



Coalition to Stop the Use of Child Soldiers

4th Floor, 9 Marshalsea Road, London SE1 1EP

Tel/Fax: +44 (0)20 7367 4110/4129

Email: info@child-soldiers.org Web: www.child-soldiers.org

Registered as a limited company (no. 4458380) in England

Registered Charity no. 1095237

This document is part of the Coalition's psychosocial web page. For more information on the psychosocial impact of armed conflict upon children go to: <http://www.child-soldiers.org/psycho-social/psychosocial>

Glossary of Statistical terms

Binary variables – see: **Variables**

Binary logistic regression – see: **Regression**

Case-Control studies: investigate a particular event, condition or phenomenon by studying one or more groups that have the condition of interest (the 'case' group(s)) and comparing them with those that do not (the 'control' group(s)).

This method has been used frequently in epidemiological research – where factors affecting the health or illness of populations have been studied and risk factors for given medical conditions identified. An example would be the well known epidemiological study by Sir Richard Doll and colleagues (1950) that established a significant association between smoking tobacco and lung cancer, through the comparison of rates of cancer in smokers (the case group) vs. non-smokers (the control group).

While case-control studies are powerful tools in public health and disease study and prevention, they cannot *in of themselves* establish *causal* links between risk factors and outcome. Different methodologies are required to establish such links. In order to demonstrate the likelihood of a causal link between tobacco and lung cancer, researchers needed to undertake animal studies and also prospective human studies (comparing smokers and non-smokers before they developed the disease and then following them up over time).

Case studies: an intensive, in depth, study of a single person, group or event – i.e. the 'case'.

Case studies may be used to gain a deeper understanding of a particular event or condition. As a research strategy, case studies can be used with single or multiple cases. They can be quantitative (gathering numerical data) or qualitative. The approach uses selective sampling rather than random sampling, choosing participants on the basis of prior information about their condition.

Traditionally, case studies have been regarded as most useful for generating hypotheses that other methodologies can then test more rigorously. It has also been argued that the case study methodology does not facilitate generalizing from an individual case, with a given condition, to the larger population with that condition. However, case study designs have become increasingly popular within the social sciences and within professional training due to their strengths in fostering a deeper understanding of phenomena. Single case studies also have the power to discredit existing theories - as illustrated in Karl Popper's famous assertion that the observation of one black swan would falsify the proposition that "all swans are white". Analogously, Aristotle's theory of gravity was challenged by Galileo's single experiment showing that the weight of an object was not a factor influencing its rate of fall from a height. See *also*: **Random Sampling**.

Chi- Square Test: a statistical test used with categorical or frequency data that have been entered into a contingency table, the simplest form of which would be a 2 x 2 table – i.e. two rows by two columns, giving rise to four cells.

For example, in a study of smoking, we might want to ascertain whether males were more or less likely to smoke than females. To check this, using a contingency table analysis, a simple 2x2 contingency table could be constructed by entering into the first table row all the males in the study, with Cell A, row 1, containing the number of male smokers and Cell B, row 1, containing the number of male non-smokers. Row two would contain all the females, with Cell C containing the number of female smokers and Cell D containing the number of female non-smokers. The row totals would, therefore, give us the number of men (Cell A + Cell B) and women (Cell C + Cell D) in the sample. The column totals would give us the number of smokers (Cell A + Cell C) and the number of non-smokers (Cell B + Cell D).

The Chi Square test is used to test whether the observed data within each table cell (the number of men and women who do or do not smoke) differs from that which would be expected under the null hypothesis, i.e. the hypothesis that the values of the *observed* data *do not* differ significantly from the values of the *expected* data. In the example above, the null hypothesis would be that smoking is *not* influenced by the sex of the smoker – and thus the *observed* values (the numbers of men and women who do or do not smoke) should *not* differ greatly from the values we would expect in each cell if there was indeed no real difference between the two groups. If the deviations between the values of the observed data and the expected values are large it suggests that these deviations are unlikely to be due to chance alone.

The Chi Square statistic is based on the probabilities of occurrence and non-occurrence in each of the groups. (Arithmetically, it is the sum of the squared difference between observed and expected data, divided by the expected data in all possible categories). Once the Chi Square statistic has been calculated, its value is compared to values contained in a probability table. If there is a low probability of a Chi Square value of this size occurring, then the null hypothesis is rejected and it is assumed that a factor other than chance alone is influencing the observed results.

(NB: As the Chi Square statistic becomes inaccurate with two by two tables containing small sample sizes, a statistic called the *Fisher's Exact* test is used for two by two contingency tables when sample sizes are small.

Co-efficient: the numerical value used to modify a mathematical expression (usually via a multiplicative process) or to define a relationship between variables..

For example, in correlation analyses, the *correlation co-efficient* would be a number between -1 and +1 that measures the degree to which two variables have a linear relationship. Where variables have a positive linear relationship (as one increases, so does the other), then the correlation will be between 0 and 1, depending on the strength of the correlation. A perfect positive relationship would give a value of 1. Where there is a negative linear relationship (as one variable increases, the other decreases), then the correlation will have a negative value up to -1. A correlation co-efficient of 0 indicates that there is no relationship between the variables. See *also*: **Correlation**.

Confidence intervals: Confidence intervals relate to the *magnitude* and *precision* of an observed effect size (see **Effect Size**).

For example, if researchers were measuring treatment effects in a drug trial administered to participants who had previously suffered a cardiac arrest, they would want to ascertain three

things. First, whether the drug treatment had had any effect. Secondly, what the size of any effect might be. Thirdly, how precise the observed effect size is. Confidence intervals address the last two questions.

