Effektstärken: Statistische, praktische und theoretische Bedeutsamkeit empirischer Studien

Effect size:
Statistical, practical, and theoretical significance of empirical studies

Georg Lind[1]

2012 [2]

"A picture is more worth than thousands of p-values: On the irrelevance of hypothesis testing in the computer age." (Loftus, G. R., 1993)

"What's wrong with significance testing? Well, among many other things, it does not tell us what we want to know, and, out of desperation, we nevertheless believe in that it does!" (Cohen, 1994, S. 997)

"Für den sinnvollen Einsatz der Inferenzstatistik ist es erforderlich, dass vor Untersuchungsbeginn eine theoretisch gut begründete Hypothese oder Fragestellung formuliert wurde." (Bortz, 1994, S. 2)

"Was wir unseren Studenten lehren sollten ist statistisches Denken: Wie man mutige Hypothesen formuliert, präzise alternative Hypothesen ableitet, Experimente so plant, dass wirkliche Messfehler minimiert werden (als sie nur zu messen und in den F-Bruch einzusetzen), Daten für jedes Individuum separat analysiert statt sie automatisch zu Mittelwerten zusammen zu fassen, und sinnvolle deskriptive Kennzahlen (Statistiken) und erkundende Datenanalysen zu verwenden" (Gigerenzer, 1998, S. 200; meine Übersetzung)

[1] Kontakt: Prof. em. Dr. Georg Lind, E-Mail: Georg.Lind@uni-konstanz.de Web: www.uni-konstanz.de/ag-moral/

[2] Dieses Papier wurde angeregt durch die Arbeiten an meiner Dissertation:
http://www.uni-konstanz.de/ag-moral/pdf/Lind-1985_Inhalt-und-Struktur.pdf
und dem Vortrag, den ich 1991 an der Katholischen Universität Eichstätt im Rahmen meiner Habilitierung gehalten hatte:
http://www.uni-konstanz.de/ag-moral/pdf/Lind-1991_Empirischer-Gehalt-von-Hypothesen.pdf.
Die erste Version dieses Papiers hat inzwischen mehrere Revisionen und Ergänzungen erfahren. Größere Korrekturen: Ergänzungen zur absoluten ES (27.10.2008). Redaktionelle Ergänzungen (16.11.2009); Korrektur von Tippfehlern auf Seite 4 (29.6.2010).

Wie wissen wir, ob ein empirisch gefundener Effekt "signifikant" oder "bedeutsam" ist?

Um diese Fragen zu beantworten, kann man mehrere Möglichkeiten in Erwägung ziehen. In der Psychologie und vielen Sozialwissenschaften wird fast ausschließlich eine formale, statistische Beantwortung in Betracht gezogen: Ist der Befund "statistisch bedeutsam"? Was aber selten (viel zu selten!) in Erwägung gezogen wird, ist die Möglichkeit, Befunde auf ihre theoretische, inhaltliche Bedeutsamkeit hin zu untersuchen: Welche Wertedifferenz ist für unser subjektives Empfinden und unsere Handlungen bedeutsam? Ab welcher Effektstärke können wir davon sprechen, dass eine Therapiemethode oder eine pädagogische Intervention wirklich etwas bringen und den Aufwand lohnen, den alle Beteiligten investieren müssen? Tritt der Effekt immer oder nur unter bestimmten Bedingungen auf? Ist er an Besonderheiten der Studie (Umfang des Samples, Streuung der unabhängigen Variablen) gebunden? Passt der Effekt zu dem, was wir bereits über die Variablen wissen, die wir untersuchen oder stellt er die fundiert geglaubten Theorien in Frage?

Kurzum: Auch wenn man einmal annimmt, dass die Bestimmung der statistischen Bedeutung ("Signifikanz") korrekt durchgeführt wird, enthebt dieses Vorgehen den Wissenschaftler und den Praktiker, der empirische Daten interpretieren und zur Grundlage von Handlungsentscheidungen machen will, nicht von der Frage, ob der Befund auch inhaltlich, fachwissenschaftlich bedeutsam ist. Wenn wir am Thermometer ablesen, dass die durchschnittliche Temperatur heute gegenüber gestern um fünf Grad gestiegen ist, können wir fragen, ob dieser Unterschied statistisch signifikant ist (eine Frage, die wir natürlich nur beantworten können, wenn wir an jedem der beiden Messzeitpunkte mehrere Messungen vorgenommen haben). Tatsächlich sind wir im Alltag an dieser Frage kaum interessiert, sondern vielleicht an der Frage, ob fünf Grad Unterschied bestimmte Verhaltensentscheidungen notwendig macht, wie zum Beispiel eine etwas weniger warme Kleidung anzuziehen. In diesem Sinne sind wir auch weniger daran interessiert, ob zwei Länder, in denen einige Tausend Schüler an Schulleistungstests teilgenommen haben, sich "signifikant" unterscheiden, sondern vielleicht eher daran, ob diese Unterschiede so groß sind, dass sie die Wirtschaftskraft der Länder oder die Lebensqualität der Menschen, die darin leben, merklich beeinflussen. Die Frage nach der statistischen Signifikanz interessiert mehr den Forscher, der überlegt, wie groß eine Untersuchungsgruppe sein muss, damit ein bestimmter Unterschied, der ihm (aus inhaltlichen Gründen!) wichtig ist, zweifelsfrei nachgewiesen werden kann, also größer ist als die Genauigkeit seines Messinstruments es erlaubt. Wenn ein Unterschied statistisch nicht "signifikant" geworden ist, heißt das also, dass er kleiner ist als der Unterschied, der als bedeutsam festgelegt wurde. (Diese inhaltliche Festlegung ist allerdings nicht mit dem so genannten "Signifikanzniveau", z.B. $\alpha = 5\%$ zu verwechseln!)

