effect size / Proportions if binary variable

For a binary variable the effect size takes the form of a change in the proportions of a population. This could be, for example, formal school enrolment. A child may be formally enrolled (1) or not formally enrolled (0). In a binary test the change measured will be in the 'proportion' of children enrolled, and the effect size will therefore be measured as a change in this proportion. For the Sample Size Calculator, you need to estimate an initial proportion (Proportion 1) and a final proportion (Proportion 2), where the difference between them is the effect size you would like to measure, also known as the minimum detectable effect (MDE). As for a continuous variable, the MDE should not be an ad hoc choice, but should be based on existing theory, models and empirical evidence available about the scale of impact that is anticipated for the given intervention. If you want to estimate a change in the take-up of vaccines and the initial take-up is 40%, you should enter 0.40 for Proportion 1. If you want to test whether a MDE of a 10% increase in vaccine take-up is achieved for the intervention group, this

implies 50% vaccine-use after the intervention, you should enter 0.50 for Proportion 2.

Clustering

Clustering is an important aspect in the statistical test for your evaluation, if not considered it can significantly reduce the rigour of the measurement of impact. You therefore have to carefully consider whether to select “Yes” to apply clustering (the default value for the E-Valuate Sample Size Calculator is “No”).

It is rare that an intervention will be completely randomised at an individual level, practically it can be very difficult and expensive to do so - for example it would be very difficult to ask a teacher to use one teaching methodology for some students and another for others within the same class. Even if it is possible, there may be a problem of contamination – for example, students may pick up from other students in their class what they are learning from the improved teaching methodology. Interventions are therefore often allocated to whole schools, hospitals, villages or other locations. If this is the case, the chosen location is known as the primary sampling unit, and is the *‘cluster’*.

In general, you should sample as many clusters as is possible, with the number of individuals per cluster the same for each. However, for logistical reasons you may be unable to sample all of the locations in which the intervention is working. For sampling purposes, you should therefore list all the clusters in the population, and from the list, select the clusters – usually with a simple random sampling strategy, assigning clearly to either the intervention group or comparison group.

For the calculator, if you select “Yes” to apply clustering, you will then be asked to enter the **number of clusters** you intend to sample in the treatment group and in the comparison group. This will be the number of schools, hospitals, villages or other defined cluster which you have defined as your primary sampling unit.

You are also asked to enter the **intra-cluster correlation (ICC) coefficient** for both your treatment group and comparison group. The ICC is a measure of how much the cluster determines the level of change experienced, and technically is defined as the ratio in the variance between clusters and the total variance for the variable of interest. The ICC takes a value between 0 and 1. The closer the ICC is to 1 the more of the variation is explained by the differences between the clusters; where a value of 1 would indicate all the variation was explained by the differences between clusters, and a value of 0 would indicate none of the variation is explained by the cluster variable. If sampling health clinics or schools for example, it is possible that when sampling in different areas that are economically diverse in terms of household income, a lot of variation between clusters in variables of interest such as health and education outcomes may be driven by these income differences – a higher ICC might be expected in such cases. This compares to a case where the health clinics or schools were taken from areas that were similar from a socio-economic point of view –a lower ICC would be expected

comparatively in this case.

The ICC chosen should ideally be based on data, and if existing data-sets are available it can be calculated with the *loneway* function on the statistical software package Stata for example. If no data on the expected ICC is available then some social sciences researchers suggest a default value of 0.1 for the ICC, but it will depend on the type of variable, the type of cluster, and the geographical variation for the intervention and comparison groups. Similar evaluations should be assessed to see what ICC may be expected in the given context for the variable of interest.

Sample attrition

The E-Valuate Sample Size Calculator allows you to factor in sample attrition. Attrition refers to a reduction of the initial sample size involved in a study, with a lower sample size reached at the end-point of the evaluation. Attrition may occur as a result of outward migration of some participants from the study area, their refusal to continue participating in the research, or other factors. In general, the longer the time between data collection points the higher attrition is likely to be. Attrition will reduce the statistical power as it decreases the sample size, making it very important to take account of.

The default value for applying attrition in the Sample Size Calculator is “Yes”, as it is advisable to consider attrition, with at least some attrition likely in almost any context. It is better to anticipate the potential attrition rate during the evaluation preparation phase and account for it through oversampling at the beginning of the study. It is also advisable to consider oversampling any particular subgroup (whether treatment or comparison group) if the subgroups have different likely rates of attrition. Researchers and evaluators should consider the likelihood of attrition based on their previous experience or the experience of others with similar interventions over similar time-periods. Some researchers may factor in 10% attrition; others may need to factor in higher levels of attrition depending on the context.

