CHAPTER 1

Introduction and Background

In empirical research one finds that one is often interested in comparing groups with one another, or with determining the relationships between variables that have been measured. The significance of the difference between these groups, or the significance of the relationship between these variables, is then usually required. The term “significant” is usually understood to imply that a so called null hypothesis, stating that there is no difference between the means (or no relationship between the variables), is rejected at a predetermined level of significance (usually 5%). In other words: the so called “*p*-value” is less than 0,05. This type of “significance”, also known as “statistical significance”, really only means that the probability of the null hypothesis being incorrectly rejected is small (for example, $\leq$ 0.05). Therefore, it indicates that the differences or relationships found in the probability sample(s) are not due to simple coincidence, because the chance of it occurring coincidentally is small (say, 5%). However, what these statements do *not* say is how *important* the differences or relationships are. To determine the importance of the differences or relationships one can make use of *effect size indices.*

The purpose of this manual is to discuss and provide effect size indices for the majority of scenarios in which empirical researchers work.

1.1 Practical significance

As with statistical significance, whereby a null hypothesis is rejected, the question which needs to be answered is: when is a difference or relationship large enough to be considered important? Effect size indices can be used in the sense that these indices are directly proportional to the importance of the differences of means or relationship between variables. If an index is large enough, then the result is said to be *practically significant.* This is a general term which can be applied in a large variety of contexts. In clinical trials it is known as “clinical significance” and when it used by Educators is often referred to as “educational significance”.

Authors, such as Cohen (1969, 1977, 1988), have attempted to assign guidelines and cut-off values for these effect size indices so that one can determine when the effect is considered to be “small”, “medium” and “large”. However, due to the fact that the choice of these cut-off values is fairly arbitrary, they have been the recipient of a great deal of criticism. Throughout this text these guidelines and cut-off values will be provided along with a motivation for their use; these motivations will also be accompanied by critical evaluations.

1.2 When are effect size indices necessary?

Suppose that the mean diastolic blood pressure of 25 hypersensitive patients is lowered by, for example, 10 mmHg, after they had received a certain treatment. Suppose further that this lowering in blood pressure was statistically significant at a significance level of 1%. The researcher could, based on this result, say that it was also a practically significant reduction in blood pressure, since the pressure was measured on the mm-Hg scale (where it is known that a difference of 10 units is enough to be important). In this case an effect size index is not necessary. Another example is when one obtains a highly significant correlation $\left(p < 0,0001 \right)$ of 0,8 between a new psychometric test and a standard test used to measure, for example, depression in a sample group of 200 individuals. This correlation indicates that the new test is a valid test for the sample group, because, from the psychologist’s experience, a correlation of 0,8 or more is large enough to indicate validity.

In both of these examples, knowledge of the scale or prior experience was sufficient to make conclusions about the practical significance of the results. Thus, calculating effect size indices and evaluating them with respect to cut-off values or guidelines was unnecessary.

However, there are many other cases where one will be required to calculate effect size indices in order for one to determine whether a result is practically significant or not. The following is a brief list of some of these cases:

Case 1:

The scale on which the various means are measured is unknown. For example, suppose that one has a questionnaire which has not been standardized and the questions are asked on a 4-point Likert scale; it would be difficult to interpret a difference between the means of, say, 0,3. For standardized scales, such as the stanine scale or the sten scale, the researcher knows the standard deviation of the scale beforehand and can interpret the differences in this context making effect size index calculations redundant. When researchers work with the so-called raw counts before standardization, the scale is usually unknown.

Case 2:

Variation of measurements on any given scale differs depending on the situation or the subjects. For instance, the standard deviation of IQs for a selected group of university students could be much smaller than that of a group of individuals selected from the general populace. Now, since effect size indices compensate for the variation of the measurements, their use would be recommended in this situation to determine the practical significance of results.

Case 3:

Probability samples (such as simple random samples or stratified samples) can be drawn from populations and statistically significant results can be obtained, but there will still be uncertainty as to the importance of differences or relationships. In these cases effect size indices are calculated as a *second step* after statistical significance is tested. If very *large* samples are used then one very often finds that the results are statistically significant, while the effect size indices will help determine if these results are practically important. When using *small* samples a statistically significant difference (or relationship) is, more often than not, also an important difference (or relationship) and calculating effect size indices is usually unnecessary. When the result is shown to be not statistically significant, then it is possible that the effect size index could still show that there is an important difference or relationship. This can mean one of two things: First, the realised result just happens to be as strong as indicated, or, second, there may well be a considerable difference or relationship, but the sample is too small to detect it. The latter would possibly lead towards redesigning the research - using larger samples and perhaps more accurate measurements. Small samples occur frequently in *new research projects* and *pilot studies*, so effect size indices are useful indicators of whether the research may continue.