Es ist also höchst problematisch, von einer *statistischen* Signifikanz von Befunden direkt auf deren *theoretische* und *praktische* Bedeutsamkeit zu schließen. So genannte Maße der "praktischen Signifikanz" oder "relativen Effektstärke" taugen dafür schon eher, obwohl auch sie rein statistische Verfahren sind, die eine theoretische und praktische Bewertung eines Befunds nicht ersetzen können. Je nach Fragestellung und Anwendungskontext können selbst sehr geringe Effektstärken von großer praktischer Bedeutung sein. So werden Beispielsweise in der Medizin dann Medikamente zugelassen, wenn ihre Effektstärke relativ gering ist (r = 0.15), wenn sie gegen besonders schlimme Krankheiten eingesetzt werden können. Bei absoluten Effektstärken kann die Beobachtung von sehr geringen Effektstärken (wie z.B. bei der Ablenkung des Lichts durch große Massen festgestellt werden) eine sehr große praktische Bedeutung haben, z.B., wenn es darum geht die genaue Startrichtung einer Mars-Sonde zu berechnen.

In diesem Papier werden Maße der relativen Effektstärke dargestellt und die Formel für ihre Berechnung aus konventionellen Maßen der statistischen Signifikanz. Außerdem wird auf das Konzept der "absoluten Effektstärke" eingegangen und auf die Frage, wie die praktische und theoretische Signifikanz eines Befundes ermittelt werden kann.

Ein Beispiel

Nehmen wir ein Beispiel: Wir wollen wissen, ob eine bestimmte Unterrichtsmethode geeignet ist, die moralische Urteilsfähigkeit zu verbessern. Vor und nach der Anwendung der Methode werden die Schüler mit dem Moralisches-Urteil-Test (MUT; Lind, 2008) gemessen. Es ergeben sich folgende Mittelwertunterschiede und Standardabweichungen:

$M_1 = 25{,}5$, $s = 12{,}1$
$M_2 = 31{,}7$, $s = 12{,}1$
Differenz = $M_1 - M_2 = 31{,}7 - 25{,}5 = 6{,}2$ (C-Punkte).
$N_1 = 15$, $N_2 = 15$

Berechnet man wie in vielen Studien nur die statistische Signifikanz (hier wird der Einfachheit halber der t-Test für unabhängige Stichproben verwendet, obwohl natürlich ein t-Test für abhängige Stichproben richtig wäre), dann ist die Antwort auf die Frage, ob das "signifikant" sei, folgende:

$$t = \frac{M_2 - M_1}{SE_{pooled}} = 6{,}2 \,/\, 7{,}1 = 0{,}87$$

Dieser t-Wert hat ein p = 0,086, d.h. der Wertezuwachs von 6,2 Punkten im MUT ist "nicht signifikant auf dem 5-Prozent-Niveau". Dafür hätte *p* gleich oder kleiner als *p* = 0,05 (α-Niveau) sein müssen.

Eine C-Wert-Differenz von 6,2 Punkten ist aber, so denkt der Experimentator, relativ viel. Warum ist die Differenz nicht signifikant geworden? Die Antwort ist klar: die Standardabweichung ist relativ groß und der Stichprobenumfang relativ klein. Die Prüfgröße für die statistische Signifikanz, d.h. die dem *t*-Wert zugehörige Wahrscheinlichkeit (*p*-Wert), wird umso größer (also eher "signifikant"),

a) desto größer der *t*-Wert ist (der wiederum umso größer wird, *je kleiner die Standardabweichung* s , d.h. die Standardabweichung S_e der Mittelwerte ist), und
b) desto größer die Stichprobengröße N ist.

Ja, sagt sich der Experimentator, jetzt weiß ich, wie ich das nächste Mal ein "signifikantes Ergebnis" bekommen kann. Ich erhöhe die Stichprobengröße und nehme 50 statt 15 Schüler *pro Gruppe*. Das fiktive Experiment führte zu folgendem Ergebnis:

$M_1 = 25{,}5$, $s = 12{,}1$
$M_2 = 31{,}7$, $s = 12{,}1$
Differenz = $M_2 - M_1 = 31{,}7 - 25{,}5 = 6{,}2$ (C-Punkte)
$N_1 = 50$, $N_2 = 50$

Tatsächlich wird dieselbe Differenz "hoch signifikant": $p = 0{,}006 \; | < 0{,}01$ (α-Niveau, zweiseitig).

Aber, so mag der Experimentator einwenden, die Versuchspersonenzahl zu erhöhen kostet viel Geld und Zeit. Ich könnte doch auch versuchen, die Streuung der C-Werte in meiner Experimentalgruppe kleiner zu machen, indem ich Personen mit sehr kleinen und sehr großen C-Werten aus dem Versuch herausnehme (Eliminierung von so genannten "Ausreißern"). Gesagt getan. Durch diese manchmal durchaus als legitim erachtete Maßnahme reduziert sich die Standardabweichung auf s = 6,2.

$M_1 = 25{,}5$, $s = 6{,}2$
$M_2 = 31{,}7$, $s = 6{,}2$
Differenz = $M_2 - M_1 = 31{,}7 - 25{,}5 = 6{,}2$ (C-Punkte)
$N_1 = 15$, $N_2 = 15$

Und auch dieses Ergebnis ist "signifikant": $p = 0{,}0106 \; | < 0{,}05$ (α-Niveau).