However rigorously samples are chosen for treatment trials, the size of any treatment effect can only be an *approximation* to the *true* effect size. To estimate the latter, we would in theory need to include everybody in the population of interest – in our example, everybody who had ever had a cardiac arrest. This is clearly not a realistic possibility. Instead confidence intervals can be statistically constructed around the measured effect size.

Confidence intervals have an upper and a lower limit (the *confidence limits*), and thus provide a range of values within which the observed value lies. For example, if an effect size of 47, with a confidence interval of +4 to -4 is reported, then it has been calculated that the *true* effect size lies within the range of 43 to 51. (Sometimes, these limits are colloquially called the 'margins of error').

More technically, the confidence level represents the calculated statistical probability that, given an observed effect size, the *true* effect size falls within a particular range of values that lie around that observed effect size – i.e. within the lower and upper limits of the calculated confidence interval.

Having constructed confidence intervals, it is then necessary to estimate the level of confidence that can be placed on the probability that the observed effect size *does not* fall within the constructed confidence interval. This is similar to hypothesis testing whereby probability values (p) for a given result are calculated (see: **Statistical significance**). A confidence interval with a 95% *confidence level* indicates that there is a 5% chance that the true effect size does *not* fall within the specified confidence limits.

Why use confidence intervals? They are used because they convey more information than a probability value alone – they indicate how large or small the *true* effect might plausibly be. In general, the *smaller* the range of the confidence interval, the more precise the observed effect is likely to be. A *wide* interval implies poor precision as there is a wider range of values within which the *true* value lies.

(When examining reported confidence intervals, look to see whether the confidence interval includes a value indicating that there is 'no difference' or 'no effect'. That is, for a confidence interval for a difference, see whether the confidence interval includes zero; with a confidence interval for a ratio, look to see whether that interval contains one. If these '*null*' values are present then this implies that there has been no statistically significant change. Note also that the *confidence level* will be influenced the size of the study sample because the precision of estimation increases with larger sample sizes, giving rise to narrower confidence levels). To help understanding of Confidence Intervals, see *also* **Effect size**.

Control group: the term used in clinical trials to refer to those participants who do not receive the active treatment that is being studied.

Control participants may be chosen randomly from the general population, or they may be matched to those receiving the treatment on variables thought likely to affect outcome, such as, e.g., age, sex or social class. For instance, given that reading skills are likely to be affected by the age of the child, in a study examining the effects of an experimental reading intervention, it would be expected that the case and control group children would be matched on age, so that we are comparing children of roughly equivalent age - otherwise age would be a *confounding* variable. That is, while the reading intervention may be having

an effect, the outcome is also likely to be being affected by the child's age. By matching the children on age, we can control for the effect of age. See also: **Case-Control studies**.

Controlling variables: when the effect of a *confounding* variable is taken into account in statistical analyses.

For instance, a researcher measures a group of children and finds that boys on average weigh more than girls. Is this weight difference a straightforward effect of the child's sex, or are there other variables also likely to be influencing the outcome? One such variable could be the height of the child as we know that, on average, boys tend to be taller than girls – and taller children tend to weigh more than shorter children. In order to establish whether sex has an effect on the child's weight (in addition to the effect of the child's height), it is necessary to control for the child's height. To do this a statistical analysis is performed that estimates the effect of the predictive variable of interest (in this case, sex of child) whilst holding constant the effect of another predictor variable (in this case, height). The statistical procedure allows us to ask: if boys and girls had the same height, what would the difference in weight be? Another research strategy for attempting to avoid the influence of known confounding variables is to use matched samples. See also: **Variables; Matched Samples**.

Control trials: Clinical trials or studies that employ control groups. See also: **Control Groups**.

Correlation: A statistic that reflects the extent to which two or more variables are associated with one another. A positive correlation is where as one variable increases, so does the other – e.g. weight usually increases as height increases. A negative correlation is where as one variable increases, the other goes down.

The statistic measuring an association between variables is called the *correlation coefficient* (see: *coefficient*). There are differing ways of measuring correlations, including *Intraclass Correlation Coefficients* (ICC), the *Pearson product-moment correlation coefficient*, and the *Spearman rank order correlation*. Researchers choose between these statistical procedures according to various characteristics of the data and the sample size. See also: **Pearson product moment correlation and Spearman Rank Order correlation**.

Cronbach's Alpha: a statistical measure of the internal reliability (or consistency) of a measuring instrument or test. The statistic is calculated on the basis of repeated measurements of split-half reliability (see: *reliability – internal reliability*). Strong internal consistency on a test is shown by moderate correlations between test items (0.70 to 0.90). If correlations between items are too low, then it is likely that they are measuring different traits rather than the same trait. If correlations are too high, then it is likely that some of the test items are redundant and could be removed from the test.

Cross-sectional study: a study that examines relationships between independent and dependent variables at one particular point in time. As both variables are measured at the one time, these studies are limited in their ability to demonstrate cause-effect relationships. See also: **Variable**.

Cut-off threshold – see: Threshold

Dependent variable – see: Variable

Effects: The measured effect of one variable upon another. Two types of effects are usually referred to, *main effects and interaction effects*. A 'main' effect analysis examines the effect of one or more independent variables upon a dependent variable. An interaction effect analysis examines whether two independent variables have a combined effect on the

dependent variable that is more than just the sum of their individual main effects on that variable.