The following table gives the potential problems when conclusions are drawn from data as a function of effect size and significance level (Rosenthal et. al., 2000:4).

Table 1: Potential problems of inference as a function of effect sizes and significance levels

	Effect size: “Acceptable” (large enough)	Effect size: “Unacceptable” (too small)
Level of significance: “Acceptable” (low enough)	No problem	Mistaking statistical significance for practical importance
Level of significance: “Unacceptable” (too high)	Failure to perceive practical importance of “nonsignificant” results	No problem

Case 4:

When *meta-analysis* is conducted, effect size indices are required to combine the results of different studies. Determining practical significance is typically not the main objective of meta-analysis so this text will not discuss the application of effect size indices in this regard any further. Books written by Rosenthal (1991), Hedges & Olkin (1985) and Hunter & Schmidt (2004) can be consulted for further details on this topic.

Case 5:

If the realised *power* of a statistical test is to be determined after the completion of the experiment (i.e., post-hoc), effect sizes are then necessary. Cohen (1969, 1977, 1988) provides power tables for nearly all statistical tests where effect sizes are used.

Case 6:

In the planning of a study it is possible to determine the sample size which will produce a specified power (typically 80%) at a certain significance level. Effect sizes are used in this calculation and, if they are not known, they can be estimated from a pilot study. Once again, Cohen's tables can be employed to aid in these calculations.

Case 7:

For *complete surveys* (censuses) where the entire population is studied, effect sizes are essentially the only method to determine the practical importance of results. In practice *complete surveys* appear quite frequently. A few examples (Steyn, 1999) include:

(a) A study to compare rapists and armed robbers from 3 different prisons. Both populations were so small that all of the individuals were included in the study.

(b) First year Psychology students undergo psychometric tests to determine the validity of the tests. Instead of randomly sampling from the class of students it is, in practice, simpler to allow the entire class to take the tests. The class is now the study population which is tested in its entirety. Even if a random sample were to be drawn, the inferences and conclusions drawn would only have been applicable to this study population. A full class test thus provides the exact results for the given population.

(c) When distributing a questionnaire to the individuals selected in a probability sample from a certain target population, it is found that the response is so deficient (say only 20%) that while the sample itself was representative of the population, the realised responses are not. This is due to the fact that the subjects in the sample "chose themselves" when they decided to respond or not. The researcher is thus forced to treat the respondents as a sub-population of the target population. The respondents now form a complete survey of this sub-population of respondents. Results obtained from this study can no longer be generalised to the target population, even though that was the original idea.

(d) In a study where young people with Adjustment disorder (the experimental group) were compared to a control group, it was found that it was difficult to obtain people from the experimental group, so all of the available people from a number of institutions over a certain period of time were tested using psychometric tests – thus the entire population was tested. Subjects from the control group from the same working environments and age group as the experimental group were, on the other hand, randomly selected.

(e) The full client database of a bank is available electronically and, based on this information and correlations, risk factors are identified. With modern computer technology at the researcher's disposal there are no major impediments to analysing the data obtained from these millions of clients. Even if the study was restricted to a random sample of 10 000 clients, the sample is still so large that any correlations will be statistically significant. Indeed, a sample of this size is so large that it should represent the underlying population very well and could even be treated as the population itself.

(f) Example B in Chapter 3 provides an additional example of a complete population.

To clearly distinguish these last examples of complete populations this manual will, in the discussion of effect size indices, provide the effect size indices for the population after which the sample estimates for the indices (when working with probability samples) will also be provided.

1.3 Requirements for good effect size measures

Preacher & Kelley (2011) proposed the following requirements to hold for good effect size measures:

1. On an appropriate scale. It is difficult to decide whether an effect size is large enough to be sensible without an interpretable scale. So is the standardised mean difference (difference in means divided by the standard deviation) independent of the measuring scale and e.g. can standardised mean differences in IQ be interpreted in the same way as those of blood pressure. A correlation coefficient as an effect size is another example which is easy to interpret.
2. A confidence interval available. When an effect size is determined for a sample, its value will differ from that of the population from which the sample is drawn. A confidence interval then indicates the possible values in between which the population effect size can vary, with high probability.
3. Independent from sample size. When a sample is used to estimate the population value of an effect size, it rather should not depend on the sample size. Effect size values based on different sample sizes should therefore interpreted in the same way.