Wie wir gesehen haben, ist der Effekt der pädagogischen Intervention eigentlich immer gleich geblieben: die Testwerte sind genau 6,2 Punkte gewachsen. Aber die "Signifikanz" des Ergebnisses hat sich jedes Mal geändert. Wenn die Stichprobe vergrößert oder die Streuung der Werte verkleinert wurde, wurde aus einem "nicht signifikanten" Ergebnis ein "signifikantes". Man könnte auch sagen, während die psychologische Signifikanz gleich geblieben ist, hat sich nur die statistische geändert. Das kann eigentlich nur jemanden verwundern, der nicht weiß, wofür statistische Signifikanztests eigentlich gut sind.

Noch ein Beispiel dafür, wie sich bei Vergrößerung der Stichprobe die "Signifikanz" ändert, ohne dass sich etwas an dem Zusammenhang zweier Variablen ändert.

N = 26

Zwischen bei Variablen (A und B) besteht folgender Zusammenhang:

	A = 0	A = 1
B = 0	10	5
B = 1	5	6

Zusammenhang: Phi = **0,21**

Statistische Signifikanz: n.s. (**p = 0,279**)

N = 260

Zwischen bei Variablen (A und B) besteht folgender Zusammenhang:

	A = 0	A = 1
B = 0	100	50
B = 1	50	60

Zusammenhang: Phi = **0,21**

Statistische Signifikanz: sig. 1%-Niveau (**p = 0,00063**)

Wofür brauchen wir statistische Signifikanztests?

Die statistischen Signifikanztests dienen dazu, die Präzision eines Stichprobenergebnisses im Hinblick auf eine wohldefinierte Grundgesamtheit und einen (auf Grund von inhaltlichem Vorwissen) bestimmten Schwellenwert für die theoretische Signifikanz eines Differenzwertes zu bestimmen (vgl. u.a. Hays, 1963). Das ist zum Beispiel der Fall, wenn wir die Frage untersuchen möchten, ob sich die Testwerte "moralische Urteilsfähigkeit" zwischen den Populationen aller 15-jährigen Deutschen und Franzosen inhaltlich bedeutsam unterscheiden und wir statt der gesamten Population aus jeder Grundgesamtheit nur eine Zufallsstichprobe ziehen möchten. Was eine inhaltlich oder praktisch bedeutsame Differenz ist, muss hier *vor* der Untersuchung festgelegt werden, damit ermittelt werden kann, wie große die Stichprobe sein muss, um diesen Unterschied mit ausreichender Präzision messen

zu können. Statistische Signifikanztests sind dagegen fehl am Platz, a) wenn gar keine Grundgesamtheit definiert ist (oder definierbar ist) b) wenn aus einer solchen wohldefinierten (!) Grundgesamtheit keine wirkliche Zufallsstichprobe gezogen wurde, und wenn nicht klar ist, welcher Differenzwert wirklich bedeutsam ist. Diese Bedingungen sind in den Sozialwissenschaften oft nicht gegeben, weswegen das, was eigentlich zuerst geklärt werden müsste, nämlich die Frage nach der praktischen und theoretischen Signifikanz eines Unterschieds, oft zuletzt kommt oder völlig ausgeklammert wird, weil man irrigerweise meint, die statistische Signifikanz gäbe uns darüber Auskunft.

Aus diesen Gründen empfiehlt Thompson (1994), einer der führenden Experten auf dem Gebiet der sozialwissenschaftlichen Statistik, folgende Sprachregelung, die ich sehr hilfreich finde:

"Die Überwindung von drei Sprachgewohnheiten kann unbewusste Fehlinterpretationen verhindern helfen:

* Sprich immer von 'statistischer Signifikanz' und nicht einfach von 'Signifikanz'. Dies kann helfen, die irrtümliche Assoziation zwischen der Rückweisung einer Nullhypothese und einem wichtigen Befund aufzubrechen.

* Sprich nicht von Dingen wie "mein Ergebnis nähert sich der statistischen Signifikanz". Eine solche Sprache macht im Rahmen der statistischen Test-Logik kaum einen Sinn.

* Sprich nicht von Dingen wie 'der statistische Signifikanztest sagt uns, ob die Ergebnisse zufällig waren'. Diese Sprache erzeugt den Eindruck, als wenn statistische Signifikanztests etwas über die Replizierbarkeit eines Befundes aussagen würden." (*Thompson, 1994, Meine Übersetzung*)

Auch die Einführung von "Power analysis" (Cohen, 1988) ändert wenig an diesem Missbrauch der statistischen Signifikanztests. Zwar korrigiert sie die Voreingenommenheit des üblichen Gebrauchs von Signifikanztests gegenüber der Nullhypothese (die man dort immer zu widerlegen versucht, als wenn das ein wichtiges Ziel in der Forschung sei). Aber auch diese Analyse hat wenig mit der praktischen oder theoretischen Signifikanz von Forschungsergebnissen zu tun.

Als Alternative zu den statistischen Signifikanztests wurden Maße der "praktischen Signifikanz" (Bredenkamp, 1970) bzw. der "Effektstärke" (Effect size) vorgeschlagen. Diese Maße haben große Vorteile gegenüber dem reinen Gebrauch von statistischen Signifikanztests, aber auch sie sind nicht optimal, da sie – trotz des Anscheins den der Begriff "praktische Signifikanz" erweckt – weder über praktische noch über theoretische Signifikanz etwas aussagen. Ich verwende daher lieber den Begriff der Effektstärke, der sich in der Literatur auch durchgesetzt hat.

