REVIEW ARTICLE

TRENDS in
Sport Sciences
2014; **1**(21): 19-25.
ISSN 2299-9590

The need to report effect size estimates revisited. An overview of some recommended measures of effect size

MACIEJ TOMCZAK[1], EWA TOMCZAK[2]

[1] *University School of Physical Education in Poznań, Department of Psychology, Poland*

[2] *Adam Mickiewicz University in Poznań, Faculty of English, Department of Psycholinguistic Studies, Poland*

Recent years have witnessed a growing number of published reports that point out the need for reporting various effect size estimates in the context of null hypothesis testing (H_0) as a response to a tendency for reporting tests of statistical significance only, with less attention on other important aspects of statistical analysis. In the face of considerable changes over the past several years, neglect to report effect size estimates may be noted in such fields as medical science, psychology, applied linguistics, or pedagogy. Nor have sport sciences managed to totally escape the grips of this suboptimal practice: here statistical analyses in even some of the current research reports do not go much further than computing *p*-values. The *p*-value, however, is not meant to provide information on the actual strength of the relationship between variables, and does not allow the researcher to determine the effect of one variable on another. Effect size measures serve this purpose well. While the number of reports containing statistical estimates of effect sizes calculated after applying parametric tests is steadily increasing, reporting effect sizes with non-parametric tests is still very rare. Hence, the main objectives of this contribution are to promote various effect size measures in sport sciences through, once again, bringing to the readers' attention the benefits of reporting them, and to present examples of such estimates with a greater focus on those that can be calculated for non-parametric tests.

KEY WORDS: sport science, effect size calculation, parametric tests, non-parametric tests, methodology.

Received: 12 September 2013
Accepted: 15 February 2014

Corresponding author: maciejtomczak5@gmail.com

What is already known on this topic?
Estimates of effect size allow the assessment of the strength of the relationship between the investigated variables. In practice, they permit an evaluation of the magnitude and importance of the result obtained. An effect size estimate is a measure worth reporting next to the *p*-value in null hypothesis testing. However, not every research report contains it. After the null hypothesis has been tested with the use of parametric and non-parametric tests (statistical significance testing), measures of effect size can be estimated.

A few remarks on statistical hypothesis testing

Studies in sport sciences have addressed a wide spectrum of topics. Empirical verification in these areas often makes use of correlation models as well as experimental research models. Just like other scholars conducting empirical research, researchers in sport sciences often rely on inferential statistics to test hypotheses. From the point of view of statistics, the hypothesis verification process often comes down to determining the probability value (*p*-value), and to deciding whether the null hypothesis (H_0) is rejected (a test of statistical significance) [1, 2, 3, 4]. In the case of rejecting the null hypothesis (H_0), a researcher will accept an alternative hypothesis (H_1), which is often referred to as the so-called substantive hypothesis as a researcher formulates it based on various criteria applicable to their own studies. Such an approach to hypothesis verification has its origin in Fisher's approach (*p-value approach*) and the Neyman-Pearson framework to hypothesis testing that was developed later (*fixed-α approach*). Below, based on Aranowska and Rytel [5, p. 250], we present the two approaches (Table 1).

Rejecting the null hypothesis (H_0) when it is in fact true is what Neyman and Pearson call making a Type I error (known as "false positive" or "false alarm"). To control for Type I error, or in other words, to minimize the chance of finding a difference that is not really there in the data, researchers set an appropriately low alpha level in their analyses. By contrast, failing to reject the null hypothesis (H_0) when it is actually false (and should be rejected) is referred to as a Type II error (known as "false negative"). Here, increasing the sample size is an effective way of reducing the probability of obtaining a Type II error [1, 2, 3].

The presented approach to hypothesis testing has been a common practice in many disciplines. However, reporting the *p*-value alone and drawing inferences based on the *p*-value alone is insufficient. Hence, statistical analyses and research reports should be supplemented with other essential measures that carry more information about the meaningfulness of the results obtained.

