ISSN 1750-9823 (print)
International Journal of Sports Science and Engineering
Vol. 03 (2009) No. 03, pp. 175-179

Outcome Effects and Effects Sizes in Sport Sciences

Michael Fröhlich [1], Eike Emrich [1], Andrea Pieter [2] and Robin Stark [3]

[1] Institute for Sports Sciences, Saarland University, 66123 Saarbrücken, Germany
[2] Institute for Health Promotion, DHfPG-University of Applied Sciences, 66123 Saarbrücken, Germany
[3] Institute for Education, Saarland University, 66123 Saarbrücken, Germany

(*Received July 31, 2009, accepted August 22, 2009*)

Abstract. In addition to statistical validation of an intervention in the context of experimental and quasi-experimental designs for hypothesis testing, the practical relevance of an intervention plays a major role. Practical relevance is considered a measure of an experimental effect with respect to various practical issues. Cohen's effect size has become the standard for assessment. However, empirical studies show that effect sizes should not be interpreted statically, but rather dynamically. Furthermore, it seems that prior experience, the target group, the way questions are posed and the context of the study influence the outcome. In the future, in addition to statistical validation, greater consideration should be given to effect sizes to allow a qualitative assessment of a measure's practical relevance. However, the applicable study context and theoretical criteria of the respective research domains must be taken into account.

Key Words: statistical validation, practical relevance, effect size, strength training.

1. Introduction

In experimental and quasi-experimental research designs, it is generally not sufficient to account for the outcome in terms of change measured in comparison with a control group, a further treatment group or longitudinally (statistical significance). It is, rather, the practical relevance of these effects that plays a key role. While this seems obvious for high performance sports and competitive sports, the practical relevance of training measures is increasingly being called for in recreational and fitness sports. Furthermore, the focus shifts to the practical relevance of an intervention when general statements are to be made about a research field on the basis of several individual studies. This means that it must be possible to assess and interpret the effectiveness of individual studies in the context of meta-analyses. In terms of practical relevance, the size of an experimental effect is assessed with respect to various practical issues – inter- and intraindividual assessment, comparison with standard values, changes over time, etc. The statistically significant difference between the experimental treatment groups, including any control groups, is generally a prerequisite for being able to interpret practical relevance with respect to content for the purpose of hypothesis testing. Statistical proof is frequently based on testing the null hypothesis (alpha error probability) with corresponding probability of error, but far less on testing the β-error probability (on the problem of classical hypothesis testing according to Fisher vs. Neyman and Pearson, see current Conzelmann and Raab, 2009). Here, however, statistical significance testing is dependent on the variance and the sample size, such that statistical significance can always be brought about by increasing the sample size *n* and reducing the test variance. Statistical significance thus indicates merely the existence of an effect, but says nothing about its practical relevance.

The practical relevance or relevance of significant effects in assessing the outcome achieved is given by the dimensionless measure "effect size", *ES*. Effect size is a name given to a family of indices that measure the magnitude of a treatment effect. Unlike significance tests, these indices are independent of sample size. Furthermore, effect size measures are the common currency of meta-analysis studies that summarize the findings from a specific area of research. The detailed explanations refer to measures of distance for group differences, and effect sizes are excluded in the following as measures of association.

In the simplest case, the effect size *d* in the population normalizes the differences between the

[1] *Corresponding Author*. Dr. Michael Fröhlich, Institute for Sports Sciences, Saarland University, Campus Building B8 1, 66123 Saarbrücken, Germany, *E-mail address:* m.froehlich@mx.uni-saarland.de

independent experimental groups (also experimental group and control group) to the variance of the test values (t-test for independent samples) according to the formula:

$$d = \frac{(\mu_A - \mu_B)}{\sigma}$$

If the effect size estimate is based on the statistical values of the samples of the two experimental groups, or the treatment group and the control group, then, if the standard deviations of both groups are approximately homogeneous, Cohen's *d* is calculated according to Cohen (1969):

$$d = \frac{(\overline{X}_{EG} - \overline{X}_{CG})}{s}$$

Various recommendations are given with respect to the variance used. Glass et al. (1981) prefer the standard deviation of the control group $s = s_{CG}$. Bortz and Döring (2006) calculate the joint variance of experimental group (*EG*) and control group (*CG*) according to the formula:

$$s = \frac{\sqrt{s_{EG}^2 + s_{CG}^2}}{2}$$

Hedges and Olkin (1985) showed that standardizing mean differences using a pooled standard deviation of both groups optimizes the estimate of the effect size. This pooled standard deviation s_{pooled} is calculated according to:

$$s_{pooled} = \frac{\sqrt{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}}{n_1 + n_2 - 2}$$