Maße für die relative Effektstärke

Effektestärkemaße sind so konstruiert, dass sie unabhängig von der Größe der Stichprobe sind und damit einen wichtigen Nachteil der statistischen Signifikanzmaße ausschalten. Mit anderen Worten: die Effektstärke einer Intervention ist unabhängig davon, an wie vielen Versuchspersonen man sie getestet hat. Damit machen diese Maße auch Untersuchungen miteinander vergleichbar, die verschieden große Samples benutzt haben. Diese Maße hängen aber nach wie vor stark von der Streubreite der Werte der abhängigen Variablen in der jeweiligen Untersuchungsgruppe ab, weshalb wir man auch als *relative* Effektstärkemaße bezeichnen sollte.

Am häufigsten werden zwei Maße verwendet, der sogenannte *d*-Wert und der Korrelationskoeffizient *r*. Der *d*-Wert wurde von Glass et al. (1978) vorgeschlagen, die ihn in Meta-Analysen verwendet haben, wo er auch heute noch vielfach anzutreffen ist. Er ist definiert als die Differenz von Mittelwerten in Relation zur "gemeinsamen" (gepoolten) Standardabweichung (gemeinsame SD aus erster und zweiter Messreihe):

$$d = \frac{M_2 - M_1}{SD_{pooled}}$$

An dieser Formel erkennt man auch schön die Abhängigkeit dieses Maßes von der Streuung der Werte in dem Sample. Der *d*-Werte ist nach unten und oben nicht begrenzt, weswegen er oft schwer einzuschätzen ist.

Diese Manko behebt der den meisten empirisch forschenden Sozialwissenschaftlern vertraute (nichtlineare) Korrelationskoeffizient *r,* der von -1.0 bis +1.0 variieren kann. Bei +1 liegt eine maximale Effektstärke vor, bei 0 absolut keine und bei -1 eine maximal negative. Das Maß wird von vielen heute viel verwendet, um Effektstärke auszudrücken (Thompson 1996). Es wird – im Unterschied zu der so genannten Korrelationsforschung – nicht darauf beschränkt, den linearen Zusammenhang zwischen einer Intervention und den gefundenen Messwerten auszudrücken, sondern auch, um andere, nichtlineare Zusammenhänge darzustellen. Der Zusammenhang mit dem Maß *d* ist über folgende Formel leicht hergestellt (Cohen, 1988, S. 23):

$$r = \frac{d}{\sqrt{d^2 + 4}}$$

Bei kleinen Stichproben und bei ungleichen Stichprobengrößen wird die etwas kompliziertere Formel von Aaron, Kromrey und Ferron (1998) vorgeschlagen:

$$r = \frac{d}{\sqrt{d^2 + (N^2 - 2N)/(n_1 n_2)}}$$

Im Anhang werden weitere Formeln angegeben, mit denen man Signifikanzwerte (Chi^2, t, F, etc.) in den Korrelationskoeffizient r als Maß für Effektstärke umrechnen kann, sofern die dafür notwendigen Informationen (vor allem der Stichprobenumfang N) vorhanden sind (siehe Anhang).

Maße der relativen statistischen Effektstärke sind besser als die der statistisches Signifikanz, weil sie nicht von der Stichprobengröße abhängig sind. Damit lassen sich Studien besser miteinander vergleichen, die unterschiedliche Samplegrößen nutzten. Diesen Vorteil macht man sich in Meta-Analysen zunutze.

Aber auch relative Effektstärkemaße sind nicht optimal:

1. Das Problem, dass die Streuungen der abhängigen Variablen von Studie zu Studie oft stark unterscheidet.[3] Damit ist die Vergleichbarkeit stark eingeschränkt (was viele Autoren aber nicht davon abhält, Vergleiche anzustellen und Meta-Analysen zu rechnen, weil ihnen die Voraussetzungen dieser Methode gar nicht bewusst sind oder sie meinen, dass das keinen großen Einfluss auf das Ergebnis hat, was aber jeweils zu beweisen wäre.) Unterschiedliche relative Effektstärken (*d* oder *r)* müssen also keine unterschiedlichen Effekte anzeigen. Sie können auch beeinflusst sein durch unkontrollierte oder unkontrollierbare oder bewusst herbeigeführten Unterschiede bezüglich der Streubreite der Werte der abhängigen Variablen.

2. Die weit verbreiteten relativen Effektstärkemaße "d" und "r" bilden nur *symmetrische* und *monokausale* Beziehung optimal ab. Symmetrisch ist eine empirische Beziehung dann, wenn sie in beiden Richtungen (von der unabhängigen Variablen auf die abhängige und umgekehrt) gleich hoch ist. Dies ist meist nur dann der Fall, wenn die unhängige Variable (z.B. Schulbildung) für die abhängige Variable (z.B. moralische Urteilsfähigkeit) sowohl eine *notwendige* wie auch eine *hinreichende* Bedingung darstellt. Das hieße zum Beispiel, dass eine hohe moralische Urteilsfähigkeit nie auftreten dürfte, wenn die Bildung niedrig ist (da Bildung als *notwendig* angesehen

[3] Lehrbücher der Statistik behandeln meist nur den Fall, dass die Streuung der abhängigen Variablen zwischen den Kategorien verschieden groß sind (Homoskedastizität). Das viel wichtiger Problem der verschiedener Varia nzen der unabhängigen Variablen scheint kaum Wert behandelt zu werden, da man offenbar mehr daran interessiert ist, von Sample-Daten auf Grundgesamtheiten zu schließen als, daran, geeignete Maße für die Bewertung von Kausalhypothesen oder Behandlungseffekten bereit zu stellen.

wird), und dass sie auch keiner anderen Bedingung bedarf, also allein den Effekt auslösen kann (da Bildung *hinreichend* angesehen wird). Man sieht, dass diese Annahmen eher unrealistisch, da Beziehung zwischen Variablen meist asymmetrisch und multi-kausal oder durch andere Variablen bedingt sind.