For instance, we might want to look at whether both exercise and diet have an effect upon weight loss when we are examining the effects of two different diets and two levels of exercise. The results indicate that diet alone successfully results in weight loss (a *main effect* for diet), as does exercise alone (*main effect* for exercise). Those who follow diet A lose an average of 2 lbs; those who follow diet B lose an average of 4 lbs. Those who follow a low exercise regime lose 3lbs, those who follow a high exercise regime lose 6lbs. We might expect that the effects of these two types of intervention would combine in an additive way – that is, those individuals who followed both Diet B and the high exercise regime would lose in the region of 10 lbs. In fact, we find that this group of participants actually loses 15lbs – they got the effects of both Diet B and the effects of high exercise plus a bonus – that is, the effect of diet B was larger where participants followed a high exercise regime. This is an interaction effect: the *main effect* of diet B, the *main effect* of exercise level and the ‘bonus’ of the effect of Diet B in the context of the level of exercise undertaken.

Effect Size: quantifies the size of a relationship between variables (e.g. the strength of correlation) or the size of the difference between two or more groups on a particular outcome measure (e.g. weight loss following a dietary intervention). When examining differences between groups, it is a standardized, scale-free measure of the relative size an intervention effect that takes into account the overlap between group scores. That is, it takes group variance into account (see: **Variance**).

To illustrate the usefulness of this, we could consider an example where we want to compare two types of remedial intervention for reading difficulties in a group of same aged children who were randomly allocated to either one of two equally sized groups. In group A, the children receive remedial intervention type I and in group B they receive remedial intervention type II. A simple comparison of the group average (or ‘*mean*’) scores after the intervention could allow us to see whether either intervention had had an effect, and what the size of that effect was. We might deduce from any difference in group means that one intervention was more effective than another – if Group B had a higher post-intervention mean score than group A, and that difference was fairly large, we might conclude that remedial intervention type II seemed more effective than type I.

However, a simple comparison of mean scores does not tell us whether every child in group B did better than every child in group A, or whether just a few children in group B did very well indeed, while most children in the group performed at similar levels to the majority of children in Group A. If that were the case, we might be less certain of the real effect of the type II intervention.

It would be helpful, therefore, to know to what extent did the scores between the two groups overlap? If there were few overlapping scores between the groups, with the majority of the children in group B doing better than the majority of the group A children, then the effect of the type II intervention would seem a substantial one. If, on the other hand, there were a lot of overlapping scores in both groups, then the effect of the type II intervention would seem more limited.

The effect size is particularly valuable for comparing the magnitude of the intervention effects of different treatments, whilst taking such group variance into account. It is calculated by subtracting the mean score of one group from another group’s mean score and then dividing by the *standard deviation* (a measure of the spread or variance of the scores). The emphasis is on the *size* of the treatment effect rather than on the statistical significance of any group outcome differences. The effect size statistic has the advantage that, unlike

measures of statistical significance, it does not conflate effect size and sample size (see: **Statistical significance**).

Cohen (1988) provided rules of thumb for characterizing whether effect sizes would be regarded as 'small', 'medium' or 'large' – thus, an effect size of 0.2 qualifies as 'small', of 0.5 as 'medium' and 0.8 as 'large', although he recognized that the meaning of these effect sizes would vary according to the specific area of inquiry. Further, the effectiveness of interventions can only be interpreted in relation to other interventions that seek to produce the same effect. Thus, in interpreting their findings, researchers will have to exercise a degree of value judgment as to the practical or clinical importance of the recorded effects.

Epidemiology: in human populations is the study of how, when and where diseases occur. Epidemiologists attempt to identify what are the *risk* factors for the occurrence of a disease, and what factors may *protect* populations against disease (*protective factors*). For example, an epidemiological strategy was employed to examine the possibility of smoking being a risk factor for cancer, whereby rates of lung cancer in smokers were compared with the rates of cancer in non-smokers. While an epidemiological strategy can demonstrate a correlation between a risk factor and a disease outcome, it cannot establish that the risk factor actually *causes* the disease. Establishing such a causal link would involve a rigorously designed experimental study in which one group was exposed to the risk factor and another group was not (with exposure to the risk factor being the only difference between the two groups). As ethical issues would rule out this type of design with human participants, such studies are usually conducted with animals (as in the case of establishing the causal link between smoking and cancer) – though this strategy, of course, raises another set of ethical questions! In general, the higher the correlation between a risk factor and the adverse outcome, the more certain is the association between the two.

Fisher's Exact Test: A statistic used in contingency table analyses with small sample sizes. For details of contingency table analyses, see: **Chi Square Test**.

Independent variable – see: Variable

Interaction effects – see: Effects

Main effects – see: Effects

Matched samples: see: Samples

Mean value: is the average value – calculated by adding all the sample participants' scores on a measure and dividing the derived total by the number of participants.

Meta-analysis: is a statistical technique used to amalgamate, summarize and review previously undertaken quantitative research studies that have all examined a particular topic. It could be used to compare, for instance, whether medication for a given condition is more or less effective than a clinical therapeutic intervention or which of several types of medication appears to be the most effective. The procedure employs a statistical analysis that summarizes and compares the *effect size* of each type of intervention in the studies considered. In essence, this is equivalent to combining all the research on one topic into one large study with many participants. However, the utility of such analyses will depend on the methodological rigour of each study that is included and sometimes differences between studies - for example, in how outcomes are measured - can make it hard to interpret the final results in a meaningful way. See also: **Effect Size**

Null hypothesis: The hypothesis that the values of the *observed* data *do not* differ significantly from the values of the *expected* data. That is, there is no true difference or effect of the independent variables being studied. For an example, see **Chi Square Test**

Odds ratio: Odds can be regarded as a measure of probability. They are derived by calculating the probability that an event will occur divided by the probability of that event not occurring.

For example, the odds that a single throw of a dice will produce a six are given by the probability of throwing a six (1) divided by the probability of not throwing a six (5), i.e. 1/5 or, in betting terminology, one in five. (In this calculation, the assumption is being made that on any one throw of the dice, each number from 1 to 6 has an equal probability of occurrence).