Why the p-value alone is not enough? – or On the need to report effect size estimates

Thanks to some of its advantages, the concept of statistical significance testing has prevailed in the empirical verification of hypotheses to the extent that many areas have still seen other vital statistical measures go largely unreported. In spite of recommendations not to limit research reports to presenting the null hypothesis testing and reporting the *p*-value only, to this day a relatively large number of published articles have not gone much beyond that. By way of illustration, a meta-analysis of research accounts published in one prestigious psychology journal in the years 2009 and 2010 showed that almost half of the articles reporting an Analysis of Variance (ANOVA) did not contain any measure of effect size, and only a mere quarter of the surveyed research reports supplemented Student's t-test analyses with information about the effect size [6]. Sport sciences have seen comparable practices every now and then. As already pointed out, giving the *p*-value only to support the significance of the difference between groups, or measurements, or the significance of a relationship is insufficient [7, 8]. The *p*-value alone merely indicates what the probability of obtaining a result as extreme as or more extreme than the one actually obtained, assuming that the null hypothesis is true [1]. In many circumstances, the computed *p*-value depends (also) on the standard error (SE) [9]. It is now well established that the sample size affects the standard error and, as a result of that, the *p*-value. As the size of a sample increases, the standard error becomes smaller, and the *p*-value tends to decrease. Due to this dependence on sample size, *p*-values are seen as confounded. Sometimes a result that is statistically significant mainly indicates that a huge sample size was used [10, 11]. For this reason, the value of the *p*-value does not say whether the observed result is meaningful or important in terms of (1) the magnitude of the difference in the mean scores of the groups on some measure, or (2)

Table 1. Fisher's and Neyman-Pearson's approaches to hypothesis testing

The Fisher approach to hypothesis testing (also known as the *p-value approach*)	The Neyman-Pearson approach to hypothesis testing (also known as the *fixed-α approach*)
– formulate the null hypothesis (H_0) – select the appropriate test statistic and specify its distribution – collect the data and calculate the value of the test statistic for your set of data – specify the *p*-value – if the *p*-value is sufficiently small (according to the criterion adopted), then reject the null hypothesis. Otherwise, do not reject the null hypothesis.	– formulate two hypotheses: the null hypothesis (H_0) and the alternative hypothesis (H_1) – select the appropriate test statistic and specify its distribution – specify α (alpha) and select the critical region (R) – collect the data and calculate the value of the test statistic for your set of data – if the value of the test statistic falls in the critical (rejection) region, then reject the null hypothesis at a chosen significance level (α). Otherwise, do not reject the null hypothesis.

the strength of the relationship between the investigated variables. Relying on the *p*-value alone for statistical inference does not permit an evaluation of the magnitude and importance of the obtained result [10, 12, 13].

In general terms, there are good enough reasons for researchers to supplement their reports of the null hypothesis testing (statistical significance testing: the *p*-value) with information about effect sizes. Given statistical measures, a large number of effect size estimates have been developed and used to this day. As reporting effect size estimates is beneficial in more than one way, below we list the benefits that seem most fundamental [6, 12, 14, 15, 16, 17, 18]:

1. They reflect the strength of the relationship between variables and allow for the importance (meaningfulness) of such a relationship to be evaluated. This holds both for relationships explored in correlational research and the magnitude of effects obtained in experiments (i.e. evaluating the magnitude of a difference). On the other hand, applying a test of significance only and stating the *p*-value may solely provide information about the presence or absence of a difference, its impact and relation, leaving aside its importance.
2. Effect size estimates allow the results from different sources and authors to be properly compared. The *p*-value alone, which depends on the sample size, does not permit such comparisons. Hence, the effect size is critical in research syntheses and meta-analyses that integrate the quantitative findings from various studies of related phenomena.
3. They can be used to calculate the power of a statistical test (power statistics), which in turn allows the researcher to determine the sample size needed for the study.
4. Effect sizes obtained in pilot studies where the sample size is small may be an indicator of future expectations of research results.