3. Diesem Einwand kann man zum Teil Rechnung tragen, indem man für die Berechnung der relativen Effektstärke (im Fall von experimentellen Designs) die statistische Methode der *Varianzanalyse* oder (im Fall von Feldstudien) die Methode der *Regressionsanalysen* heranzieht. Bei der Regressionsanalyse zeigt der ß-Koeffizient (ß = beta) an, um wie viele Einheiten sich die abhängige Variable (im Beispiel: moralische Urteilsfähigkeit) vermehrt, wenn die unabhängige Variable (im Beispiel der Umfang an Bildung) um eine Einheit erhöht wird. Aber auch diese Maße sind nicht unproblematisch, da sie voraussetzen, dass die unabhängige Variable eine hinreichend Bedingung für den Effekt ist (also keine bedingten Effekte oder so genannte 'Interaktions-Effekte' abgebildet werden) und weil auch sie von der Streuung der Werte der unabhängigen Variable abhängen. Bei der Varianzanalyse kann der Anteil der Varianz zwischen den Gruppen in ähnlicher Weise interpretiert werden.

4. Ein weiteres Problem entsteht bei allen Maßen der *relativen Effektstärke*, wenn man die Wirkung mehrerer "unabhängiger"[4] Variablen miteinander vergleichen will, da diese unabhängigen Variablen untereinander korrelieren können. Um eindeutige Werte zu erhalten, müssen diese Korrelationen "kontrolliert", das heißt, auf Null gebracht werden. Die kann auf zwei Wegen erreicht werden, zum einen, indem man die Studie als *Experiment* anlegt, das heißt, indem man durch eine geeignete Versuchsanlage die unabhängigen Variablen statistisch unabhängig macht ("orthogonalisiert"). Wenn z.B. die Einflüsse von Geschlecht und Bildung auf die moralische Entwicklung verglichen werden sollen, kann man die mögliche Korrelation zwischen Geschlecht und Bildung dadurch neutralisieren, dass man für jede Kombination dieser Variablen genau dieselbe Zahl von Fällen einbezieht, wie das in folgender Tabelle illustriert wird:

[4] "Unabhängig" bezeichnet nicht die Beziehung dieser Variablen oder Faktoren zu der "abhängigen" Variablen, wie in manchen Publikationen irrtümlich steht, sondern auf die statistische Beziehung der Faktoren untereinander. Mit unabhängig und abhängig ist auch zunächst nur eine statistische Beziehung gemeint. Die Frage, ob dieser statistischen Beziehung auch eine wirklich kausale Beziehung entspricht, kann man nur durch Heranziehen anderer Befunde und einer erklärenden Theorie beantworten. Wie Popper (1968) nachwies, kann aber niemals absolute Gewissheit erlangt werden. Jede als sicher angenommene Erkenntnis kann sich aufgrund neuer Forschung später einmal als falsch herausstellen.

	Männer	Frauen
Ohne Abitur	20	20
Mit Abitur	20	20

Hier ist die Korrelation zwischen Geschlecht und Bildung genau null. Wenn jetzt eine der beiden Variablen, zum Beispiel Geschlecht, mit einer dritten Variablen (z.B. Schulerfolg) korreliert, kann man diese Korrelation ganz dieser Variablen zuschreiben, ohne daran denken zu müssen, dass ein Teil der Korrelation sich vielleicht durch die Korrelation zwischen den beiden "unabhängigen Variablen" "erklären" lässt. Es geht hier also um ein rein methodisches Verfahren, um die statistischen Korrelationen klarer zuordnen zu können. Ein solches *experimentelles* Design ("experimentell", weil es absichtlich so eingerichtet wurde) bietet dazu noch die Möglichkeit, mittels Varianzanalyse *bedingte* (Interaktions-) Effekte zu messen.

Wenn keine experimentelle Anlage der Studie möglich ist, sondern eine *Feldstudie* vorliegt, bei der die unabhängigen Variablen untereinander korrelieren, können die Korrelationen zwischen den unabhängigen Variablen heraus gerechnet ("kontrolliert") werden. In solchen Fällen, die eher die Regel als die Ausnahme sind, hängt die Größe der Effektstärke aber davon ab, in welche Reihefolge man die unabhängigen Variablen kontrolliert und welche Variablen man kontrolliert. Auch hier bestehen wieder viele Einflüsse, die mit dem "wahren" Effekten nichts zu tun haben. Dem Verdacht der Beliebigkeit oder gar der Manipulation kann man nur entgehen, wenn präzise inhaltliche Theorien zur Verfügung stehen, aus denen man gehaltvolle Hypothesen ableiten kann.

