Michael Fröhlich[1] und Andrea Pieter[2]

[1] Sportwissenschaftliches Institut, Universität des Saarlandes
[2] Deutsche Hochschule für Prävention und Gesundheitsmanagement, Saarbrücken

Cohen's Effektstärken als Mass der Bewertung von praktischer Relevanz – Implikationen für die Praxis

Die Originalpublikation wird in englischer Sprache im Int. J. Sports Sci. Eng. erscheinen; eine deutsche Version zusätzlich in PT Zschr. Physiother.

Zusammenfassung

Neben der statistischen Absicherung einer Intervention im Rahmen von experimentellen und quasi-experimentellen Versuchsdesigns im Sinne des Hypothesentestens, spielt die Bewertung der praktischen Relevanz einer Intervention (z.B. Training, medizinische Behandlung, therapeutische Massnahme usw.) im Sinne der Evidence Based Medicine bzw. Evidence Based Therapy eine entscheidende Rolle. Zur Abschätzung von praktischer Relevanz hat sich die Effektstärke nach Cohen (1969) etabliert. Aktuelle Metaanalysen zeigen jedoch, dass die Effektstärken nicht statisch, als feststehende Grösse, sondern dynamisch im Sinne einer zeitlichen Veränderung zu interpretieren sind. Des Weiteren scheinen die Vorerfahrung, die Zielgruppe, die Fragestellung und der Untersuchungskontext einen Einfluss auszuüben, was bei der Bewertung der Effektstärke zu berücksichtigen ist. Um die praktische Bedeutsamkeit einer Massnahme qualitativ abschätzen zu können, sollten zukünftig die Effektstärken bei Trainingsinterventionen stärker beachtet werden. Dabei sind jedoch der jeweilige Untersuchungskontext und theoriegeleitete Kriterien der jeweiligen Forschungsdomäne sowie bereits vorliegende Effektstärken zu berücksichtigen.

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 (e.g. training, medical attendance, therapeutic measure etc.) in the context of evidence based medicine or evidence based therapy plays a major role. Cohen's (1969) effect size has become the standard for assessment. Current meta-analysis show that effect sizes should not be interpreted statically, as a fixed dimension, 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, 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 as well as present effect sizes must be taken into account.

Schweizerische Zeitschrift für «Sportmedizin und Sporttraumatologie» 57 (4), 139–142, 2009

Einleitung

Im Rahmen experimenteller und quasi-experimenteller Versuchsdesigns reicht es in der Regel nicht aus, den Outcome bzw. Output gegenüber einer Kontrollgruppe, einer weiteren Treatmentgruppe oder im Längsschnitt im Sinne der Veränderungsmessung nachzuweisen, sondern die praktische Bedeutsamkeit dieser Effekte spielt eine entscheidende Rolle (Bossmann, 2008). Während dies für den Leistungs- und Spitzensport offensichtlich erscheint und im Kontext der Evidence Based Medicine bzw. Evidence Based Therapy zahlreich dokumentiert wird (u.a. Moher et al., 2001), wird die praktische Relevanz von Trainingsmassnahmen oder allgemein von Interventionen zunehmend im Freizeit-, Breiten- und Gesundheitssport eingefordert. Des Weiteren steht die praktische Bedeutsamkeit einer Intervention im Fokus der Betrachtungen, wenn es darum geht, über verschiedene Einzelstudien hinweg verallgemeinerte Aussagen zu einem Forschungsfeld zu treffen, d.h. wenn die Effektivität von Einzelstudien im Rahmen von Metaanalysen abgeschätzt und interpretiert werden soll (Bortz & Döring, 2006; Rustenbach, 2003).

Unter dem Aspekt der praktischen Bedeutsamkeit wird das Ausmass eines experimentellen Effektes im Hinblick auf verschiedene praktische Belange – u.a. inter- und intraindividuelle Einschätzung, Vergleich zu Normwerten, Veränderungen über die Zeit, Kosten-Nutzen-Relationen usw., – bewertet. Notwendige Voraussetzung, um die praktische Bedeutsamkeit inhaltlich interpretieren zu können, stellt in der Regel der statistisch signifikante Unterschied zwischen den experimentellen Treatmentgruppen, einschliesslich eventueller Kontrollgruppen (randomized controlled studies), im Sinne des Hypothesentestens dar. Die statistische Signifikanzprüfung ist hierbei jedoch von der Varianz und der Stichprobengrösse abhängig, sodass durch wachsende Stichprobengrösse n (u.a. bei Multi-Center-Studien, gross angelegten Panel-Studien usw.) und Senkung der Prüfvarianz die statistische Signifikanz immer herbeigeführt werden kann (Bortz & Döring, 2006). Die statistische Signifikanz sagt somit lediglich etwas über die Existenz eines Effektes, nicht jedoch über

