Physical Therapy in Sport 32 (2018) 87–97

Contents lists available at ScienceDirect

Physical Therapy in Sport

journal homepage: www.elsevier.com/ptsp

The use of parametric effect sizes in single study musculoskeletal physiotherapy research: A practical primer

Nikolas Pautz [a, *], Benita Olivier [b], Faans Steyn [c]

[a] University of KwaZulu-Natal (UKZN), King Edward Ave, Scottsville, Pietermaritzburg, 3209, Durban, KwaZulu-Natal, South Africa
[b] University of the Witwatersrand, Richard Ward, 1 Jan Smuts Ave, Braamfontein, 2000, Johannesburg, Gauteng, South Africa
[c] Statistical Consultation Services, Potchefstroom Campus, North-West University, Hoffman St, Potchefstroom, 2520, South Africa

ARTICLE INFO

Article history:
Received 13 April 2017
Received in revised form
14 November 2017
Accepted 3 May 2018

Keywords:
Physiotherapy
Between-subject
Within subject
Differences
Associations
Effect sizes

ABSTRACT

Many researchers often do not report effect sizes at all, and, if they do report them, often do not report the correct measure for the design that has been used in the research. With the increased level of attention being given to the reporting of effect sizes and their corresponding confidence intervals, it is important that there is field-specific literature pertaining to the calculation and reporting of these measures. This paper acts as a practical primer for the calculation and reporting of effect size measures aimed at, but not limited to, the field of musculoskeletal physiotherapy research. This primer involves a discussion on which effect sizes are appropriate for within and between-subject single study research, illustrating, through examples based on musculoskeletal research data, how these measures are calculated, interpreted, and reported.

1. Introduction

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 in significance testing 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, ≤ 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.*

Effect sizes express the magnitude of an effect, such as, but not limited to, treatment differences and associations. There are multiple definitions for ES measures which has led to variability and uncertainty regarding a standardized definition of effect size (Kazis, Anderson, & Meenan, 1989; Nakagawa & Cuthill, 2007). The definition of effect size (ES) that will be used in this research is as follows: "a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest" (Kelley & Preacher, 2012, p. 140). In other words, ES express either the sizes of associations (e.g., correlations) or the sizes of differences (e.g., means, medians) and are directly related to a research question of interest. Unlike significance tests which are dependent on sample sizes, ES indices are independent of sample size (Kelley & Preacher, 2012). This allows for the comparison of estimates of ES regardless of the sample size that was used to estimate the population value (Steyn, 2012). ES indexes are useful for comparison due to their inherent comparability. For instance, effect sizes are used to calculate responsiveness and can be used to compare different measures (Husted, Cook, Farewell, & Gladman, 2000).

When dealing with probability samples drawn from populations, estimates of the unknown population ES's have to be determined from those samples. To know how accurate these

* Corresponding author.
E-mail addresses: nikolas.pautz@ntu.ac.uk (N. Pautz), benita.olivier@wits.ac.za (B. Olivier), Faans.Steyn@nwu.ac.za (F. Steyn).

https://doi.org/10.1016/j.ptsp.2018.05.002
1466-853X/

estimates are, it is necessary to obtain their confidence intervals (CI's). Such an interval has limits or bounds which can be expected to cover the unknown population ES with a predetermined high probability (usually 95%).

ES measures are of the most important outcomes of empirical studies (Cohen, 1988). Indeed, the call for authors to report and interpret effect sizes, as well as their corresponding CI's, has never been stronger (Kelley & Preacher, 2012). The Publication Manual of the American Psychological Association (APA) states that null hypothesis significance testing is "but a starting point and that additional reporting elements such as ES's, CI's, and extensive description are needed to convey the most complete meaning of the results" (APA, 2010).

Lakens (2013) writes that ES's are useful for three reasons: (1) ES reporting aids researchers in communicating the magnitude of the effect in a uniform manner. (2), Researchers are able to draw meta-analytic conclusions about a variety of studies using the same standardized ES. (3) Previous studies which have reported ES's can be used in order to calculate appropriate sample sizes for future studies a priori.