5. Alle Maße der relativen Effektstärke (wie auch die Maße der absoluten Effektstärke) hängen von der Verteilung der *unabhängigen* Variablen ab. Im Extremfall gilt: Wo keine Varianz in der unabhängigen Variable ist, kann sich auch kein Effekt zeigen. Beispiel: Wenn nur Jungen untersucht werden, aber keine Mädchen, kann auch nicht der Effekt der Variable 'Geschlecht' auf die moralische Urteilsfähigkeit oder die Mathematikleistung studiert werden. Aber auch sonst gilt: je geringer die Varianz der Werte der unabhängigen Variablen, umso geringer ist meist der gemessene Effekt (wohlgemerkt: nicht der 'wirkliche' Effekt), da der 'wahre' Effekt umso stärker von Mess- und Stichprobenfehlern überlagert ist, umso geringer die Varianz der unabhängigen Variable ist. Diese Problem kann man zu lösen versuchen, indem man die Untersuchungsgruppen bezüglich bestimmter unabhängiger Variablen sorgfältig (zum Beispiel: gleich viele Männer und Frauen oder gleich viele Personen mit geringer, mittlerer und hoher Bildung) auswählt. Das ist aber nicht immer möglich. Versucht werden sollte es aber auf jeden Fall.

Was kann man tun, wenn die Untersuchung bereits durchgeführt ist und daher keine Änderungen mehr am Design der Studie vorgenommen werden können? Wie kann man solche Studien möglichst gut analysieren? Hier einige Tipps, die aus den obigen Erläuterungen folgen:

- Nicht allein auf statistische Signifikanzen vertrauen! Sie hängen von der Größe der Stichprobe und der Varianz der abhängigen Variablen ab. Beide Größen schwanken stark von Studie zu Studie und können auch bewusst "manipuliert" worden sein, um große Effekte zu erhalten.
- Maße der relativen Effektstärke sind besser und sie können meist aus der statistischen Signifikanz nachträglich berechnet werden. Aber auch sie haben ihre Probleme. Meist unterschätzen sie den 'wahren Effekt', da die Varianzen der abhängigen Variablen oft stark von einander abweichen, da die Varianz der unabhängigen Variablen oft gering ist, und da sehr restriktive Vorstellungen über die Beziehung zwischen Variablen (symmetrisch und mono-kausal) vorausgesetzt werden, ohne dass der Forscher sich dessen oft bewusst ist.
- Eine besser Vorstellung von der praktischen Bedeutsamkeit eines Effekts bekommt man durch die *binomialen Effektstärkeabbildung (BESD – binomial effect size display)* von Rosenthal und Rubin (1982) und durch absolute Effektstärken wie (Mittelwert-) Differenzen.

Maße für die absolute Effektstärke

Im Alltag, aber auch in den Naturwissenschaften wird meist anders vorgegangen als in der psychologischen Forschung. Statt statistische Signifikanzwerte und relative Korrelationen auszurechnen, wird dort oft einfach die Messdifferenzen abgelesen (oder es werden Mittelwerte über kleine Messreihenberechnet, wenn Einzelbeobachtungen zu ungenau sind). Die Effektstärke ist dann die Differenz zwischen zwei Messungen oder zwei Mittelwerten von Messreihen. Maße für die absolute Effektstärke werden also gebildet, indem Differenzwerte zwischen Vorher-Nachher-Messungen berechnet werden, ohne dass dafür die Streuung der abhängigen Variablen herangezogen wird.

Beispiel: Wir wollen wissen, ob es heute wärmer als gestern ist. Dazu vergleichen wir die (durchschnittliche) Temperatur von gestern mit der (durchschnittlichen) Temperatur von heute. Wenn es gestern 18 Grad war und heute 23 Grad, dann beträgt die Erwärmung 5 Grad (= 23 - 18). Im Alltag sind wir an der *praktischen Signifikanz* dieser Differenz interessiert. 5 Grad Differenz scheinen für die meisten Menschen bedeutsam zu sein, da hieraus die Veränderung verschiedener Verhaltensweisen folgt: Man "spürt die Erwärmung. Man zieht sich nicht mehr so warm an. Bei zwei Grad Differenz würde vermutlich nichts verspürt und kein anderes Verhalten folgen. In der Meteorologie spielen

hingegen schon wesentlich geringere Schwankungen der Temperatur eine Rolle, zum Beispiel die Veränderung der mittleren Jahrestemperatur. Schon wenige Zehntel Grad Erwärmung kann das Abschmelzen großer Gletscher und Überschwemmungen in vielen Küstenbereichen bedeuten.

Die Berechnung der absoluten Effektstärke (aES) basiert auf einfachen Mittelwertvergleichen bzw. Differenzbildungen: $M_2 - M_1$. Welche Mittelwerte verglichen werden, hängt von der Fragestellung und dem Forschungsdesign ab. Uns interessieren hier vor allem Interventionsstudien, in denen der Effekt einer bestimmten Intervention (Maßnahme, Therapie, Lehrmethode usw.) gemessen werden soll. Eine gewisse Verbreitung hat der Vergleich der Mittelwerte einer Experimentalgruppe mit einer Vergleichs- oder Kontrollgruppe: $aES = M_{experimtal} - M_{kontrolle}$. Es ist offensichtlich, dass ein solcher Vergleich uns nur dann eine zuverlässige Auskunft über den Effekt gibt, wenn die Anfangswerte in beiden Gruppen gleich waren und auch sonst die Teilnehmer in beiden Gruppen keine Unterschiede aufwiesen, die den Effekt begünstigen oder erschweren können.