The effect size can be reliably calculated for the intervention using the individual formulas. Leonhart (2004) and Morris (2008) discuss the individual formulas with respect to homogeneous and heterogeneous variances, different sample sizes and content aspects, as well as with regard to the test design used. As regards the calculation of effect size using certain test statistics, e.g. t-test, ANOVA, etc., reference is made to the relevant literature (Cohen, 1969, 1992; Glass et al. 1981; Hedges & Olkin, 1985; Hunter & Schmidt, 2004; Rosnow & Rosenthal, 1996). Rhea (2004a, p. 918) points out that the calculation and indication of effect size in the context of training interventions offers numerous advantages:

(1) Effect size is a standardized parameter for assessing and interpreting changes in individual or multiple groups.

(2) Effect size allows various training methods to be compared within one study.

(3) Effect size is easy to calculate and thus reveals the impact of an individual study for theory and practice.

In the context of meta-analyses calculating effect sizes allows the results of individual primary studies to be put in relationship to one another. Meta-analyses use some estimates of effect size because effect size estimates are not influenced by sample sizes. The most common estimate found in current meta-analyses is Cohen's *d* (cf. Thomas et al., 2005, p. 248 f.). In this way, effect sizes ultimately allow individual studies on the same topic to be evaluated and global statements made on a research subject. Hall and Rosenthal (1995, p. 394) speak euphorically of a revolution in the sciences when they point out that "At a time when the need to integrate empirical knowledge has never been greater, we now have statistical tools that permit the summary and analysis of entire research literatures, using replicate quantitative methods. It is nearly impossible to overestimate the gains in knowledge that can result from the application of meta-analysis to scientific literature." In addition, moderator variables and their impact on a global effect can be estimated (Hall & Rosenthal, 1995; Morris, 2008; Peterson et al., 2004; Rhea, 2004b; Rhea et al., 2003).

2. Assessing practical relevance

Assessing practical relevance has a long tradition, dating back to the studies of Cohen (1969). In the article "*A Power Primer*" from 1992, Cohen describes the intent that guided him in assessing effect sizes and criteria that are now established convention (Cohen, 1992, p. 156): "My intent was that medium ES represent an effect likely to be visible to the naked eye of a careful observer. […] I set small ES to be noticeably smaller than medium but not so small as to be trivial, and I set large ES to be the same distance above medium as small was below it. Although the definitions were made subjectively, with some early minor

adjustments, these conventions have been fixed since the 1977 edition of SPABS and have come into general use."

The effect size classification that has since become convention exhibits the following values: *small* effect $d = 0.20$, *medium* effect $d = 0.50$ and *large* effect $d = 0.80$ (cf. Cohen, 1969, p. 38; 1992, p. 157). Cohen (1969, p. 12) himself puts the convention-based effect size classification into perspective: "For each statistical test's ES index, the author proposes, *as a convention*, ES value to serve as operational definitions of the qualitative adjectives "small," "medium," and "large." This is an operation fraught with many dangers: The definitions are arbitrary, such qualitative concepts as "large" are sometimes understood as absolute, sometimes as relative; and thus they run a risk of being misunderstood."

Generally speaking, effect sizes greater than 0.50 are interpreted as *large*, effect size of 0.50-0.30 as *medium*, effect size of 0.30-0.10 as *small*, and those < 0.10 as *trivial*. However, the classification proposed by Cohen is just an initial orientation aid. In contrast, Sedlmeier (1996, p. 55) emphasizes that the classification, which was not originally established empirically, reflects the average effects in various areas of psychology. The extent to which this applies to other research disciplines and domains will be debated and discussed by way of example based on practical relevance in strength training research.

3. Problems in assessing practical relevance and effect size classification

Current meta-analyses by Fröhlich and Gießing (2008), Fröhlich and Schmidtbleicher (2008), Fröhlich et al. (2008), Peterson et al. (2004), Rhea et al. (2003) and Rhea and Alderman (2004) showed that effect sizes in strength training research depend on the subjects' level of training in the dichotomous categorization "trained" vs. "untrained," but are also influenced by gender and, in part, even the duration of the intervention. Furthermore, effect sizes are determined by the training method used, so indirectly by individual training parameters, like trainings intensity, training frequency, and rest pause (Rhea et al., 2003; Peterson et al., 2005). Thus, in this field, effect sizes as an indicator of practical relevance should be viewed dynamically rather than statically. Furthermore, empirical results show that effect sizes must be interpreted in different ways depending on the context, prior experience, research domains, etc. The following two tables show effect sizes as a factor of training status, respectively training experience and the number of series, as well as the duration of the strength training study.