Rounding (if clustering applied)

If you have applied clustering in the Sample Size Calculator, you will have one final option which is whether to round up the final sample size estimate to ensure that you have an equal number of participants per cluster. This can be useful to reduce the complexity of sampling different numbers of households per village – for an implied sample size of 7.5 per cluster you would need to sample 8 households for half the villages and 7 households for other half. The calculator sets the default value to “No” so consider if you would like to change this to “Yes”.

Final sample size

The final sample size provides the result of the calculation. This should automatically

provide you with a required sample size for your Intervention group and Comparison group given all of the variable values you have provided.

If you have applied clustering the final sample size will also provide you with the average sample per cluster to one decimal place.

In some circumstances the entry may read n/a (not applicable) - this will mean that it is not possible to compute a sample size for the variables you have entered. Usually this will be because the effect size is too small to be detected given you the values for variables you have entered.

3. Power Calculator

The E-valuate Power Calculator provides a measure of the statistical significance of your evaluation for a given sample size and minimum detectable effect size.

The Power Calculator provides all the key variables you need to complete the calculation. Each of the key variables is discussed below.

Type of Test - binary or continuous variable

The variable of interest is the outcome or impact you want to measure a change in. This is the variable for which you would like to test whether the intervention or treatment causes the change. There are two main types of variable that you are likely to encounter and each requires a different statistical test.

A **binary variable** is one that can take two values – e.g. 1 or 0. This could be, for example, formal school enrolment or the take-up of a vaccine among a population. A child may be formally enrolled (1) or not formally enrolled (0); a vaccine can either be applied (1) or not applied (0). A binary test between a treatment and comparison group would then seek to test whether enrolment or vaccine-use is increased by the intervention. In a binary test the change will be in the *'proportion'* of children enrolled or individuals using vaccines. The effect size will therefore be measured as a change between two proportions. For the Power Calculator, you will need to estimate an initial proportion (Proportion 1) and a final proportion (Proportion 2), where the difference in between is the minimum detectable effect size you will be able to measure. The bigger the difference between the two proportions and therefore the effect size, the larger the calculated statistical power will be.

A **continuous variable** is one that can take many numerical values and these may range widely. This could be, for example, the distribution of scores on a school test. The effect size for an evaluation using a continuous variable is usually measured in terms of the number of standard deviations, a measure of change relative to the existing distribution. For example, if you wanted to estimate whether an improvement in school test scores in an intervention group compared to a comparison group of 5 was significant, and the standard deviation of test scores was 10, the effect size in this case would be 0.5 standard deviations (5 divided by 10). The bigger the effect size, the larger the calculated statistical power will be.

Type of test - one-sided or two-sided

Statistical tests can be performed as one-sided or two-sided tests. Specifying which will be used is important for power calculations as one-sided tests require smaller sample sizes compared with two-sided tests. The choice between the two should be made based on an understanding if the type of test being conducted and the existing evidence

of similar evaluations conducted.

One-sided tests are used when the direction of change is expected to be in one direction, i.e. the researcher can argue that the intervention is expected to either raise or lower the value of the outcome variable of interest. Ideally, a strong body of evidence should be available, leading the researcher to predict that the new intervention will improve outcomes when compared to a comparison group. This normally justifies using a one-sided test to interpret the statistical results.

A **two-sided statistical test** should be used when the direction of change could be in either direction, i.e. when the researcher cannot predict whether the intervention will have a positive or negative impact. This can be justified when there is a lack of pre-existing evidence on the type of intervention and the level and direction of impact it might have.

Sample sizes

For the Power Calculator, you need to know the sample sizes for your intervention group and comparison group. The sample sizes for intervention group and comparison group should be the final sample sizes you will be able to use for the evaluation. I.e. they should factor in any attrition experienced between the initial and follow-up data collection.

Level of significance

There are two measures of statistical significance that are important in a power calculation, measuring the probability of different errors taking place with the evaluation. When using the E-valuate Power Calculator, you only need to enter the level of significance, as the goal of the calculation is to calculate the power (see below).