dessen Bedeutsamkeit und Relevanz im Kontext der Fragestellung aus. Pointiert ausgedrückt: Signifikante Ergebnisse müssen auch praktisch bedeutsam sein (Bortz & Döring, 2006). So würde wohl niemand eine neue Trainingsmethode – obwohl signifikant einer bisherigen Trainingsmethode überlegen – zur Anwendung bringen, wenn sich die erzielten Effekte als nicht praktisch relevant zeigen würden (Drinkwater, 2008).

Abschätzung der praktischen Relevanz von signifikanten Effekten

Die praktische Bedeutsamkeit oder Relevanz von signifikanten Effekten im Sinne der Einschätzung des erzielten Outcomes wird im Allgemeinen über das dimensionslose Mass der Effektstärke oder Effektgrösse vorgenommen, da die Effektstärke kaum von der Stichprobengrösse *n* beeinflusst wird (Bortz & Döring, 2006; Leonhart, 2004). Die weiterführenden Ausführungen beziehen sich auf Abstandsmasse bei Gruppenunterschieden, Effektstärken als Zusammenhangsmasse werden nachfolgend ausgeklammert.

Die Effektstärke *d* in der Population normiert im einfachsten Fall die Unterschiede zwischen den unabhängigen experimentellen Gruppen (auch Experimentalgruppe [EG] und Kontrollgruppe [KG]) auf die Streuung der Testwerte (t-Test für unabhängige Stichproben) nach der Formel:

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

Bezieht sich die Effektstärkenschätzung auf die Stichprobenkennwerte der beiden Experimentalgruppen bzw. Treatmentgruppe und Kontrollgruppe wird *d*, wenn die Standardabweichungen der beiden Gruppen annähernd homogen sind, nach Cohen (1969) berechnet:

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

Bezüglich der zu verwendenden Streuung im Nenner der Formel werden unterschiedliche Empfehlungen ausgesprochen (Leonhart, 2004). Glass et al. (1981) bevorzugen die Standardabweichung der Kontrollgruppe $s = s_{KG}$. Bortz und Döring (2006) berechnen die gemeinsame Streuung aus EG und KG nach der Formel:

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

Hedges und Olkin (1985) zeigten, dass eine Standardisierung der Mittelwertsdifferenzen anhand einer gepoolten Standardabweichung beider Gruppen die Schätzung der Effektstärke optimiert (vgl. Rustenbach, 2003). Die gepoolte Standardabweichung $s_{gepoolt}$ berechnet sich dabei nach Leonhart (2004, S. 243):

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

Anhand der einzelnen Formeln lässt sich somit die Effektstärke zuverlässig für die Intervention bzw. das Treatment beispielsweise einer physiotherapeutischen Behandlung, einer Medikation, einer durchgeführten Operation oder eines Trainings usw. im Prä-Post-Testdesign bzw. Prä-Post-Kontrollgruppendesign berechnen. Darüber hinaus weist Rhea (2004a, S. 918) darauf hin, dass die Berechnung und Angabe der Effektstärke im Rahmen von Interventionsmassnahmen zahlreiche Vorteile bietet:

1. Die Effektstärke repräsentiert eine standardisierte Grösse zur Einschätzung und Interpretation von Veränderungen von einzelnen oder mehreren Gruppen (z.B. hat sich die Effektstärke vergrössert oder ist sie gleich geblieben).
2. Die Effektstärke erlaubt den Vergleich verschiedener Trainings-/Interventions- bzw. Behandlungsmethoden innerhalb einer Studie (z.B. lässt sich VO_2max durch ein High Intensity Training genauso gut verbessern wie durch ein aerobes Grundlagenausdauertraining).
3. Die Effektstärke ist leicht zu berechnen und zeigt somit den Einfluss einer einzelnen Studie für die Theorie und Praxis auf.

