Journal of Strength and Conditioning Research, 2004, 18(4), 918–920

Research Note

Determining the Magnitude of Treatment Effects in Strength Training Research Through the Use of the Effect Size

Matthew R. Rhea

Department of Physical Education, Southern Utah University, Cedar City, Utah 84720.

Abstract. Rhea, M.R. Determining the Magnitude of Treatment Effects in Strength Training Research Through the Use of Effect Size. *J. Strength Cond. Res. 18(4)000–000. 2004.*—In order to improve the applicability of research to exercise professionals, it is suggested that researchers analyze and report data in intervention studies that can be interpreted in relation to other studies. The effect size and proposed scale for determining the magnitude of the treatment effect can assist strength and conditioning professionals in interpreting and applying the findings of the strength training studies.

Key Words. statistics, data interpretation, meta-analysis, meaningfulness

Introduction

Bridging the gap between research and practice is vital to the advancement of both theory and the application of theory among strength and conditioning professionals. In fact, this has been one of the long-standing goals of the National Strength and Conditioning Association. A method for decreasing this gap is for researchers to analyze and report their data in a more practical and professionally applicable manner. Most researchers in exercise science report only the statistical probability (the *p* value) of their results. This statistic represents the reproducibility of the study. For instance, at the 0.05 level, the results of a particular study can be expected to occur 95 times out of 100. While the *p* value is important in determining the amount of confidence we can place in the findings, the value provides no measure of the magnitude of the treatment effect and is of little value when reported alone.

Probability values are highly affected by sample size and variance (11). Therefore, this statistic may be misleading in that large differences in groups or treatments may fail to be identified at the 0.05 level simply due to small sample sizes or large variance. In contrast, trivial differences may reach the 0.05 level if the study had a large enough sample size. Cohen explained what he considered to be "a fact widely understood among statisticians" (3) that, in the real world, the null hypothesis is always false and that a large enough sample size will produce significant results. If a measure of meaningfulness or magnitude of the treatment effect was reported, researchers would be able to evaluate the data more thoroughly and precisely. In fact, if the null hypothesis is always false in the real world, we should be more concerned about identifying the actual magnitude of a treatment effect.

Previous authors (12) have reported the need to convince researchers of the importance of reporting an estimate of the magnitude of the differences between groups as well as the significance of the effects. This approach to evaluating research findings is especially relevant for researchers examining strength development. For the purposes of applying strength training research into practice, the magnitude of the treatment affect may be more important than, or at least just as important as, the reproducibility of the study. Therefore, the first step to improving the applicability of research to practice is to calculate and report statistics that examine the actual magnitude of a treatment effect.

Various methods have been described for estimating the magnitude of a treatment effect (4) or an effect size (ES). Among these, eta squared, omega squared, and Cohen's d have been presented as the most common and, generally, the most appropriate. Regardless of the chosen calculation for an ES, researchers should report some calculation of the magnitude of a treatment.

In terms of ease of calculation, practicality, and application to practice, Cohen's d and the standardized mean difference may be the preferred methods of ES calculation in the area of strength and conditioning research. These methods for calculating an ES can easily be calculated by the following formulas (1):

$$d = (M_E - M_C)/SD_C \quad (1)$$

where M_E = the mean of the experimental group, M_C = the mean of the control group, and SD_C = the standard deviation of the control group.

$$\text{Pre-Post ES} = \frac{\text{Posttest mean} - \text{Pretest mean}}{\text{Pretest SD}} \quad (2)$$

As seen above, these ESs represent the difference between two means, divided by the variability among the sample. They are therefore reported in standard deviation units such that an ES of 0.5 represents a difference of ½ of a standard deviation. The ES provides several benefits to researchers and professionals. First, it represents a standard unit for measuring and interpreting changes in one or more groups that can easily be calculated by both researchers and strength professionals. Second, it allows for comparisons of different training methods within a single study. It also provides a method for comparing treatments in separate but related studies. ESs can be directly compared across studies because they have been normalized (10). Finally, the ES enables a single study to have a greater impact on theory and practice. As previous authors (12) have stated, a single study resulting in a "yes/no" decision at the 0.05 level is unlikely to have an impact on theory or practice. However, including the ES offers a valid method for comparisons with

past and future research, and thus has the potential to contribute greater information to the body of research and the practices among professionals.

The calculation of percentage increases in strength (the difference between pretest and posttest scores divided by the pretest score) is common in an attempt to accomplish the same goal as the ES: to determine the magnitude of the changes in strength. However, the calculation of percentage increases does not take into consideration the variance of strength improvements among subjects and, therefore, cannot be accurately compared either within or across research studies. By including the variance in the calculation, the ES accounts for the variation within and across samples, making it a standardized and more accurate description of the treatment effect.

The second step to providing more applicable information for professionals is to determine the relative magnitude of an ES in comparison with other treatment effects in strength training research. The development of an ES scale would provide researchers and professionals with benchmarks to which calculated ESs in strength training research could be applied in order to determine their relative magnitude. Cohen (1) estimated such a scale for behavioral and social sciences. His scale identified 0.2 as representing a small effect, 0.5 a moderate effect, and 0.8 or greater as a large effect. Cohen (2) later revisited this scale and proposed that less than 0.41 represent a small ES, 0.41–0.70 a moderate ES, and greater than 0.70 a large ES. However, Cohen arbitrarily assigned these magnitudes to be used in the behavioral sciences, and it is uncertain as to whether Cohen's scale accurately represents magnitudes in strength training research.