Some recommended effect size estimates

In the present section we provide an overview of a number of effect size estimates for statistical tests that are most commonly used in sport sciences. Since parametric tests are frequently used, measures of effect size for parametric tests are described first. Then, we describe effect size estimates for non-parametric tests. Reporting measures of effect size for the latter is more of a rarity. Aside from that, in the overview below we omit the measures of effect size that are most popular and widely reported for parametric tests. In sport sciences examples of the most popular estimates of effect size include correlation coefficients for relationships between variables measured on an interval or ratio scale such as the Pearson's correlation coefficient (*r*). Nor do we present effect size measures popular and widely used, among others, in sport sciences, calculated for relationships between ordinal variables such as the Spearman's coefficient of correlation. Some measures of effect size presented below can be calculated automatically with the help of statistical software such as *Statistica*, *the Statistical Package for the Social Sciences (SPSS)*, or *R*. Others can be calculated by hand in a quick and easy way.

Effect size estimates used with parametric tests

The Student's *t*-test for independent samples is a parametric test that is used to compare the means of two groups. After the null hypothesis is tested, one can easily and quickly calculate the value of the point-biserial correlation coefficient with the help of the Student's *t*-test (provided that the *t*-value comes from comparing groups of relatively similar size). This coefficient is similar to the classical correlation coefficient in its interpretation. Using this coefficient one can calculate the popular r^2 (η^2). The formula used in computing the point-biserial correlation coefficient is presented below [1, 6, 19]:

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

$$r^2 = \eta^2 = \frac{t^2}{t^2 + df}$$

t – value of Student's *t*-test, df – the number of degrees of freedom ($n_1 - 1 + n_2 - 1$); n_1, n_2 – the number of observations in groups (group 1, group 2)

r – point-biserial correlation coefficient

r^2 (η^2) – the index assumes values from 0 to 1 and multiplied by 100% indicates the percentage of variance in the dependent variable explained by the independent variable

Often used here are the effect size measures from the so-called *d* family of size effects that include, among others, two commonly used measures: Cohen's *d* and Hedges' *g*. Below we provide a formula for calculating Cohen's *d* [1, 19, 20, 21]:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sigma}$$

d – Cohen's index
$\bar{x}_{1,2}$ – means of the first and second sample
σ – standard deviation of a population

Normally, we do not know the population standard deviation and we estimate it based on the sample. Given that, it is possible here to use the estimate of standard deviation of the total population. In this case, to estimate the effect size one can compute the *g* coefficient that uses the weighted pooled standard deviation [22]:

$$g = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}}$$

n_1, n_2 – the number of observations in groups (group 1, group 2)
s_1, s_2 – standard deviation in groups (group 1, group 2)
rough arbitrary criteria for Cohen's *d* and Hedges' *g* values: *d* or *g* of 0.2 is considered small, 0.5 medium, and 0.8 large [21]

When it comes to the dependent-samples Student's *t*-test, it is possible to compute the correlation coefficient *r*. For this purpose, the above-presented formula for calculating *r* for independent samples is adopted. However, the *r* coefficient "is no longer the simple point-biserial correlation, but is instead the correlation between group membership and scores on the dependent variable with indicator variables for the paired individuals partialed out" [23, p. 447]. Additionally, once the dependent-samples Student's *t*-test has been used, it is possible to calculate the effect size estimate *g*, where [1, 22]:

$$g = \frac{\bar{D}}{\sqrt{\frac{SS_D}{n-1}}}$$

$\bar{D}$ – mean difference score
SS_D – sum of squared deviations (i.e. the sum of squares of deviations from the mean difference score)

In turn, to compare more than two groups on ratio variables or interval variables, Analysis of Variance (ANOVA) is used, be it one-way or multi-factor ANOVA (provided that the samples meet the criteria). The effect size estimates used here are the coefficient η^2 or ω^2. To compute the former (η^2), we may use the ANOVA output from popular statistical software packages such as *Statistica* or *SPSS*. Below we present the formula [1, 6, 24]:

$$\eta^2 = \frac{SS_{ef}}{SS_t}$$

SS_{ef} – sum of squares for the effect
SS_t – total sum of squares
η^2 – the index assumes values from 0 to 1 and multiplied by 100% indicates the percentage of variance in the dependent variable explained by the independent variable