Odds ratios are used when comparing the probabilities of an event occurring in different groups - for example, the odds that one group of people exposed to a risk factor for a disease (group A) develop that disease as compared to the odds of another group of people, who have *not* been exposed to that risk factor (group B), developing that disease. The odds ratio for this would be calculated by dividing the probability of the occurrence of the disease in group A by the probability of occurrence of the disease in group B. The odds ratio is used in logistic or binomial regression, in which a dichotomous outcome is predicted by one or more variables. See *also*: **Regression analyses**

Non-controlled design: studies that employ systematic data collection, but do not follow a case-control design. See *also*: **Control groups; Case-Control studies; Randomized Control trials.**

Pearson product-moment correlation: see *also*: **correlation and co-efficient.** A statistical procedure that produces a correlation coefficient measuring the association between two variables, giving a value between -1 and +1, inclusive. It is usually denoted by *r*. It measures the linear relationship between two variables that have been measured on interval or ratio scales (eg height in inches and weight in pounds). It is assumed that the data are normally distributed (most values lie around the mean, with relatively few extreme scores at each end of the distribution, producing a Bell shaped distribution curve).

Peer review: a refereeing process whereby an author's research or academic work is subjected to the scrutiny of others who are experts in the same field and who are considered capable of giving an impartial review. Research publications that have not undergone the process of independent peer review are not regarded as highly as those that have done so.

Pilot study/pilot testing: involves undertaking a preliminary, small scale, project in order to assess whether data collection tools and procedures work effectively in the real world. This allows potential problems in their use to be identified and corrections or adjustments to be made before the real study commences.

Prevalence: In statistics, prevalence refers to the number of cases of a condition that are present in a particular population at a given time.

Quasi-experimental design: refers to research studies where the conditions required for a randomized control trial are not fully met as, for example, when groups of participants are chosen on the basis of availability rather than randomly selected from the general population of interest.

An example of the need to work with available participants is given by research into the outcome of former child soldiers where participants can only be selected when they are prepared to identify themselves as formerly belonging to an armed group – the researcher has no access to the general population of child soldiers which is in reality unknown. If those who are prepared to identify themselves differ systematically in some way from those who are not, then the study findings will reflect this limitation – for example, former girl soldiers may be less likely to want to identify themselves than former boy soldiers due to the social stigma attaching to girl soldiers as a result of the sexual abuse many have suffered.

Within the social sciences, quasi-experimental designs are used frequently as double blind randomized controlled trials are simply not feasible in many situations. Ethical considerations may also mean that a quasi-experimental design is preferred. For example, if one wanted to study the effect of maternal alcohol consumption on later child development, it would not be ethically acceptable to allocate women randomly to treatment and control groups and then administer them alcohol. Instead, a quasi-experimental design would be used – for instance, women in their last trimester of pregnancy could be asked about their consumption of alcohol throughout the earlier months of their pregnancy, and then divided into study groups on that basis. As long as the shortcomings of the design are recognized, such studies can remain powerful tools in social investigations. Quasi-experimental designs tend to be less time consuming and expensive than randomized control designs. They are a good way to obtain a general overview, or to identify general trends, that that can then be examined in more detail in, e.g., single case or qualitative studies. See also: **Randomized control trials.**

Randomization: refers to a process by which participants are chosen at random from the larger population of interest for entry into a study, so giving each person on the larger population list an equal chance of being chosen for inclusion.

A sample chosen randomly frees the selection procedure from any inherent bias deriving from the researcher or the selection process. Given a sufficient sample size, it also ensures that the chosen sample is representative of the larger group from which it was drawn. For example, if you have a list of 1,000 potential participants and randomly select 100 people as participants, the choice will be unbiased and the chosen participants are likely to be representative of the list as a whole.

By contrast, if you take the first 100 people off the list it will be possible that they share a common factor that is not equally distributed throughout the remainder of those on the list and that may have a bearing upon the results. For instance, recent research suggests that August born children who enter their first year of school in the September immediately following their fourth birthday do less well academically throughout their school life than those children in the same year group at school who were up to ten months older at their time of school entry. We could imagine a situation in which groups of children are chosen from each year group in different schools to take part in an intervention to improve their mathematical skills. Children in school A will receive one intervention, those from school B another. If school A ordered its class lists on the basis of birth date (youngest first), whilst School B ordered theirs alphabetically, and the researchers chose the first ten children from the class lists within each year group throughout each school, then the sample from School A could well have an overrepresentation of August born children – which could potentially bias the results of the study.

The term ‘randomized selection’ generally refers to the process of selecting people to enter into a study, while ‘random allocation’ generally refers to the allocation of participants to different study groups subsequent to their entry into the study. A Cochrane review suggests that medical trials that did not use concealed random allocation of participants to study groups, resulted in findings that were biased in unpredictable ways.

See: <http://www.cochrane.org/reviews/en/mr000012.html>; last assessed as up-to-date, February, 2007; last accessed February, 2011).

Randomized control trials: are considered the most rigorous way of determining whether a given treatment or intervention has a causal effect upon patient outcome. In relation to mental health, for example, one might ask whether Cognitive Behavioural Treatment (the '*intervention*') is more effective in decreasing symptoms of depression than other types of existing treatments. In such a treatment trial, the extent of the decline in depressive symptoms would be regarded as an '*outcome*' measure. Those who take part in such treatment trials are known as *participants*.