4. Be unbiased, consistent and efficient. Unbiased effect sizes is when the expected sample effect size is equal to the population effect size, while consistency requires convergence of a sample effect size to its population counterpart with increasing sample size. Efficiency on its turn requires low variation of the sample effect size, which lowers with increasing sample sizes.

In the discussions from Chapter 4 onwards, effect sizes will be considered which usually comply with above requirements.

1.4 Use of effect sizes in the application of Statistics in various fields

Due to the ongoing debates over the years in psychological journals concerning the use of statistical significance tests, the *American Psychological Association* (APA) brought into being the *Task Force on Statistical Inference* (TFSI). Their report (Wilkinson & TFSI, 1999) offers recommendations concerning data analysis. A selection of the main recommendations of Kline (2004a:13) are:

1. Make use of only the most necessary and simplest statistical analyses.
2. Do not report on statistical results obtained from computer packages without first understanding the meaning of these results.
3. Document the assumptions pertaining to the population effect sizes, sample sizes or measurements underlying an a priori estimation of statistical power. If one reports a Post hoc calculation of power one should rather use a confidence interval based on observed results rather than calculating a straightforward “Post hoc” power.
4. Report observed effect sizes of primary outcomes or when *p*-values are provided. This promotes better research and supplies additional information for future meta-analysis.
5. Report confidence intervals of effect sizes.

6. Provide, as far as is possible, some proof that the statistical assumptions that are made are indeed valid.

Note that the topic of effect sizes forms a substantial portion of the issues that are addressed in the report. The fifth edition of the *APA's Publication Manual* (APA, 2001: 21-26) provides, among other things, the following recommendations deduced from the TFSI-report (see Kline, 2004a: 13):

1. Report appropriate descriptive statistics, such as means, variances, group sizes and the common variances and covariances between groups for comparative studies, or a correlation matrix for a regression analysis. This information is required for meta-analysis and further analysis by researchers.
2. Effect sizes should almost always be reported. Their absence are considered to be a characteristic of a flawed study.
3. The use of confidence intervals is strongly recommended.

The sixth edition of the *APA's Publication Manual* (APA, 2010: 33) states further: "... complete reporting of all tested hypotheses and estimates of appropriate effect sizes and confidence intervals are the minimum expectations for all APA journals".

Kline also makes the following recommendations derived from the TFSI and *APA Publication Manual*:

1. Researchers should not consider statistically significant results as being particularly informative, i.e., they do not automatically indicate significance and repeatability.
2. However, a statistical result which is not statistically significant should not be ignored out of hand, i.e., not rejecting the null hypothesis does not

necessarily mean that the population effect is zero. Possible advantageous effects in research are often overlooked in this way.

3. Effect sizes should always be reported and, when possible, should also be reported with their associated confidence intervals. This means that effect sizes are not just calculated as a supplement to statistically significant results. The emphasis should be on the effect sizes themselves so that they are not only reported, but also interpreted.

Huberty (2002) lists 19 journals in educational psychology where the use of effect sizes are encouraged. Bruce Thompson provides a summary of the requirements relating to effect sizes of 9 journals on his website: http://www.coe.tamu.edu/~bthompson/index.htm.

Bartlett (1997) states in an editorial that effect sizes are noticeably absent from research articles in sport and exercise sciences. Effect sizes should, according to him, be seen as a more important part of the statistical testing procedure. Citing Thomas, et.al. (1991) as a motivation, the editor of *Research Quarterly for Exercise and Sport* encourages authors to include effect sizes (or statistics that make their calculation possible) in articles. Fern & Monroe (1996) provide readers of the *Journal of Consumer Research* an overview of the use of effect sizes in various applications, how to interpret them and also the relationships between different effect sizes.

While the above discussion provides an overview of the use of effect sizes in certain application fields of statistics, there is no attempt in any of that literature to compile a comprehensive list of these methods. However, it does give one an idea as to how strongly these methods are supported by the journals. The use of effect sizes in research is struggling to find a foothold. Thompson (2001) attests to this fact when he is able to references only 11 empirical studies in 23 journals published after 1994 which make use of effect sizes. Kirk (1996) also finds in 4 psychological journals that a small percentage of articles which contained work done using statistical inference contained effect sizes. More recently, Cumming et al. (2007) investigated whether statistical practices in psychology have been changing since 1998 after the APA Task Force in Statistical Inference (TFSI) advocated improved statistical practices, including reporting effect sizes and confidence intervals.