One of the disadvantages of η^2 is that the value of each particular effect is dependent to some extent on the size and number of other effects in the design [25]. A way out of this problem is to calculate the partial eta-squared statistic (η_p^2), where a given factor is seen as playing a role in explaining the portion of variance in the dependent variable provided that other effects (factors) present in the analysis have been excluded [6]. The formula is presented below [1, 6, 24]:

$$\eta_p^2 = \frac{SS_{ef}}{SS_{ef} + SS_{er}}$$

SS_{ef} – sum of squares for the effect
SS_{er} – sum of squared errors

In the same way, one can calculate the effect size for within-subject designs (repeated measures). However, both coefficients η^2 and (η_p^2) are biased and they estimate the effect for a given sample only. Therefore, we should compute the coefficient ω^2 that is relatively unbiased. To calculate it by hand one can use the ANOVA output that contains values of mean square (*MS*), sum of squares (*SS*), and degrees of freedom (*df*). For between-subject designs the following formula applies [24]:

$$\omega^2 = \frac{df_{ef}(MS_{ef} - MS_{er})}{SS_t + MS_{er}}$$

MS_{ef} – mean square of the effect
MS_{er} – mean square error
SS_t – the total sum of squares
df_{ef} – degrees of freedom for the effect

For within-subject designs ω^2 is calculated using the formula [24]:

$$\omega^2 = \frac{df_{ef}(MS_{ef} - MS_{er})}{SS_t + MS_{sj}}$$

MS_{ef} – mean square of the effect
MS_{er} – mean square error
MS_{sj} – mean square for subjects
df_{ef} – degrees of freedom for the effect

The partial omega-squared (ω^2_p) is computed in the same way both for the between-subject designs and within-subject designs (repeated measures) using the formula below [24]:

$$\omega^2_p = \frac{df_{ef}(MS_{ef} - MS_{er})}{df_{ef}MS_{ef} + (n - df_{ef})MS_{er}}$$

Both η^2 and ω^2 are interpreted similarly to R^2. Hence, these measures multiplied by 100% indicate the percentage of variance in the dependent variable explained by the independent variable.

Effect size estimates used with non-parametric tests

Now we turn to non-parametric tests. Various effect size estimates can be quickly calculated for the Mann-Whitney *U*-test: a non-parametric statistical test used to compare two groups. In addition to the *U*-value, the Mann-Whitney test report (output) contains the standardized *Z*-score which, after running the Mann-Whitney *U*-test on the data, can be used to compute the value of the correlation coefficient *r*. The interpretation of the calculated *r*-value coincides with the one for Pearson's correlation coefficient (*r*). Also, the *r*-value can be easily converted to r^2. The formulae for calculating *r* and r^2 by hand are presented below [6]:

$$r = \frac{Z}{\sqrt{n}}$$

$$r^2 = \eta^2 = \frac{Z^2}{n}$$

Z – standardized value for the *U*-value
r – correlation coefficient where *r* assumes the value ranging from –1.00 to 1.00
r^2 (η^2) – the index assumes values from 0 to 1 and multiplied by 100% indicates the percentage of variance in the dependent variable explained by the independent variable
n – the total number of observations on which *Z* is based

Following the computation of the Mann-Whitney *U*-statistic, one can also calculate the Glass rank-biserial correlation using average ranks from two sets of data ($\overline{R}_1, \overline{R}_2$) and sample size in each group. Some statistical packages next to the test score produce the sum of ranks that can be used to calculate mean ranks. To interpret the calculated value one can draw on the interpretation of the classical Pearson's correlation coefficient (*r*). Here the following formula applies [1]:

$$r = \frac{2(\overline{R}_1 - \overline{R}_2)}{n_1 + n_2}$$

$\overline{R}_1$ – mean rank for group 1
$\overline{R}_2$ – mean rank for group 2
n_1 – sample size (group 1)
n_2 – sample size (group 2)
r – correlation coefficient where *r* assumes the value ranging from –1.00 to 1.00