Im Rahmen von Metaanalysen können mit der Berechnung der Effektstärken die Ergebnisse von einzelnen Primärstudien bspw. zur Wirkungsweise einer Behandlung, eines Trainings oder Medikamentation in Beziehung zu einander gesetzt werden, indem die Effektstärken verglichen werden (Hall & Rosenthal, 1995; Hunter & Schmidt, 2004; Rustenbach, 2003). Letztendlich können so durch die Effektstärken einzelne Studien zu einem gleichen Thema, beispielsweise Behandlungskonzepte bei chronischer Dysfunktion der Tibialissehne (Leumann et al., 2007) bewertet und globale Aussagen zu einem Forschungsgegenstand getroffen werden (Bortz & Döring, 2006). Hinzu kommt, dass Moderatorvariablen wie Geschlecht, Alter, Compliance usw. und deren Einfluss auf einen globalen Effekt abgeschätzt werden können (Hall & Rosenthal, 1995; Morris, 2008; Peterson et al., 2004; Rhea, 2004a; Rhea et al., 2003).

Die einfache Berechnung der Effektstärke *d* sowie deren praktische Bewertung soll anhand eines Beispiels verdeutlicht werden (vgl. Drinkwater, 2008): Durch die Einführung eines neuartigen Sprinttrainings wurde die 100-m-Sprintzeit signifikant um 0.05 s verbessert (Prä-Post-Differenz = 0.05 s). Wie ist nun diese signifikante Zeitverbesserung durch das neuartige Sprinttraining zu interpretieren? Soll die neuartige Trainingsmethode eingeführt werden? Zur Beantwortung dieser Fragen werden zwei Szenarien skizziert:

1. Die Teilnehmer einer Breitensportveranstaltung (Sportabzeichen) haben eine mittlere 100-m-Zeit von 13.04 s (SD = 2.02) und verbessern sich um 0.05 s. Die Berechung der Effektstärke *d* ergibt 13.04 – 12.99/2.02 = 0.025.
2. Die durchschnittliche Finalzeit im 100-m-Sprint der Männer bei den Olympischen Spielen 2008 betrug 9.92 s (SD = 0.11) und verbessert sich um 0.05 s. Die Effektstärke *d* berechnet sich auf 9.92 – 9.87/0.11 = 0.45.

Aus der beispielhaften Effektstärkenberechnung lässt sich somit ableiten, dass die Effektstärke *d* bei den Finalteilnehmern deutlich grösser ist als bei den Teilnehmern der Breitensportveranstaltung und die Einführung der neuen Trainingsmethode bei Weltklasse-Sprintern, im Gegensatz zu durchschnittlich Trainierten, eine praktische Relevanz besitzt. Wie die Effektstärke inhaltlich quantitativ zu bewerten ist, wird nachfolgend erörtert.

Bewertung der praktischen Bedeutsamkeit

Die Bewertung von praktischer Bedeutsamkeit – wie sie im Sprintbeispiel exemplarisch verdeutlicht wurde – hat eine lange Tradition und geht auf die Untersuchungen von Cohen (1969) zurück. Die seitdem auf Konvention festgelegte Effektstärkeklassifizierung weist dabei folgende Werte aus: kleiner Effekt *d = 0.20,* mittlerer Effekt *d = 0.50* und grosser Effekt *d = 0.80* (Cohen, 1969, S. 38; 1992, S. 157). Allgemein werden Effektstärken > *0.50* als gross interpretiert. Effektstärken von *0.50–0.30* als moderat und Effektstärken von *0.30–0.10* als klein bzw. < *0.10* als trivial (Bortz & Döring, 2006, S. 606). Die von Cohen (1969, 1992) vorgeschlagene Klassifikation stellt jedoch nur eine erste Orientierungshilfe dar. Somit wäre die Effektstärke – die Wirkungsweise des neuartigen Sprinttrainings – bei Freizeitsportlern als trivialer Effekt, ohne praktische Bedeutsamkeit, zu interpretieren. Da die Effektstärke bei den Weltklasse-Sprintern bei *d* = 0.45 liegt, ist ein moderater bzw. mittlerer Effekt, mit mittlerer praktischer Bedeutung bzw. Relevanz festzustellen.