Wenn einem sozialwissenschaftlichen Experimenten eine reine Zufallsaufteilung auf Experimental- und Kontrollgruppe nicht möglich ist, weil sie nicht nur sehr aufwendig und teuer ist, sondern auch die externe oder ökologischen Validität der Ergebnisse in Frage stellt, ist der obige Mittelwertvergleich nicht brauchbar. Es bietet sich statt dessen ein Vorher-Nachher-Vergleich der Mittelwerte an:
$aES = M_{nachher} - M_{vorher}$ oder einfacher $M_2 - M_1$.

Die Einschätzung von absoluten Effektstärken bei sozialwissenschaftlichen Skalen.

Die Einschätzung der Bedeutsamkeit einer bestimmten absoluten Effektstärke hängt in erster Linie davon ab, wie viel wird über die Skala wissen, mit der wir den Effekt messen. Wenn die Skala noch "jung" ist und wir wenig Forschung und Alltagserfahrung dazu haben, stehen uns zur Orientierung nur formale Eigenschaften der Skala selbst zur Verfügung. Bei endlichen Skalen, wie sie in den Sozialwissenschaften (einschließlich Psychologie und Erziehungswissenschaft) üblich sind, sind das die absoluten Endpunkte und der absolute Mittelpunkt. Je mehr länger eine Skala im Einsatz ist und je mehr wir über die empirische Bedeutung einer Skala wissen, um so besser können wir die praktische Bedeutsamkeit von Skalenwerten und Wertedifferenzen einschätzen. "Empirische Bedeutung" meint hier sowohl die Bedingungen, die zum Erreichen bestimmter Skalenwerte oder für bestimmte Wertedifferenzen notwendig sind, also auch die diversen Effekte, die bestimmte Skalenwerte oder Wertedifferenzen haben. Beispiel: Die Temperaturskala ist uns Menschen seit langem gute vertraut.

Wir wissen z.B. wie viel Energie notwendig ist, die Temperatur eines Liters Wasser von 20 Grad Celsius Raumtemperatur auf 100 Grad Kochtemperatur anzuheben. Wir wissen auch, welche Konsequenzen ein Anstieg der Körpertemperatur auf 40 Grad hat und wie wir unsere Bekleidung ändern müssen, wenn die Außentemperatur um ca. 5 Grad steigt oder fällt.

Anders als in den Naturwissenschaft, wo wir es mit wenigen, seit langem bekannten und immer wieder verbesserten Messskalen (für Länge, Zeit, Energie, Masse) zu tun haben, gibt es in den Sozialwissenschaften ungeheuer viele Skalen und sind diese meisten sehr jung. Ständig kommen neue hinzu und verschwinden alte. In den Naturwissenschaften gibt es (von einigen Ausnahmen im Alltag abgesehen) für jede Messskala immer nur eine "Operationalisierung" (Messoperation) gibt. Variationen betreffen m.W. nur die Genauigkeit und Größenordnung der Messung (Micro-, Meso- und Makrobereich). In den Sozialwissenschaften spielt Genauigkeit dagegen eine untergeordnete Rolle. Hier finden sich dagegen oft verschiedene Operationalisierungen derselben Messdimension, die nicht – wie man erwarten müsste – perfekt miteinander korrelieren und oft noch nicht einmal dem Augenschein nach große Ähnlichkeit miteinander aufweisen. Beispiel: Moralische Urteilsfähigkeit (MU). "Fähigkeit" ist im Alltag und in der Wissenschaft eindeutig definiert als etwas, das sich bei bestimmten Aufgabenarten bewähren muss (MU-Fähigkeit ist die Fähigkeit, moralische Aufgaben zu bewältigen), MU-Fähigkeit von vielen Autoren als moralische *Einstellung* operationalisiert (z.B. als Präferenz für moralische Prinzipien, wie beim Defining-Issues Test, DIT). Einstellungen sind aber in jede Richtung simulierbar und geben allenfalls indirekte Hinweise auf eine Fähigkeit. Zum Beispiel besteht meist eine hohe empirische Korrelation zwischen der Fähigkeit in einem Fachgebiet und dem Interesse daran. Diese Korrelation findet man aber nur unter günstigen Bedingungen. Interessen und andere Einstellungen sind kein *verlässlicher* Indikator für Fähigkeiten. Die Messung von Fähigkeiten kann daher meist nicht durch die Messung entsprechender Einstellungen ersetzt werden, wenn die Messung verlässlich sein soll.

Gerade im Bereich der Moralpsychologie liegt mit dem Moralisches Urteil-Test (MUT) von Lind (2008) eine Messskala vor, die seit über 30 Jahren besteht und also vergleichsweise alt ist, und deren Bedeutung im zweifachen Sinne des Worten (siehe oben) schon relativ gut erforscht ist: a) Wir können gut abschätzen, wieviel Bildung notwendig ist, um eine bestimmte Punktzahl auf dieser Skala zu erreichen, und wie gut diese Bildung sein muss, um eine bestimmte absolute Effektstärke zu erreichen; b) wir können auch relativ gut abschätzen, wie hoch der Grad der Moralentwicklung sein muss, um das Risiko dysfunktionalen Verhaltens (wie Kriminalität, Drogensucht, Hilfeverweigerung, Lernprobleme, Entscheidungsunfähigkeit) auf ein Minimum zu bringen. Die Schätzungen sind zwar noch immer mit einer relativ großen Unsicherheit behaftet, aber diese Unsicherheit konnte durch die Forschung mit dem MUT gegenüber früher, als es diese Skala noch nicht gab, deutlich verringert

werden. Vermutlich geht diese Unsicherheit nicht nur auf die Qualität der Skala zurück, sondern auch auf die Komplexität des Verhaltens, das hier zur Debatte steht. Auch in den Naturwissenschaften gibt es viele Bereiche wie z.B. die Wettervorhersage, in denen trotz hochpräziser Messskalen und trotz sehr leistungsfähiger Computermodelle Unsicherheit besteht.