1.5 Effect sizes in statistical literature and computer packages

Very few standard textbooks written in statistical methods contain any material relating to effect sizes. One notable exception is the *Handbook of parametric and nonparametric statistical procedures* (Sheskin, 2000) which discusses a large variety of effect size index calculations. Other examples are the 4th edition of *Using multivariate statistics* (Tabachnick & Fidell, 2000) and *Applied discriminant analysis* by Huberty (1994), both of which concentrate on multivariate effect size indices. A result of their underexposure in standard literature is that well known statistical packages such as SAS (SAS Institute, Inc.2002-2003) and STATISTICA (StatSoft Inc., 2011) do not include any options to calculate effect sizes. The package SPSS (SPSS Inc., 2007), on the other hand, does include some ability to calculate effect sizes, but is limited to the reporting of the eta-squared value in MANOVA.

With the aim of calculating power and sample sizes by making use of an assortment of statistical significance tests, Cohen (1969, 1977, 1988) discusses effect sizes in some detail. However, since his purpose was not to use the effect sizes as a measure of practical significance, these discussions do not make any mention of how these effect sizes might be estimated. Meta-analysis also requires effect sizes and, towards this end, the books of Hedges & Olkin (1985) Rosenthal (1991), and Hunter & Schmidt (2004) should be consulted. In an editorial on the journal *Statistics in Medicine,* D'Agostino (1999) proposes some guidelines for practical significance (he, however, calls it “quantitative or clinical significance”). An article written by Feinstein (1999) spurred interest and a broader investigation into practical significance in situations where one compares two groups with one another.

It is clear from the above discussion that, even in statistical literature, effect sizes are seen as a “foreign” concept. It is thus no surprise that it not been utilized more extensively in Applied Statistics. To the best of my knowledge effect sizes and practical significance are also almost never included in introductory statistics courses for first year students at university, nor can it be found in any statistical “support” courses. An exception to this is the course STTN124 presented at the Northwest University’s Potchefstroom campus. It is regrettable that this rather important aspect of Statistics is not taken up for further investigation by the majority of statisticians, but rather, it is left to the users – specifically those in Psychology and Education - to be this cause’s champions!

1.6 Objectives and structure of this manual

- The very fact that there is not a great deal written about effect sizes in the statistical literature means that one of the main purposes of this manual is to expose the statistical consultant to this subject. This is done because many of the sources relating to the theory are concealed in, for example, Psychological literature.
- For the researcher using statistical methods as a tool in their research, this manual will attempt to collect the vast assortment of effect sizes applied in different research problems, into a single package.
- Focus will concentrate on the background and interpretation of effect sizes and not so much on the statistical theory behind it. A sufficient number of examples are provided to illustrate how they would be applied.

- The calculation of most effect size indices are straightforward if one assumes that certain statistics such as the arithmetic mean, standard deviations (SD's) and correlation coefficients have already been calculated (with the help of computer packages such as EXCEL, STATISTICA, SAS or SPSS). The goal is thus to help the reader calculate effect sizes from results that have already been obtained from a computer's output.
- Many of the confidence intervals for effect sizes are too complicated to calculate by hand, and so programs custom written in SAS are employed to do this. These programs have been made available on the web page for this manual. The purpose of this manual is thus also to aid the researcher in the use of confidence intervals and to make it easier for him/her to calculate these intervals.

Chapter 2 discusses measurement scales and assumptions that need to be made by a researcher, after which it provides an overview of the literature on effect size indices. There are a host of examples of empirical research used throughout this text; Chapter 3 will be dedicated to describing these examples so that they may be used again in later chapters. Chapter 4 discusses effect size indices as a standardized difference between two group means, while the effect sizes supplied in Chapter 5 concern the various relationships between variables. The comparison between more than two groups also has associated effect size indices and these will be reviewed in Chapter 6. Chapter 7 will elaborate on the comparison of two or more groups, extending the ideas to the multivariate case.

A further chapter about effect size indices for overlapping groups is also considered. To my knowledge, there are a large number of tests which do not have an associated effect size index, including tests for normality and tests for homogeneity of variances. These issues are the topics of further research.

In some very specialized cases there exist effect size indices. The following is list of these indices coupled with the references to their articles:

1. Effect sizes in multi-factor designs – an extension to Chapter 6: Kline (2004a: Chapter 7), Fidler & Thompson (2001), Olejnik & Algina (2000).
2. Effect sizes for inter-rater agreement: Shoukri (2004).