For another non-parametric test, the Wilcoxon signed-rank test for paired samples, again, the *Z*-score may be used to calculate correlation coefficients employing the formula given below (where *n* is the total number of observations on which *Z* is based) [6].

$$r = \frac{Z}{\sqrt{n}}$$

On the other hand, once the Wilcoxon signed-rank test has been computed, one can also calculate the rank-biserial correlation coefficient using the formula [1]:

$$r = \frac{4\left|T - \left(\frac{R_1 + R_2}{2}\right)\right|}{n + (n+1)}$$

R_1 – sum of ranks with positive signs (sum of ranks of positive values)
R_2 – sum of ranks with negative signs (sum of ranks of negative values)
T – the smaller of the two values (R_1 or R_2)
n – the total sample size
r – correlation coefficient (which is the same as *r* coefficient in its interpretation)

For the Kruskal-Wallis *H*-test, a non-parametric test adopted to compare more than two groups, the eta-squared measure (η^2) can be computed. The formula for calculating the η^2 estimate using the *H*-statistic is presented below [26]:

$$\eta^2_H = \frac{H - k + 1}{n - k}$$

H – the value obtained in the Kruskal-Wallis test (the Kruskal-Wallis *H*-test statistic)
η^2 – eta-squared estimate assumes values from 0 to 1 and multiplied by 100% indicates the percentage of variance in the dependent variable explained by the independent variable
k – the number of groups
n – the total number of observations

In addition, once the Kruskal-Wallis *H*-test has been computed, the epsilon-squared estimate of effect size can be calculated, where [1]:

$$E^2_R = \frac{H}{(n^2 - 1)/(n + 1)}$$

H – the value obtained in the Kruskal-Wallis test (the Kruskal-Wallis *H*-test statistic)
n – the total number of observations
E^2_R – coefficient assumes the value from 0 (indicating no relationship) to 1 (indicating a perfect relationship)

Also, for the Friedman test, a non-parametric statistical test employed to compare three or more paired measurements (repeated measures), an effect size estimate can be calculated (and is referred to as *W*) [1]:

$$W = \frac{\chi^2_w}{N(k - 1)}$$

W – the Kendall's *W* test value
χ^2_w – the Friedman test statistic value
N – sample size
k – the number of measurements per subject

The Kendall's *W* coefficient assumes the value from 0 (indicating no relationship) to 1 (indicating a perfect relationship).

Also, in sport sciences it is quite common practice to use the chi-square test of independence (χ^2). Having tested the null hypothesis (H_0) with a χ^2 test of independence, one may assess the strength of a relationship between nominal variables. In this case, *Phi* (ϕYoula, computed for 2 × 2 tables where each variable has only two levels, e.g. the first variable: male/female, the second variable: smoking/non-smoking) can be reported, or one can report Cramer's *V* (for tables which have more than 2 × 2 rows and columns). The values obtained for the estimates of effect size are similar to correlation coefficients in their interpretation. Again, popular statistical software packages calculate *Phi* and Cramer's *V*. Below we present the formulae for such calculations [1, 6]:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

and for Cramer's *V*:

$$V = \sqrt{\frac{\chi^2}{n(df_s)}}$$

df_s – degrees of freedom for the smaller from two numbers (the number of rows and columns, whichever is smaller)
χ^2 – the calculated chi-square statistic
n – the total number of cases

The *Phi* coefficient and the Cramer's *V* assume the value from 0 (indicating no relationship) to 1 (indicating a perfect relationship).

Conclusions

In the present contribution we have re-emphasized the need to report estimates of effect size in conjunction with null hypothesis testing, and the benefits thereof. We have presented some of the recommended measures of effect size for statistical tests that are most commonly used in sport sciences. Additional emphasis has been on effect size estimates for non-parametric tests, as reporting effect size measures for these tests is still very rare. The present paper may also serve as a point of departure for further discussion where practical (e.g. clinical) magnitude (importance) of results in the light of the conditionings in a chosen area will come into focus.

What this paper adds?