Bewertung der praktischen Bedeutsamkeit im Kontext einer Forschungsdomäne

Aktuelle Metaanalysen u.a. von Fröhlich und Giessing (2008), Fröhlich und Schmidtbleicher (2008), Fröhlich et al. (2008), Peterson et al. (2004), Rhea (2004b), Rhea et al. (2003), Rhea und Alderman (2004) konnten zeigen, dass die Effektstärken beispielsweise im Bereich der Krafttrainingsforschung einerseits vom Trainingszustand der Probanden in der dichotomen Kategorie «trainiert» vs. «untrainiert» abhängen *(Tab. 1)* und anderer-

seits durch das Geschlecht und partiell sogar von der Interventionsdauer beeinflusst sind *(Tab. 2)*. Des Weiteren wird die Effektstärke durch die verwendete Trainingsmethode, also indirekt durch die einzelnen Belastungsnormativa, wie Belastungsintensität und Belastungsdauer, bestimmt (Rhea et al., 2003; Peterson et al., 2004). So sind die Effektstärken als Indiz für praktische Bedeutsamkeit im Bereich der Krafttrainingsforschung nicht statisch, sondern eher dynamisch zu verstehen. Des Weiteren zeigen die empirischen Ergebnisse, dass je nach Kontext, Vorerfahrung, Forschungsdomäne usw. die Effektstärken unterschiedlich zu interpretieren sind.

Serien	Trainierte	n	Untrainierte	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

Tabelle 1: Effektstärken (MW ± SD) von trainierten und untrainierten Probanden (Rhea et al., 2003, S. 458).

	Studiendauer [Wochen]				
	1–6	7–12	13–18	19–24	25–30
Einsatztraining	0.76 ± 0.32	1.02 ± 0.71	0.89 ± 1.07	0.76 ± 0.69	1.24 ± 0.34
Mehrsatztraining	0.87 ± 0.38	1.05 ± 0.62	1.23 ± 0.64	0.81 ± 0.47	3.42 ± 2.04

Tabelle 2: Effektstärken (MW ± SD) beim Einsatz- bzw. Mehrsatztraining in Abhängigkeit von der Studiendauer (Fröhlich et al., 2009).

Praktische Konsequenzen

Welche Konsequenzen lassen sich nun aus den bisherigen Ausführungen zur praktischen Bedeutsamkeit und der Effektstärkenklassifizierung ziehen? Erstens sollten neben den statistischen Kennwerten und Prüfgrössen Effektstärken als Masse für praktische Bedeutsamkeit in den publizierten Studien angegeben werden und zweitens müssen die ermittelten Effektstärken je nach Forschungsdisziplinen bzw. Forschungsdomäne beurteilt und eingeschätzt werden. Dies bedeutet, die jeweilige Scientific Community muss letztlich entscheiden, was als «kleiner», «mittlerer» oder «grosser» Effekt zu klassifizieren ist *(Tab. 3)*. Dabei sind theoriegeleitete Kriterien der jeweiligen Domäne ausdrücklich zu berücksichtigen, beispielsweise in analogen Studien zu vergleichbaren Rahmenbedingungen.

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

Tabelle 3: Skala zur Einschätzung der Effektstärken in der Krafttrainingsforschung (Rhea, 2004a, S. 919).

Zusammenfassung

Neben der Angabe von statistischer Signifikanz sollten in empirischen Studien die Effektstärken als quantitative Masse für die Einschätzung von praktischer Bedeutsamkeit angegeben werden (Leonhart, 2004). Die Effektstärken sind jedoch je nach Forschungsdisziplin bzw. Domäne zu spezifizieren und dynamisch zu interpretieren, das bedeutet, bereits vorliegende Effektstärken von Interventionen sind für die eigene Ergebnisdiskussion zu nutzen. So resultieren beispielsweise bei Krafttrainingsanfängern im Vergleich zu Fortgeschrittenen bzw. Könnern hohe Effektstärken, welche die hohen Krafttrainingszuwächse in den ersten Monaten erklären. Im weiteren Trainingsverlauf fallen die entsprechenden Leistungszuwächse – dies dürfte auch für Lernprozesse gelten – entsprechend geringer aus, was in geringeren absoluten Effektstärken zum Ausdruck kommt. Wie exemplarisch anhand der Krafttrainingsforschung gezeigt wurde, müssen die Effektstärken je nach Zielgruppe, Setting und Fragestellung unterschiedlich interpretiert werden. Existieren im jeweiligen Forschungsgebiet übliche Effektstärken, so sollten diese zur Einschätzung der eigenen Studienergebnisse herangezogen werden. Gerade die angewandten Wissenschaften, hier explizit die Sportwissenschaft, sollten nicht nur die gefundenen Ergebnisse statistisch auf Signifikanz absichern, sondern zusätzlich die praktische Relevanz von einzelnen Effekten dokumentieren und dies vor dem Hintergrund der hier diskutierten Faktoren reflektieren.