Table 1: Effect sizes per group and condition by volume (Rhea et al., 2003, p. 458)

Sets	Trained (mean ± sd)	n	Untrained (mean ± sd)	n
1	0.47 ± 0.57	25	1.16 ± 1.59	233
2	0.92 ± 0.52	14	1.75 ± 1.98	82
3	1.00 ± 1.26	122	1.94 ± 3.23	399
4	1.17 ± 0.81	12	2.28 ± 1.96	321
5	1.15 ± 0.99	23	1.34 ± 0.89	38
6	-	-	0.84 ± 0.42	46

Table 2: Effect sizes of single-set and multiple-set interventions for different study durations (cf. Fröhlich, Emrich & Schmidtbleicher, in press)

(mean ± sd)	Study duration [weeks]				
	1-6	7-12	13-18	19-24	25-30
Single set	0.76 ± 0.32	1.02 ± 0.71	0.89 ± 1.07	0.76 ± 0.69	1.24 ± 0.34
Multiple set	0.87 ± 0.38	1.05 ± 0.62	1.23 ± 0.64	0.81 ± 0.47	3.42 ± 2.04

What implications can be drawn from the above statements regarding practical relevance and effect size classification? First, in addition to the statistical values and test statistics, effect sizes should be provided as measures of practical relevance, and second, the calculated effect sizes must be judged and evaluated according to the research disciplines or research domains. This means that each scientific community must ultimately decide what is to be classified as a "small," "medium" or "large" effect in a specific research context. In doing so, theoretical criteria for the respective domain must be specifically taken into account.

The following example will illustrate this (Drinkwater, 2008): In the context of a novel training program, the 100m sprint time was improved significantly ($p < 0.05$) by 0.05 seconds. Now how should this statistically significant improvement, and thus the training method, be assessed in terms of practical

relevance? Should the novel training method be introduced or not? For practical assessment, the effect size is calculated: At a school sports day, it was determined that boys had an average 100m time of 13.04 seconds, with a standard deviation of 2.02 seconds. Calculating Cohen's *d* gives an effect size of $d = 0.05 / 2.02 = 0.025$. Thus, in terms of practical relevance, the new method must be considered not worthwhile. Applying the same statistically significant training improvement of 0.05 seconds to the participants in the men's 100m finals in the 2008 Olympics yields a very different result. The finalists had an average 100m time of 9.92 (SD = 0.11). Here, calculating the effect size of the new training method shows a medium effect ($d = 0.05 / 0.11 = 0.45$), with the training method sometimes having a significant impact on standings in the finals.

4. Possible aids for assessing practical relevance and effect size classification in strength training research

In 2004, Rhea proposed, based on approximately 3,000 effect sizes from more than 400 primary studies, the effect size scale for strength training interventions given in table 3 (Rhea, 2004a). In principle, this proposal is justifiable and should be validated in further studies. However, the reported effect sizes should not be viewed as static values, but rather dynamically, as a function of the duration of the intervention, prior experience, and the associated research domain. This means that strength training research would be called upon to analyze the effect sizes documented by Rhea (2004a), similar to the data presented in table 2, depending on the duration of the intervention. It can reasonably be assumed here that the effect sizes are a good indicator of adaptation to the temporal dynamics of training processes.

Table 3: Scale for determining the magnitude of effect sizes in strength training research (Rhea, 2004a, p. 919)

Magnitude	Untrained	Recreationally trained	Highly trained
Trivial	< 0.50	< 0.35	< 0.25
Small	0.50-1.25	0.35-0.80	0.25-0.50
Moderate	1.25-1.9	0.80-1.50	0.50-1.0
Large	> 2.0	> 1.50	> 1.0

Untrained = individuals who have not been consistently trained for 1 year; recreationally trained = individuals training consistently from 1-5 years; highly trained = individuals training for at least 5 years.

5. Conclusion

In addition to indicating statistical significance, empirical studies should specify effect sizes as a quantitative measure for assessing practical relevance (Leonhart, 2004). However, effect sizes should be specified for individual research disciplines or domains, and should be interpreted dynamically. For example, for strength training beginners, this would result in high effect sizes that explain the high increases in strength training in the first few months compared with those at an advanced or expert level. In the further course of training, the corresponding performance increases are smaller – a fact that likely also applies to learning processes – which is expressed by lower absolute effect sizes. However, as shown with the example of the 100m sprint, effect sizes must be interpreted differently according to the target group, setting and issue. For the field of strength training research, the effect sizes presented in table 3 should be validated in the scientific community. If common effect sizes exist in a given research field, then they should be used to assess study results in that field. Particularly the applied sciences, and explicitly sports science or strength training research, should statistically underpin the significance of not only the results found, but also document the practical relevance of individual effects and consider this against the background of the factors discussed here.