Randomized control trials have several important and distinctive features. First, participants are chosen *randomly* from the population of interest and then allocated *at random* to either a '*treatment group*' (where participants receive the 'new' treatment of interest) or a '*control group*', where participants will receive an alternative intervention of some kind. For example, they might receive a placebo treatment (a sham intervention such as a sugar pill), genuine medication, or another standard treatment already available. Random allocation ensures that there are no systematic differences between the intervention groups in known or unknown factors (such as age, sex, previous medical history) that could affect participant outcome. Second, in order to try to eliminate conscious or unconscious bias amongst participants, they should be unaware of whether they are in fact receiving the new, experimental treatment or the substitute treatment until the trial finishes. Where this condition holds, the trial is regarded as a 'single blind trial'. Where those conducting the trial are also 'blind' as to who is receiving the 'new' or 'control' treatments, the 'double blind' condition holds. Double blind trials ensure that the preconceived views of subjects and clinicians as to the efficacy of particular treatments or other treatment factors, cannot systematically bias the assessment of participant outcomes. Thirdly, except for the given variations in treatment type, all participants in both groups should be treated in the same way in order to ensure that there is no other systematic variation in research conditions that could influence the trial results.

The methodological rigor of randomized control trials means that they are often referred to as the 'gold standard' for the assessment of interventions. While their strengths are clear, they also have a number of limitations. The choice of treatments can arouse ethical concerns – for example, when control group participants would be required to receive a treatment that is regarded as 'inferior' to the new treatment, or when the trial design appears to involve withholding effective treatments from some participants. There may also be practical difficulties in trial design – it may not in reality be possible to select participants randomly from a larger population, or to keep investigators 'blind' as to which research groups participants belong. Randomised control trials also tend to take longer and to be more expensive than other, simpler interventions.

Regression: Regression is a statistical procedure used to determine and measure the predictive relationship between one (dependent) variable and other (independent) variables. That is, it indicates the extent to which you can predict some variables by knowing others. Regression procedures could be employed, for instance, when wanting to ascertain whether exam results are influenced by, say, the age and/or the sex of the exam takers. More specifically, regression analysis helps us understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. See also: **Variables – independent, dependent and controlling variables.**

Ordinal regression: a form of regression analyses used when the dependent variable is ordinal (e.g. exam results graded A-F) and where the independent variables can be categorical (e.g. male or female) or continuous (e.g. age).

Binary (or binomial) Logistic regression: a form of regression analyses used when the dependent variable is dichotomous or binary (e.g. 'yes' or 'no', or '0' or '1').

Reliability: refers to the *consistency* of a measure – the extent to which we get the same result repeatedly when employing that measure with a given individual under the same conditions. For example, when using a test to measure, say, a psychological trait in an individual on two separate occasions, we anticipate that a *reliable* test will give approximately the same result each time. If the test indicates that the individual is an 'introvert' at time one, we would expect that same individual to be classed as an 'introvert' at time two. Reliability cannot be measured precisely – it can only be estimated. The main types of reliability measurement used in the social sciences are:

Test-retest reliability: assesses the consistency of a given measure from one time to another. The shorter the time interval between the two assessments, the higher the consistency of the test is likely to be.

Inter-Rater or Inter-Observer reliability: assesses the degree to which different raters or observers give consistent estimates of the same phenomenon.

Internal consistency reliability: assesses the consistency of results across different items from within the *same* test. The reliability of the measuring instrument is judged by estimating the extent to which items measuring the same construct yield similar results. One frequently used statistical measure of internal consistency reliability is Cronbach's alpha. See: **Cronbach's alpha**)

A variety of internal consistency measures can be used: e.g. *average inter-item correlation; split-half reliability* etc. In the latter, all items measuring the same construct in an instrument are randomly divided into two sets. The whole instrument is then administered to the sample population and subsequently the total score for each randomly divided half is calculated. Reliability is assessed by calculating the size of the statistical correlation between each half of the instrument. Strong internal consistency is shown by moderate correlations between items (0.70 to 0.90). If correlations between items are too low, then it is likely that they are measuring different traits. If correlations are too high, then it is likely that some items are redundant and could be removed from the test.

Parallel-Forms reliability: is assessed by comparing the consistency of measurement of two tests constructed in the same way from a larger set of questions that are regarded as measuring the same construct. One way to construct parallel forms would, for example, be to generate a large set of questions measuring the same construct and then devising two measuring instruments by randomly dividing these questions into two sets. Reliability is judged by the extent to which these parallel forms give consistent results when administered to the same population.

Reliability co-efficient: See: **co-efficient; reliability**

Repeated measures: Where the same variable is measured more than once for each participant over time, or under difference circumstances. For example, where blood pressure is measured before participants undergo an exercise, and then after they have completed the exercise task.

Sample: The participants selected for study. Participants are usually chosen from a larger 'target population' which is the entire group about which the researcher wishes to draw conclusions.

Matched samples: Where participants in the groups being studied are matched on the basis of one or more criteria - for example, their IQ levels or birth order. This strategy is employed to try to eliminate the effect of *confounding variables* (i.e. other variables that could also be responsible for any observed effect in the outcome variable). See also: **Controlling variables.**

Randomized sample: see: **Randomization and Randomized control trials.**

Stratified sample: when the sample of participants is drawn from the population of interest that has first been divided into non-overlapping classes (or strata) from which participants are then drawn at random.

The number of participants drawn from each strata is proportional to the number of people within each. Strata can be based on variables such as social class, age or geographical locality. The method works best when variability *within* the strata is minimized, when variability *between* the strata is maximized, and the variables used to stratify the population are strongly correlated with the dependent or outcome variable.

One advantage of this type of sampling is that it allows researchers to make inferences about specific subgroups of participants in a way that more generalized random sampling would not.

Sensitivity and specificity: refers to the accuracy with which tests can establish the presence or absence of a given condition.