The **level of significance** (or 'alpha') gives the probability of detecting an effect that is not present, i.e. a *false positive result*, also known as a 'Type 1 error'. This would be a result inferring that the intervention lead to a change when in reality the measured change occurred by chance. The significance level is known as α (alpha); its converse, the *confidence level,* is defined as (1 – α). The confidence level is the probability that you do not find a statistically significant effect if the treatment effect is zero. The default value of the level of significance in the E-valuate Power Calculator is 5% (0.05), equivalent to a confidence level of 95%. However, some researchers may require less rigorous experimental results and therefore set it at 10% (0.10), while others may look for a more rigorous result and set it at 2.5% (0.025) or 1% (0.01). This decision should be made based on the severity of consequences and costs of producing a false positive result, which might mean concluding a treatment is useful and spending resources on expanding it, when in fact it is not an effective treatment.

Effect size if continuous variable (Minimum Detectable Effect)

For a continuous variable, the effect size should be entered in terms of the number of standard deviations of impact expected. The calculator will provide the power achieved if this 'minimum detectable effect' (MDE) is achieved. For example, if you want to estimate whether an improvement in school test scores in an intervention group compared to a comparison group of 5 is significant, and the standard deviation of test scores is 10, the effect size in this case will be 0.5 standard deviations (5 divided by 10) – you enter 0.50 for the effect size.

Effect size / Proportions if binary variable

For a binary variable the effect size takes the form of a change in the proportions of a population. This could be, for example, formal school enrolment. A child may be formally enrolled (1) or not formally enrolled (0). In a binary test the change measured will be in the 'proportion' of children enrolled. The effect size will therefore be measured as a change between the proportions. For the Power Calculator, you need to estimate an initial proportion (Proportion 1) and a final proportion (Proportion 2), where the difference between them is the effect size you would like to measure, also known as the minimum detectable effect (MDE). If you want to estimate a change in the take-up of vaccines and the initial take-up is 40%, you should enter 0.40 for Proportion 1. If you want to test whether a minimum detectable effect of a 10% increase in vaccine take-up is achieved for the intervention group, this would imply 50% vaccine-use after the intervention, so you should enter 0.50 for Proportion 2.

Clustering

Clustering is an important aspect in the statistical test for your evaluation, if not considered it can significantly reduce the rigour of the measurement of impact. You therefore have to carefully consider whether to select "Yes" to apply clustering (the default value for the E-Valuate Power Calculator is "No").

It is rare that an intervention will be completely randomised at an individual level, practically it can be very difficult and expensive to do so - for example it would be very difficult to ask a teacher to use one teaching methodology for some students and another for others within the same class. Even if it is possible, there may be a problem of contamination – for example, students may pick up from other students in their class what they are learning from the improved teaching methodology. Interventions are therefore often allocated to whole schools, hospitals, villages or other locations. If this is the case, the chosen location is known as the primary sampling unit, and is the *'cluster'*.

In general, you should sample as many clusters as is possible, with the number of individuals per cluster the same for each. However, for logistical reasons you may be

unable to sample all of the locations in which the intervention is working. For sampling purposes, you should therefore list all the clusters in the population, and from the list, select the clusters – usually with a simple random sampling strategy, assigning clearly to either the intervention group or comparison group.

For the calculator, if you select "Yes" to apply clustering, you will then be asked to enter the **number of clusters** you intend to sample in the treatment group and in the comparison group. This will be the number of schools, hospitals, villages or other defined cluster which you have defined as your primary sampling unit.

You are also asked to enter the **intra-cluster correlation (ICC) coefficient** for both your treatment group and comparison group. The ICC is a measure of how much the cluster determines the level of change experienced, and technically is defined as the ratio in the variance between clusters and the total variance for the variable of interest. The ICC takes a value between 0 and 1. The closer the ICC is to 1 the more of the variation is explained by the differences between the clusters; where a value of 1 would indicate all the variation was explained by the differences between clusters, and a value of 0 would indicate none of the variation is explained by the cluster variable. If sampling health clinics or schools for example, it is possible that when sampling in different areas that are economically diverse in terms of household income, a lot of variation between clusters in variables of interest such as health and education outcomes may be driven by these income differences – a higher ICC might be expected in such cases. This compares to a case where the health clinics or schools were taken from areas that were similar from a socio-economic point of view –a lower ICC would be expected comparatively in this case.

The ICC chosen should ideally be based on data, and if existing data-sets are available it can be calculated with the *loneway* function on the statistical software package Stata for example. If no data on the expected ICC is available then some social sciences researchers suggest a default value of 0.1 for the ICC, but it will depend on the type of variable, the type of cluster, and the geographical variation for the intervention and comparison groups. Similar evaluations should be assessed to see what ICC may be expected in the given context for the variable of interest.