6. References

[1] J. Bortz, N. Döring. *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Berlin, Heidelberg u. a.: Springer, 2006.

[2] J. Cohen. *Statistical power analysis for the behavioral sciences*. New York, London u. a.: Academic Press, 1969.

[3] J. Cohen. A power primer. *Quantitative Methods in Psychology*. 1992, **12**(1): 155-159.

[4] A. Conzelmann, M. Raab. *Datenanalyse: Das Null-Ritual und der Umgang mit Effekten in der Zeitschrift für Sportpsychologie*. Zeitschrift für Sportpsychologie. 2009, **16**(2): 43-54.

[5] E. J. Drinkwater. Applications of confidence limits and effect sizes in sport research. *The Open Sports Sciences Journal*. 2008, **1**(1): 3-4.

[6] M. Fröhlich, J. Gießing. The effectiveness of single-set vs. multiple-set training – A meta-analytical consideration. *In J. Gießing, M. Fröhlich (Eds.), Current results of strength training research. A multi-perspective approach*. Göttingen: Cuvillier Verlag, 2008, pp. 9-33.

[7] M. Fröhlich, D. Schmidtbleicher. Trainingshäufigkeit im Krafttraining - ein metaanalytischer Zugang. *Deutsche Zeitschrift für Sportmedizin*. 2008, **59**(2): 34-42.

[8] M. Fröhlich, E. Emrich, D. Schmidtbleicher. Outcome effects of single-set vs. multiple-set training – an advanced replication study. *Research in Sports Medicine: An International Journal*. in press.

[9] M. Fröhlich, J. Gießing, D. Schmidtbleicher, E. Emrich. A comparison between 2 and 3 days of strength training per week – A metaanalytical approach. *In J. Gießing, M. Fröhlich (Eds.), Current results of strength training research. A multi-perspective approach*. Göttingen: Cuvillier Verlag, 2008, pp. 151-166.

[10] G. V. Glass, B. McGaw, M. L. Smith. *Meta-analysis in social research.* Beverly Hills: CA: Sage, 1981.

[11] J. A. Hall, R. Rosenthal. Interpretation and evaluating meta-analysis. *Evaluation & the Health Professions*. 1995, **18**(4): 393-407.

[12] L. V. Hedges, I. Olkin. *Statistical methods for meta-analysis.* New York, London u. a.: Academic Press, 1985.

[13] J. E. Hunter, F. L. Schmidt. *Methods of meta-analysis: correcting error and bias in research*. Newbury Park: Sage, 2004.

[14] R. Leonhart. *Effektgrößenberechnung bei Interventionsstudien.* Rehabilitation. 2004, **43**: 241-246.

[15] S. B. Morris. Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*. 2008, **11**(2): 364-386.

[16] M. D. Peterson, M. R. Rhea, B. A. Alvar. Maximizing strength development in athletes: a meta-analysis to determine the dose-response relationship. *Journal of Strength and Conditioning Research*. 2004, **18**(2): 377-382.

[17] M. R. Rhea, B. Alderman. A meta-analysis of periodized versus nonperiodized strength and power training programs. *Research Quarterly for Exercise and Sport*. 2004, **75**(4): 413-422.

[18] M. R. Rhea. Determining the magnitude of treatment effects in strength training research through the use of the effect size. *Journal of Strength and Conditioning Research*. 2004a, **18**(4): 918-920.

[19] M. R. Rhea. Synthesizing strength and conditioning research: the meta-analysis. *Journal of Strength and Conditioning Research*. 2004b, **18**(4): 921-923.

[20] M. R. Rhea, B. A. Alvar, L. N. Burkett, S. D. Ball. A Meta-analysis to determine the dose response for strength development. Medicine and Science in Sports and Exercise. 2003, **35**(3): 456-464.

[21] R. L. Rosnow, R. Rosenthal. Computing contrasts, effect sizes, and counternulls on other people´s published data: general procedures for research consumers. *Psychological Methods*. 1996, **1**(4): 331-340.

[22] P. Sedlmeier. Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. Methods of Psychological Research Online. 1996, **1**(4): 41-63.

[23] J. R. Thomas, J. K. Nelson, S. J. Silverman. *Research methods in physical activity*. Champaign, Illinois: Human Kinetics, 2005.