Researchers wishing to identify those with a given disease, want to know the extent to which their test for that disease is 'sensitive' to its presence, i.e., the extent to which it accurately identifies those people who truly have the disease ('true positives'). Tests in reality cannot be 100% accurate, and will give some 'false positives' – i.e. they will indicate that someone has a particular disease when, in fact, they do not. They can also give 'false negative' results – i.e. when the test says that someone does not have the disease when, in fact, they do.

The *sensitivity* of the test refers to the extent to which a test can accurately identify those who truly *do* have the disease. Arithmetically, the sensitivity of the test is calculated by the number of 'true positives' (TP) divided by the sum of the number of 'true positives (TP) + the number of 'false negatives' (those 'true positives' that the test falsely identifies as not having the disease – FN). That is: $TP/(TP+FN)$. In other words, the calculated sensitivity equals the number of people with the disease who got a positive test result divided by the total number of people with the disease.

The *specificity* of the test refers to the extent to which the test accurately identifies those who truly *do not* have the disease. Numerically, this is calculated by dividing the number of 'true negatives' (TN) by the sum of the number of 'true negatives' (TN) + the number of 'false positives' (i.e. those 'true negatives' that the test falsely identifies as having the disease - FP). That is: $TN/(TN+FP)$. In other words, this is a measure of how well the test excludes people if they are healthy – it is the proportion of people without the disease who get a negative test result.

Spearman Rank Correlation Co-efficient: **see: correlation and coefficient.** This statistical procedure is used to measure the association between ranked variables. It can be used when the data are not normally distributed. It can also be used when the relationship between two variables is non-linear.

Split-half reliability – see: reliability – internal reliability.

Standard Deviation – *this definition will be more easily understood if **Variance** is consulted first.* The standard deviation is the most commonly used measure of the *spread* or *variance* in a set of values.

Mathematically, the standard deviation is the square root of the variance. There is an arithmetic reason for needing to find this square root. When the variance is being calculated it is necessary to square the difference of each observed value from the mean value in order that positive and negative values of the same size do not cancel each other out.

For example, two scores can have a difference of 1 from the mean, but one score is 1 above the mean (+1), while the other is 1 below the mean (-1). If these were added, they would sum to zero, indicating that there was no variation around the mean! Squaring these difference scores gets rid of this problem ($-1 \times -1 = +1$). To correct for this squaring process, however, once the variance has been calculated its square root is found. This square root is known as the standard deviation. As with variance, the larger the standard deviation, the more spread out is the distribution, while smaller standard deviations reflect distributions clustered closer to the mean.

In normally distributed data sets (where most values lie closer to the mean, with fewer scores at the extremes – as for example, in population distributions of height and weight values), we know that about 68% of scores will lie within 1 standard deviation of the mean score, while about 95% of scores will be within two standard deviations of the mean. See *also: Variance.*

Statistical significance: Levels of statistical significance indicate *the likelihood* that the null hypothesis is true - i.e. *that there is no true difference or effect* of one or more variables upon the outcome that we are measuring.

A significance level of 0.05 indicates that there is only a 5% chance that the null hypothesis is true. To put it another way, if the variables we are examining really **do** have no effect upon the outcome that we are interested in, then these results are pretty unlikely to have arisen just by chance.

The smaller the p-value obtained, the more *unlikely* it is that there is *no* effect – and the *more* likely it is that there *is* an effect.

In general, within the social sciences, a p value of 0.05 is the minimum level taken to indicate that it would be reasonable to assume that any differences found **do** reflect a genuine difference between the groups or variables of interest. If statistical testing produces a p-value of less than 0.05, then in these cases, we say the result is 'statistically significant' (*i.e. it is unlikely to have arisen by chance*). Note, though, that sometimes different significance levels are used (e.g. 0.01).

An example is given in the following data set, where researchers pose the following question:

Is the consumption of gasoline X affected by the area in which respondents live, or the type of vehicle that they drive? Respondents to this question provide the data found in Table 1, where gasoline consumption is considered in relation to the area in which respondents to the question live, and table 2 where consumption is considered in relation to the vehicles that respondents drive.

Table 1 Area effect

Consumes gasoline X	City		Suburb		Total
	N	(%)	N	(%)	
Yes	215	(60)	213	(61)	428
No	146	(40)	139	(40)	285
Total	361		352		713

Chi Square: 0.07 p= 0.795

Table 2 Vehicle effect

Consumes gasoline X	Type of vehicle					Total
	Car	Truck	Bus	Van		
	N	(%)	N	(%)	N	(%)
Yes	131	(53)	74	(49)	29	(66)
No	116	(47)	76	(51)	15	(34)
Total	247		150		44	

Chi Square 24.4 p=0.001

(data taken from: <http://www.surveysystem.com/signif.htm> - last accessed, February 2011)

Looking at table 1, we see that the area in which respondents live appears to exert little effect on the consumption of gasoline X – it is used by 60% of city dwellers and 61% of suburban dwellers. This is reflected in the p value of 0.795 – indicating that there is a 79% likelihood that there is no true difference between the two groups (i.e. that the null hypothesis is true).

In contrast, if we look at table 2 we see that the type of vehicle driven does appear to have an effect – while respondents who drive cars or trucks are roughly evenly divided between those who do and do not use gasoline X, the proportions of bus and van drivers who do use gasoline X are substantially larger than the proportions of bus and van drivers who do not. The small p value of 0.001 indicates that there is only a 1% likelihood that this pattern of consumption of gasoline X according to the type of vehicle driven is due to chance.