Power achieved

The final calculation for the Power Calculator is the statistical power your evaluation will achieve.

The **statistical power** is the probability of correctly concluding that an intervention does not have any statistically significant effect. Statistically it is formally expressed as (1 – β), with β as the probability of concluding an observed effect is due to chance when in reality a genuine effect was present. This would be a *false negative result,* known as a 'Type 2 error'. Overall, the higher the value of statistical power you calculate the more

robust the results of the evaluation will be.

4. Effect Size Calculator

The E-valuate Effect Size Calculator allows you to measure what your 'minimum detectable effect' size will be for a given sample size and measure of statistical significance.

The Effect Size Calculator provides all the key variables you need to complete the calculation. Each of the key variables is discussed below.

Type of Test - binary or continuous variable

The variable of interest is the outcome or impact you want to measure a change in. This is the variable for which you would like to test whether the intervention or treatment causes the change. There are two main types of variable that you are likely to encounter and each requires a different statistical test.

A **binary variable** is one that can take two values – e.g. 1 or 0. This could be, for example, formal school enrolment or the take-up of a vaccine among a population. A child may be formally enrolled (1) or not formally enrolled (0); a vaccine can either be applied (1) or not applied (0). A binary test between a treatment and comparison group would then seek to test whether enrolment or vaccine-use is increased by the intervention. In a binary test the change will be in the *'proportion'* of children enrolled or individuals using vaccines. The effect size will therefore be measured as a change between two proportions. For the Effect Size Calculator, you will need to estimate an initial proportion (Proportion 1), and the calculator will then give you a final proportion (Proportion 2), where the difference in between is the minimum detectable effect size you will be able to measure.

A **continuous variable** is one that can take many numerical values and these may range widely. This could be, for example, the distribution of scores on a school test. The effect size for an evaluation using a continuous variable is usually measured in terms of the number of standard deviations, a measure of change relative to the existing distribution. For example, if you wanted to estimate whether an improvement in school test scores in an intervention group compared to a comparison group of 5 was significant, and the standard deviation of test scores was 10, the effect size in this case would be 0.5 standard deviations (5 divided by 10).

Type of test - one-sided or two-sided

Statistical tests can be performed as one-sided or two-sided tests. Specifying which will be used is important for your effect size calculations as one-sided tests require smaller sample sizes compared with two-sided tests. The choice between the two should be made based on an understanding if the type of test being conducted and the existing evidence of similar evaluations conducted.

One-sided tests are used when the direction of change is expected to be in one direction, i.e. the researcher can argue that the intervention is expected to either raise or lower the value of the outcome variable of interest. Ideally, a strong body of evidence should be available, leading the researcher to predict that the new intervention will improve outcomes when compared to a comparison group. This normally justifies using a one-sided test to interpret the statistical results.

A **two-sided statistical test** should be used when the direction of change could be in either direction, i.e. when the researcher cannot predict whether the intervention will have a positive or negative impact. This can be justified when there is a lack of pre-existing evidence on the type of intervention and the level and direction of impact it might have.

Sample sizes

For the Effect Size Calculator, you need to know the sample sizes for your intervention group and comparison group. The sample sizes for intervention group and comparison group should be the final sample sizes you will be able to use for the evaluation. I.e. they should factor in any attrition experienced between the initial and follow-up data collection.

Proportions (if binary variable)

If you are running an evaluation with a binary variable, the change will be expressed as a change in proportions. For example, if you want to estimate a change in the take-up of vaccines and the initial take-up is 40%, you should enter 0.40 for Proportion 1.

If you have selected a one-sided test and therefore the direction of change should be either positive or negative, you have to then enter whether you expect the Proportion 2 to be higher or lower. So in the case above if you expect vaccine take-up to increase, you should enter “Higher”. If the variable of interest is, say, prevalence of malaria, and you expect this to go down, then the value you enter for the direction of change for Proportion 2 should be “Lower”.

Power & Significance

There are two measures of statistical significance required in assessing the effect size achieved, which measure the probability of different errors taking place with the evaluation.

The **level of significance** (or ‘alpha’) gives the probability of detecting an effect that is not present, i.e. a *false positive result*, also known as a ‘Type 1 error’. This would be a result inferring that the intervention lead to a change when in reality the measured

change occurred by chance. The significance level is known as α (alpha); its converse, the *confidence level,* is defined as (1 – α). The confidence level is the probability that you do not find a statistically significant effect if the treatment effect is zero. The default value of the level of significance in the E-valuate Effect Size Calculator is 5% (0.05), equivalent to a confidence level of 95%. However, some researchers may require less rigorous experimental results and therefore set it at 10% (0.10), while others may look for a more rigorous result and set it at 2.5% (0.025) or 1% (0.01). This decision should be made based on the severity of consequences and costs of producing a false positive result, which might mean concluding a treatment is useful and spending resources on expanding it, when in fact it is not an effective treatment.