Other reasons for the use of ES's are (Steyn, 2012, Chapter 1): (4) To supplement the results of statistically significance testing to determine the practical importance. The p values calculated by statistical tests allow for the interpretation of statistical significance. ES indices are useful in aiding the interpretation of results in the sense that they are directly proportional to the importance of the difference of means or relationships between variables (Steyn, 2012). If an ES index is large enough, the difference or relationship that it represents can be said to be *practically significant*. Practical significance is a rather general term which changes according to context; in clinical trials it is known as *clinical* significance, and when it is used in educational research, it is typically referred to as *educational* significance. For example, Bartolucci, Tendera, and Howard (2011) reported high statistical significance ($p < .00001$) on the effect of aspirin to prevent myocardial infarction (MI), concluding with a recommendation that aspirin should be used for general prevention. While the p-value was highly *statistically significant*, the ES of the relationship was extremely small ($r^2 = 0.001$), indicating that the *practical significance* of the effect of aspirin on MI was low and that the risk of aspirin actually outweighed the possible benefit. (5) If the realised *power* of a statistical test is to be determined after the completion of the experiment (i.e., post-hoc), ES measures are then necessary. (6) For complete surveys (censuses) where the entire population is studied, ES are essentially the only method to determine the practical importance of results (see Steyn, 2012, for detailed examples of these points).

This paper will be focusing on several aspects of ES's that are not commonly produced in standard statistical analyses outputs generated by software (e.g., SPSS®, and Statistica®). Kelley and Preacher (2012) call for a more comprehensive classification of ES's, and this primer seeks to address some aspects of this call. The typical classification for measures of ES is divided between the d (differences) and the r (correlations) family (Rosenthal, Cooper, & Hedges, 1994). These families and their corresponding measures will be addressed in this primer. We will also be discussing which ES's are appropriate to use for within and between-group designs. Within-group designs apply the same variations of conditions to each subject, for example, a pre-test and a post-test, while between-group designs are used for experiments that have two or more groups of participants each being administered the same test. Additionally, how we can interpret ES measures, including when findings are not statistically significant and how we can go about reporting them, as well as their confidence intervals, will be covered. Practical examples in musculoskeletal physiotherapy will be used to show how these measures can be calculated, and illustrations of how previous papers could have improved upon their results by calculating ES measures will be presented (see Boxes 1-3).

2. The *D* (difference between two means) family

2.1. Between-group design

Perhaps the most commonly known, and therefore used (likely mainly due to the relative simplicity of the calculation and interpretation), measure of ES is d, sometimes known as Cohen's d. Cohen's d is a measure of effect size that is used to describe the standardized mean difference of an effect. d can range from 0 to infinity, where higher scores are equated to a greater magnitude of effect. It must be noted that d can have negative scores. Researchers have the option of using negative ES values when utilizing multiple outcome measures. For example, improvement in function may result in a positive ES, while improvement in pain may result in a negative ES (Becker, 2000). Cohen (1988) started the convention of using subscripts to denote the different versions of Cohen's d, a practice which will be continued in this primer as it helps to prevent confusion.

Cohen (1994) defined d as a population parameter and not an estimate based on samples. Thus, many of the formulae which derived from the original d, such as d_s, are estimates of the ES on a specific population from which the sample derived. If a complete survey is used (such as a census) and the entire population one is investigating is included in the analysis, then Cohen's d in its original state would be appropriate (Steyn, 2012). As population data is not frequently collected in physiotherapy research, as well other fields, this paper has focused on discussing sample estimates.

Box 1

Jenkins, Williams, Williams, Hefner, and Welch (2017) investigated sex differences in total frontal plane knee movement and velocity during a functional single-leg landing. In one section of the analysis, differences between groups (male and female) using independent samples t-tests were presented. No measures of effect size - apart from the absolute difference between the means (D) - were reported, and thus the readers without an intuitive understanding of the measure were left without a true impression of the magnitude of the difference between the groups. There was a significant difference ($p = 0.03$) in frontal plane knee excursion (Valgus in Table 2) between the "female" ($n = 20$; $M = 7.10$; $SD = 3$) and "male" ($n = 20$; $M = 4.39$; $SD = 1.6$) groups. This finding could have been expanded by reporting any of the above reported measures of effect size. For instance, d_s could have been calculated using **FORMULA 1**, giving us a d_s value of 1.13. From this value, it is possible to calculate the less biased estimator of effect size, g_s. Using **FORMULA 3**, a g_s value of 1.11 can be calculated, with a lower 95% CI of 0.44, and an upper limit of 1.77. The unstandardized difference in means is 2.71, which can also be interpreted by the researcher. Using Cohen's (1988) conventions, which will be discussed in detail later, we can know intuitively that the difference between these groups is large. By adding these relatively simple calculations to the outcomes section, we are able to get much more information out of the results.