This paper highlights the need for including adequate estimates of effect size in research reports in the area of sport sciences. The overview contains various types of effect size measures that can be calculated following the computation of parametric and non-parametric tests. Since reporting effect size estimates when using non-parametric tests is very rare, this section may prove particularly useful for researchers. Some of the effect size measures given can be calculated by hand quite easily, others can be calculated with the help of popular statistical software packages.

References

1. King BM, Minium EW. Statystyka dla psychologów i pedagogów (Statistical reasoning in psychology and education). Warszawa: Wydawnictwo Naukowe PWN; 2009.
2. Cohen J. The earth is round (p < .05). American Psychologist. 1994; 49(12): 997-1000.
3. Cohen J. Things I have learned (so far). American Psychologist. 1990; 45(12): 1304-1312.
4. Jascaniene N, Nowak R, Kostrzewa-Nowak D, et al. Selected aspects of statistical analyses in sport with the use of Statistica software. Central European Journal of Sport Sciences and Medicine. 2013; 3(3): 3-11.
5. Aranowska E, Rytel J. Istotność statystyczna – co to naprawdę znaczy? (Statistical significance – what does it really mean?). Przegląd Psychologiczny. 1997; 40(3-4): 249-260.
6. Fritz CO, Morris PE, Richler JJ. Effect size estimates: current use, calculations, and interpretation. Journal of Experimental Psychology: General. 2012; 141(1): 2-18.
7. Drinkwater E. Applications of confidence limits and effect sizes in sport research. The Open Sports Sciences Journal. 2008; 1(1): 3-4.
8. Fröhlich M, Emrich E, Pieter A, et al. Outcome effects and effects sizes in sport sciences. International Journal of Sports Science and Engineering. 2009; 3(3): 175-179.
9. Altman DG, Bland JM. Standard deviations and standard errors. British Medical Journal. 2005; 331(7521): 903.
10. Sullivan GM, Feinn R. Using effect size – or why the *p* value is not enough. Journal of Graduate Medical Education. 2012; 4(3): 279-282.
11. Bradley MT, Brand A. Alpha values as a function of sample size, effect size, and power: accuracy over inference. Psychological Reports. 2013; 112(3): 835-844.
12. Brzeziński J. Badania eksperymentalne w psychologii i pedagogice (Experimental studies in psychology and pedagogy). Warszawa: Wydawnictwo Naukowe Scholar; 2008.
13. Durlak JA. How to select, calculate, and interpret effect sizes. Journal of Pediatric Psychology. 2009; 34(9): 917-928.
14. Shaughnessy JJ, Zechmeister EB, Zechmeister JS. Research Methods in Psychology. 5th ed. New York, NY: The McGraw-Hill; 2000.
15. Aarts S, van den Akker M, Winkens B. The importance of effect sizes. European Journal of General Practice. 2014; 20(1): 61-64.
16. Ellis PD. The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results. Cambridge: Cambridge University Press; 2010.
17. Lazaraton A. Power, effect size, and second language research. A researcher comments. TESOL Quarterly. 1991; 25(4): 759-762.
18. Hatch EM, Lazaraton A, Jolliffe DA. The research manual: Design and statistics for applied linguistics. New York: Newbury House Publishers; 1991.
19. Rosnow RL, Rosenthal R. Effect sizes for experimenting psychologists. Canadian Journal of Experimental Psychology. 2003; 57(3): 221-237.
20. Cohen J. Some statistical issues in psychological research. In: Wolman BB, ed., Handbook of clinical psychology, New York: McGraw-Hill; 1965. pp. 95-121.
21. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1988.
22. Hedges LV, Olkin I. Statistical methods for meta-analysis. San Diego, CA: Academic Press; 1985.
23. Rosnow RL, Rosenthal R, Rubin DB. Contrasts and correlations in effect-size estimation. Psychological Science. 2000; 11(6): 446-453.
24. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Frontiers in Psychology. 2013; 4: 863.
25. Tabachnick BG, Fidell LS. Using multivariate statistics. Upper Saddle River, NJ: Pearson Allyn & Bacon; 2001.
26. Cohen BH. Explaining psychological statistics. 3rd ed. New York: John Wiley & Sons; 2008.