However, it is important to remember that if a large number of statistical tests of significance are applied to one data set, the probability that false positive results will be found is raised. If 100 tests are conducted that each show a p value of 0.05, it is likely that five of these tests will give a false positive result – although we would not be able to identify which of these particular tests do so. To counter this, researchers sometimes use smaller cut-off points as indicating statistical significance (e.g. p=0.005).

It is also worth noticing that the size of the sample employed has an effect on significance levels – the likelihood that a small difference between groups will be statistically significant increases as the size of the sample increases.

Finally, a statistically significant difference between, say, two groups, is not synonymous with an 'important' difference. Whether a difference is important, or clinically significant, is a judgement that the clinician, researcher and/or reader must make.

Two sided (tailed) tests of significance: researchers comparing two groups often start with a null hypothesis (i.e. that there is no true difference or effect of the independent variables being studied). If this hypothesis is not true, then the alternative hypothesis must be true – i.e. there is a difference. In either case, neither of these hypotheses specify a direction of effects – e.g. in a treatment trial a direction of effect would be specified if the researcher a priori hypothesized that treatment is better than no treatment.

Statistical tests of significance that can be employed with data sets where no direction of effects has been specified (e.g. the scores of group A might be higher **or** lower than those of group B) are known as two-tailed tests. When researchers a priori hypothesise that there will be a treatment effect in a given direction (e.g. group A **will have** higher scores than group B), a *one-tailed test of significance* could legitimately be employed. Two-tailed tests of significance are generally regarded as a more rigorous ways of assessing significance levels than one-tailed tests.

Target population: the larger group in the general population who match the main characteristics of the group being studied, and to which the researcher wishes to generalize conclusions.

For instance, if the effects of a given medication have been studied in males between the ages of 25-40 following an initial heart attack, then the *target population* would be males of a similar age who meet the same general conditions as those in the study.

Threshold: a term used to identify the point beyond which the condition of interest is said to be present.

For instance, when conducting a mental health assessment, a psychiatrist may decide on the basis of the person's severity and type of presenting symptoms, that they cross the threshold used to define clinical depression and so can be diagnosed as 'clinically depressed'.

Multiple thresholds may be used. For example, if a reading test that classified *poor readers* as those children scoring less than 15 on a standardized reading score, and *average readers* as those scoring 15 and above, then 15 would be regarded as the threshold score for the *average* reading level. It is the minimum score that the reader must earn in order to be classified as an *average* reader. Other points on the reading score scale could define thresholds for *above average* and *gifted*. The terms *Cut-off* and *threshold* merge when used to indicate a score on a standardized assessment scale at which criteria for a given condition are met, and these terms are often used interchangeably within the literature.

Two sided (tailed) tests of significance: see: Significance

Validity or Construct validity: refers to the extent to which a test, or measurement tool, measures what it purports to measure (i.e. the construct under examination). There are various measurements of validity. Those most commonly referred to are:

Face validity: when an instrument or test appears to be able to measure what it sets out to measure. This is an intuitive, rather than a technical, assessment of validity – does the instrument 'look' valid to an outside observer with some knowledge of the field?

Content validity: a more technical assessment of the extent to which a measure represents all aspects of the construct under examination. For example, a measure of

trauma would lack content validity if it did not include both behavioral and affective components of trauma. While it may be relatively straightforward to establish content validity for some constructs, e.g. tests of mathematical reasoning, the process becomes more complex as one moves into psychosocial and cultural domains where there may be less agreement on the component parts of a construct.

Convergent validity: examines whether two measures of the same thing correlate with each other. An example would be the extent to which standardized test scores for mathematical reasoning correlate with success in national mathematics exams, or performance on a class assignment requiring the use of mathematical reasoning. In contrast, *divergent validity*: examines whether two measures of unrelated things correlate with each other – that is, do the measures discriminate between dissimilar constructs?

Criterion related validity: Whether a new measure correctly identifies groups defined by some other established procedure, such as a clinical diagnosis. For instance, to what extent do scores on a newly developed trauma scale successfully identify those who meet, and those who fail to meet, trauma criteria on the basis of a psychiatric diagnosis? There are two types of criterion related validity:

Concurrent validity: when the results of one measure are compared to the results of an established and valid measure of the same construct, at the same point in time (concurrently).

Predictive validity: the extent to which one variable, such as test scores, correlates with some future performance or variable

Variable: the category, event or object that you are trying to measure. This could be the temperature, the rate of rainfall, a clinical condition or a person's attitude. In research, variables are usually divided into dependant and independent variables, and researchers are often trying to explore the relationships between them. The types of variables often referred to include:

Binary, or dichotomous, variable: a discrete variable that has only two possible values – e.g. male/female; pass/fail; alive/dead.

Dependant Variable: a variable that depends on other things. For instance, an individual's speed of response in a given test (the *dependent variable*) may depend on the amount of alcohol that s/he consumed the night before. Usually researchers are looking for factors that will influence the dependent variable – e.g. what factors influence whether children suffer from trauma after exposure to stress or violence? *Dependent* variables are also often known as '*outcome*' variables.

Independent variable: in contrast to a *dependent* variable, an *independent* variable is regarded as one that is not changed by the other variables that are being measured, or one that the researcher directly manipulates. For instance, a person's age or sex is not changed by the amount of alcohol that they drank the night before. Researchers are usually considering whether a given set of *independent* variables influence the *dependent*, or '*outcome*' variable(s).

Ordinal variable: a variable in which the order of the data points can be determined but not the distance between the data points – for example, when exam results are graded by letter.