With the completion of a number meta-analyses (5–8) as well the calculation and analysis of the magnitude of treatment effects in a large number of strength training sessions, it has become apparent that Cohen's scale for the social and behavioral sciences does not accurately reflect the norm for ESs in strength training research. Among the nearly 3,000 effect sizes from more than 400 studies, including various doses and modes of strength training, the average ES calculated was about 1.25 (± 1.0). Cohen (1) stated that a small ES should be one that would not occur by chance and that a large ES should be difficult, but possible, to achieve. While it represents only the average ES in strength training research, an ES of 1.25 is considered very large based on Cohen's scale for the social/behavioral sciences. Thus, Cohen's scale does not accurately reflect the relative magnitudes of treatment effects in strength training research.

The differences in the size of ESs among social/behavioral research and strength training research is most likely due to the types of treatments employed in behavioral sciences and the potential for change among dependent variables. Interventions such as visualization, imagery, and relaxation do not appear to elicit changes in measurements that are as large or drastic as the ability of a resistance training program to increase strength measures. These differences in the magnitude of treatment effects impede our ability to gain an accurate comparison to other strength training research when Cohen's scale is used.

After careful and thorough examination of the ESs calculated in a variety of strength training research, it is proposed that a new scale, specific to strength training research and the training status of the subjects being measured, be used to evaluate the relative magnitude of an ES in this area (Table 1). For the purposes of this scale, an untrained individual is considered one who has not been consistently training for at least 1 year. Recreationally trained populations have been training consistently for at least 1 year but less than 5 years. A highly trained individual is one who has been training consistently for at least 5 years. While numerous definitions of training status could be argued, a simple characterization such as the one suggested seems most applicable considering the general sense in which such characterizations will be given.

TABLE 1. Scale for determining the magnitude of effect sizes in strength training research.*

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	>1.5	>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 as least 5 years.

With this scale, based on the average ES measured in strength training research as well as the variability among such ESs, researchers can now determine the relative magnitude of an ES calculated from a strength training intervention. When this relative magnitude is reported along with the precise statistical probability, both the researcher and the reader are better able to evaluate the overall treatment effect.

To illustrate the benefit of calculating and reporting the ES, consider the following example. Readers gain much more knowledge from a statement such as "a moderate effect size was calculated (ES = 1.5, $p = 0.12$)" than if a researcher simply reported that "no significant differences were found ($p > 0.05$)." In this case, a moderate treatment effect was measured and such an effect could be expected 88 times out of 100. This information allows the professional to make a judgment as to whether or not he or she will accept an intervention that can be expected to elicit a moderate treatment effect (compared with other strength training interventions) 88 times out of 100. If only a "nonsignificant" p value is reported, it could only be concluded that the program was ineffective. In this case, such a conclusion would be incorrect.

CONCLUSION

It is imperative that researchers calculate and report some measure of the treatment effect in strength training research. With the scale provided it is also possible to determine the relative magnitude (small, moderate, or large, etc.) of the ESs calculated (primarily Cohen's d or the standardized mean difference) relative to other strength training research. The main goal of strength training research should be to determine the magnitude of a treatment effect, rather than solely the reproducibility, of the results of a study. With the effect size statistic and the scale provided, researchers can provide more

practical and applicable information to the strength and conditioning professional.

Readers are directed to the following additional suggested readings: (2, 3, 9, 12).

REFERENCES

1. COHEN, J. *Statistical Power Analysis for the Behavioral Sciences* (1st ed.). New York: Academic Press, 1969.
2. COHEN, J. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: L. Erlbaum Associates, 1988. pp. xxi, 567.
3. COHEN, J. Things I have learned (so far). *Am. Psych.* 45:1304–1312. 1990.
4. FOWLER, R. Point estimates and confidence intervals in measures of association. *Psych. Bull.* 98:160–165. 1985.
5. PETERSON, M.D., M.R. RHEA, AND B.A. ALVAR. Maximizing strength development in athletes: A meta-analysis to determine the dose-response relationship. *J. Strength Cond. Res.* 1: 11. 2003.
6. RHEA, M., AND B. ALDERMAN. A meta-analysis of periodized versus non-periodized strength and power training programs. *Res. Q. Exerc. Sport.* In press.
7. RHEA, M.R., B.A. ALVAR, AND L.N. BURKETT. Single versus multiple sets for strength: A meta-analysis to address the controversy. *Res. Q. Exerc. Sport.* 73:485–488. 2002.
8. RHEA, M.R., B.A. ALVAR, L.N. BURKETT, AND S.B. BALL. A meta-analysis to determine the dose-response for strength development. *Med. Sci. Sports Exerc.* 35:456–464. 2003.
9. ROSNOW, R., R. ROSENTHAL, AND D. RUBIN. Contrasts and correlations in effect-size estimation. *Psych. Sci.* 11:446–453. 2000.
10. SALAZAR, W., S.J. PETRUZZELLO, D.M. LANDERS, J.L. ETNIER, AND K.A. KUBITZ. Meta-analytic techniques in exercise psychology. In: *Exercise Psychology.* P. Seraganian, ed. New York: John Wiley, 1993.
11. THOMAS, J., AND J. NELSON. *Research Methods in Physical Activity* (4th ed.). Champaign, IL: Human Kinetics, 2001.
12. THOMAS, J., W. SALAZAR, AND D.M. LANDERS. What is missing in $p < .05$? *Res. Q. Exerc. Sport.* 62:344–348. 1991.

Address correspondence to Matthew Rhea, rhea@suu.edu.