Korrespondenzadresse:

Michael Fröhlich, Sportwissenschaftliches Institut, Universität des Saarlandes, Universität Campus Geb. B8.1, D-66123 Saarbrücken; Fon: +49 (0) 681-302 4911; Fax: +49 (0) 681-302 4915, Mail: m.froehlich@mx.uni-saarland.de

Literaturverzeichnis

Bortz J., Döring N. (2006): Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. Springer, Berlin, Heidelberg, 4. Auflage.
Bossmann T. (2008): Was ist ein d-Mass? Zschr Physiother. 60: 47–49.
Cohen J. (1969): Statistical power analysis for the behavioral sciences. Academic Press, New York, London.
Cohen J. (1992): A power primer. Quant. Meth. Psychol. 112: 155–159.
Drinkwater E.J. (2008): Applications of confidence limits and effect sizes in sport research. Open Sports Sci. J. 1: 3–4.
Fröhlich M., Giessing J. (2008): The effectiveness of single-set vs. multiple-set training – A meta-analytical consideration. In: Current results of strength training research. A multi-perspective approach, J. Giessing and M. Fröhlich (eds.), Cuvillier Verlag, Göttingen, pp. 9–33.
Fröhlich M., Schmidtbleicher D. (2008): Trainingshäufigkeit im Krafttraining – ein metaanalytischer Zugang. Dtsche Zschr. Sportmed. 59: 34–42.
Fröhlich M., Emrich E., Schmidtbleicher D. (2009): Outcome effects of single-set vs. multiple-set training – an advanced replication study. Res. Sports Med. (in press)
Fröhlich M., Giessing J., Schmidtbleicher D., Emrich E. (2008): A comparison between 2 and 3 days of strength training per week – A metaanalytical approach. In: Current results of strength training research. A multi-perspective approach, J. Giessing and M. Fröhlich (eds.), Cuvillier Verlag, Göttingen, pp. 151–166.
Glass G.V., McGaw B., Smith M.L. (1981): Meta-analysis in social research. Sage, Beverly Hills.
Hall J.A., Rosenthal R. (1995): Interpretation and evaluating meta-analysis. Eval. Health Prof. 18: 393–407.
Hedges L.V., Olkin I. (1985): Statistical methods for meta-analysis. Academic Press, New York, London.
Hunter J.E., Schmidt F.L. (2004): Methods of meta-analysis: correcting error and bias in research. Sage, Newbury Park.
Leonhart R. (2004): Effektgrössenberechnung bei Interventionsstudien. Rehabilitation 43: 241–246.
Leumann A., Merian M., Wiewiorski M., Hintermann B., Valderrabano V. (2007): Behandlungskonzepte der chronischen Dysfunktion der Tibialis posterior-Sehne. Schweiz. Zschr. Sportmed. Sporttraumatol. 55: 19–25.
Moher D., Schulz K.F., Altman D.G. (2001): The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trails. Lancet 357: 1191–1194.
Morris S.B. (2008): Estimating effect sizes from pretest-posttest-control group designs. Organiz. Res. Meth. 11: 364–386.

Peterson M.D., Rhea M.R., Alvar B.A. (2004): Maximizing strength development in athletes: a meta-analysis to determine the dose-response relationship. J. Strength Cond. Res. 18: 377-382.
Rhea M.R. (2004a): Determining the magnitude of treatment effects in strength training research through the use of the effect size. J. Strength Cond. Res. 18: 918–920.
Rhea M.R. (2004b): Synthesizing strength and conditioning research: the meta-analysis. J. Strength Cond. Res. 18: 921–923.
Rhea M.R., Alderman B. (2004): A meta-analysis of periodized versus nonperiodized strength and power training programs. Res. Q. Exerc. Sport 75: 413–422.
Rhea M.R., Alvar B.A., Burkett L.N., Ball S.D. (2003): A Meta-analysis to determine the dose response for strength development. Med. Sci. Sports Exerc. 35: 456–464.
Rustenbach S.J. (2003): Metaanalyse. Eine anwendungsorientierte Einführung. Verlag Hans Huber, Bern, Göttingen, Toronto, Seattle.