Continuous and discrete variables: for the purpose of statistical analyses, researchers need to consider whether quantitative variables can be regarded as 'discrete' or 'continuous'. Discrete variables are those with possible scores that represent discrete points

on a scale. For instance, the number of children in a family – these need to be measured by whole numbers; one can have one or two children, but not 1.5. In contrast, continuous variables are those measured on a scale that can be regarded as being continuous rather than being composed of discrete steps. Time and temperature are two such examples. By convention, we report the temperature in degrees, such as 30 degrees centigrade, but if we wanted to do so, we could measure it more precisely and report the temperature using decimal points – e.g. 30.534 degrees centigrade. Of course, the practicalities of measurement preclude most measured variables from being truly continuous.

Variance: measures of variance derive from the spread of values that arise whenever events or participant behavior is measured by researchers. For instance, in a study of children's reading skills, it is highly likely that children of a given age will vary in how skilled they are at reading and, therefore, they will obtain different scores on the same tests of reading skills. The variance tells us how far each individual child's score (or data point) is from the average (or mean) score calculated by summing all the individual scores and dividing by the number of children. (Technically, the variance is calculated by computing the average squared deviation of each number from its mean). The larger the variance, the more spread out the distribution is. The smaller the variance, the closer the values cluster to the mean. Understanding the variance aids in interpreting the meaning of the data (for a worked example of why understanding the data distribution can be useful, see: **Effect Size**). The variance is also important as many tests of statistical significance are based on assumptions about the spread, or distribution, of the data and their use is only valid if the obtained data distribution concurs with these assumptions. The most commonly used measure of spread is the *standard deviation* which is the square root of the calculated sample variance. See also: **Standard Deviation; Effect Size**

Acknowledgements:

Dr William Finlay was kind enough to read and comment on our Glossary of Statistical Terms. We are extremely grateful for his input, wise advice and for his sharing of his understanding and knowledge with us. If errors have crept into this document, they remain the responsibility of Dr Linda Dowdney, Editor.

Sources:**Case-control studies:**

See: http://en.wikipedia.org/wiki/Case-control_study

Reference for studies by Doll:

Doll R, Hill AB. Smoking and carcinoma of the lung. *BMJ* 1950;2(4682):739-48.

Doll R, Peto R, Boreham J, Sutherland I. Mortality in relation to smoking: 50 years' observations on male British doctors. *BMJ* 2004;328 (7455):1519

Link: [\[Abstract/Free Full Text\]](#)) – last accessed February, 2011.

Case studies:

For a robust defence of the utility and validity of case study research strategy, see:

Fyyvberg, B (2006) Five Misunderstandings About Case-Study Research. *Qualitative Inquiry* 12 (2) pp219-245.

Link: <http://flyvbjerg.plan.aau.dk/Publications2006/0604FIVEMISPUBL2006.pdf> -

Last accessed February, 2011.

Chi-Square Test:

For more details, and worked examples, see:

http://ccnmtl.columbia.edu/projects/qmss/the_chisquare_test/about_the_chisquare_test.html#Chi-Square_Test

<http://science.jrank.org/pages/1401/Chi-Square-Test.html#ixzz0ldBla7Qw> - last accessed February, 2011.

Co-efficient:

See: the following website for these and other definitions of co-efficient, and worked examples:

http://www.google.co.uk/search?hl=en&defl=en&q=define:coefficient&ei=nvCxS8q_EdH64Ab3tYHBAg&sa=X&oi=glossary_definition&ct=title&ved=0CAYQkAE - last

accessed February, 2011.

Confidence intervals For further details, see:

http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/What_are_Conf_Inter.pdf
www.childrens-mercy.org/stats/journal/confidence.asp - last accessed February, 2011

http://findarticles.com/p/articles/mi_m0689/is_12_52/ai_111614752/ - last accessed February, 2011.

Controlling variables: Based on: <http://www.sportsci.org/resource/stats/complex.html> - last accessed February, 2011.

Effect Size: Based on: Coe, R. (2002). It's the Effect Size, Stupid. What effect size is and why it is important. Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England 12-14 September, 2002.

Summary version: Coe, R. (2002) What is an Effect Size? A brief introduction. Available online at <http://www.cemcentre.org/evidence-based-education/effect-size-resources> - last accessed, February, 2011.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. NY: Academic Press

See also: <http://www.stanford.edu/~kcobb/hrp259/p-value%20comparisons.pdf> – last accessed February, 2011

Epidemiology: Based on: <http://pmp.cce.cornell.edu/profiles/extoxnet/TIB/epidemiology.html> - last accessed February, 2011.

Meta-analysis Based on <http://wilderdom.com/research/meta-analysis.html> - last accessed October, 2010

Quasi-experimental design: Based on Source:

Shuttleworth, Martyn (2008). Quasi-Experimental Design. Retrieved from Experiment Resources: <http://www.experiment-resources.com/quasi-experimental-design.html> - last accessed February, 2011.

Randomized control trials: Based on:
Randomised control trials reference: Sibbald, B. & Roland, M.. Understanding controlled trials: Why are randomized controlled trials important? BMJ 1998;316:201 (17 January)
see

<http://www.bmj.com/cgi/content/full/316/7126/201> - last accessed February, 2011

For more information on 'blind trials' see: http://en.wikipedia.org/wiki/Double_blind_trials - last accessed February, 2011.

Reliability:

Based on: <http://www.socialresearchmethods.net/kb/reotypes.php> - last accessed February, 2011.

Sampling see: [http://www.en.wikipedia.org/wiki/Sampling_\(statistics\)#Stratified_sampling](http://www.en.wikipedia.org/wiki/Sampling_(statistics)#Stratified_sampling) – last accessed February, 2011

Significance tests see: <http://www.bmj.com/cgi/content/full/309/6949/248> - last accessed February, 2011

Validity See: <http://www.medicine.mcgill.ca/stroking-assess/definitions-en.html> - last accessed February, 2011.