The **statistical power** is the probability of correctly concluding that an intervention does not have any statistically significant effect. Statistically it is formally expressed as (1 – β), with β as the probability of concluding an observed effect is due to chance when in reality a genuine effect was present. This would be a *false negative result,* known as a 'Type 2 error'. Power is usually set at 80% (0.80), which is the default value for the E-valuate Effect Size Calculator, implying the probability of making a type 2 error is 20%. This is generally seen as the minimum value for a good evaluation, but some researchers may prefer to have a higher value such as 90% (0.90) or higher. In practical terms, the value should be considered based on the severity of consequences and costs of producing a false negative result, i.e. inferring the treatment does not have a causal impact when it actually does, and therefore potentially discontinuing a useful treatment.

Clustering

Clustering is an important aspect in the statistical test for your evaluation, if not considered it can significantly reduce the rigour of the measurement of impact. You therefore have to carefully consider whether to select "Yes" to apply clustering (the default value for the E-Valuate Effect Size Calculator is "No").

It is rare that an intervention will be completely randomised at an individual level, practically it can be very difficult and expensive to do so - for example it would be very difficult to ask a teacher to use one teaching methodology for some students and another for others within the same class. Even if it is possible, there may be a problem of contamination – for example, students may pick up from other students in their class what they are learning from the improved teaching methodology. Interventions are therefore often allocated to whole schools, hospitals, villages or other locations. If this is the case, the chosen location is known as the primary sampling unit, and is the *'cluster'*.

For the calculator, if you select "Yes" to apply clustering, you will then be asked to enter the **number of clusters** you sampled in the treatment group and in the comparison group. This will be the number of schools, hospitals, villages or other defined cluster which you have defined as your primary sampling unit.

You are also asked to enter the **intra-cluster correlation (ICC) coefficient** for both your treatment group and comparison group. The ICC is a measure of how much the cluster determines the level of change experienced, and technically is defined as the ratio in the variance between clusters and the total variance for the variable of interest. The ICC takes a value between 0 and 1. The closer the ICC is to 1 the more of the variation is explained by the differences between the clusters; where a value of 1 would indicate all the variation was explained by the differences between clusters, and a value of 0 would indicate none of the variation is explained by the cluster variable. If sampling health clinics or schools for example, it is possible that when sampling in different areas that are economically diverse in terms of household income, a lot of variation between clusters in variables of interest such as health and education outcomes may be driven by these income differences – a higher ICC might be expected in such cases. This compares to a case where the health clinics or schools were taken from areas that were similar from a socio-economic point of view –a lower ICC would be expected comparatively in this case.

The ICC chosen should ideally be based on data, and if existing data-sets are available it can be calculated with the *loneway* function on the statistical software package Stata for example. If no data on the expected ICC is available then some social sciences researchers suggest a default value of 0.1 for the ICC, but it will depend on the type of variable, the type of cluster, and the geographical variation for the intervention and comparison groups. Similar evaluations should be assessed to see what ICC may be expected in the given context for the variable of interest.

Final effect size if continuous variable

The final effect size provides the result of the calculation. If using a continuous variable, the effect size calculated will represent the number of standard deviations of impact expected. This will be the 'minimum detectable effect' (MDE) achieved given the values for sample size, clustering, significance level and power you have entered.

In some circumstances the entry may read n/a (not applicable) - this will mean that it is not possible to compute an effect size for the variables you have entered. Usually this will be because the sample size is too small to detect any result given the values you have entered.

Final effect / proportions if binary variable

The final effect size for a binary variable presents a proportion (Proportion 2). The difference between Proportion 1 and Proportion 2 will be the 'minimum detectable effect' (MDE) achieved given the values for sample size, clustering, significance level and power you have entered.

If you have selected a two-sided test, there will be two options for the second proportion, one will be higher than Proportion 1, the other will be lower than Proportion 1, reflecting that the change could be in either direction.

In some circumstances the entry may read n/a (not applicable) - this will mean that it is not possible to compute an effect size for the variables you have entered. Usually this will be because the sample size is too small to detect any result given the values you have entered.