Zur konventionellen Einschätzung absoluter Effektstärke bei wenig erforschten Messskalen

Wenn die Forschung auf einem neuen Gebiet noch wenig fortgeschritten ist und wir nicht richtig einschätzen können, was ein großer oder ein geringer Effekt ist, kann man sich behelfen, indem man die absolute Effektgröße per *Konvention* bestimmt, zum Beispiel, indem man sie in Bezug zur theoretischen Skalenbreite setzt. (Ich nenne sie trotzdem "absolut" und nicht "relativ", weil sie nicht relativ zu einer *empirischen* Werteverteilung sind.) Da sozialwissenschaftliche Skalen oft sehr verschieden breit sind, empfiehlt es sich, in einem ersten Schritt die Punktwerte auf eine Standardskala von 1 bis 100 umrechnen und die gefundenen Differenzen oder Veränderungen hiermit in Beziehung setzen. Im zweiten Schritt ist festzulegen, wie groß eine Differenz bezogen auf dieser 100er-Skala sein sollte, damit sie als "bedeutend" oder "deutlich" gelten kann. Im dritten Schritt schließlich wäre eine theoretisch fundierte Abstufung der Bedeutung oder Größe eines Effekts vorzunehmen.

Wir wollen hier nur bis zum zweiten Schritt gehen und Anhaltspunkte für psychologisch-pädagogisch bedeutende Effekte zu gewinnen versuchen. Als Basis hierfür ziehen wir empirische Studien heran, die in der Wissenschaft ein hohes Ansehen genießen und deren Befunde einen gewissen Einfluss auf die Praxis genommen haben. Da wir in diesem Schritt nur eine grobe Trennlinie zwischen bedeutenden und unbedeutenden Unterschieden ermitteln wollen, genügen einige wenige Studien. Zudem gibt es leider nur wenige Studien, in denen alle notwendigen Informationen, nämlich die Mittelwertdifferenzen *und* die Skalenbreite mitgeteilt werden. Ausgewählt wurden die Schuluntersuchung von Fend und Kollegen (Fend et al. 1976), die Lehrerstudenten-Studie von Dann, Müller-Fohrbrodt und anderen (Müller-Fohrbrodt, 1973) und die Gruppeninteraktions-Forschung von Oser (1981). Aus jeder Studie haben wir einige Auswertungsergebnisse exemplarisch ausgewählt, und zwar solche, die von den Autoren als statistisch "hochsignifikant" bezeichnet wurden und denen sie auch in ihren Schlussfolgerungen große Bedeutung zugemessen haben.

Dreizehn solcher Befunde sind in die folgende Tabelle eingetragen und auf eine Bezugsskala von 1 bis 100 umgerechnet worden. Die Werte in der äußersten Spalte rechts geben an, wie groß die von den

Autoren berichtete Differenz oder Veränderung ausgedrückt in einer 100er–Skala wäre. Ein Wert von 1,75 bedeutet z.B., dass der statistisch "hochsignifikante" Effekt 1,75% oder 1/67 der gesamten Skalenbreite beträgt.

Wie wir aus der Tabelle entnehmen können, schwanken die Autoren sehr in ihrem Urteil, ab welcher Größe, gemessen an der absoluten Skalenbreite, ihnen ein Effekt als sehr bedeutsam gilt. Die Werte reichen von 1,75% bis 15% der Skalenbreite. Im Mittel ergibt sich ein Wert von 7,83 Prozentpunkten. Dieser Wert scheint typisch für die empirische Sozialforschung. Auf dieser Basis schlagen wir folgende verbale Beschreibungen für Wertedifferenzen auf Einstellungsskalen vor:

Effekt > 10% der Skalenbreite = “sehr bedeutend” oder “sehr deutlich”

Effekt > 5% der Skalenbreite = “bedeutend” oder “deutlich”.

Es sprechen also gute Gründe dafür, von *sehr bedeutenden* Differenzen oder Veränderungen nur dann zu sprechen, wenn sie *10 Prozent* der theoretisch möglichen Skalenwerte oder mehr betragen. Bei einer 5er–Skala zum Beispiel, wie sie in der sozialwissenschaftlichen Forschung häufig benutzt wird, wären 0,5 Punkte, also eine halbe Skaleneinheit eine sehr bedeutende Differenz. Bei einer Differenz von einem Viertel Punkt können wir bei dieser Skalenbreite noch immer von einem bedeutenden Unterschied sprechen. Die einschränkenden "guten Gründe" können vielfältiger Natur sein. Insbesondere wenn ein bestimmtes Gebiet theoretisch und empirisch gut durchdrungen ist, können Gründe für einen anderen (höheren oder niedrigeren) Kriteriumswert oder eine stärkere Differenzierung der Bewertung sprechen. In diesem Fall können wir auf rein konventionell begründete Kriterien verzichten.

Autor(en)	Interpretation der Effekte	Vergleich	Skalenbreite	Effekt absolut	Effekt in %
Fend et al. 1976	"Schüler mit hohem Leistungsstatus entsprechen den schulischen Disziplinforderungen eher als Schüler mit niedrigem Leistungsstatus. Die Unterschiede ... sind jeweils hochsignifikant." (S. 79)	mittel/ niedrig	16	0,28	1,75
		hoch/ mittel	16	0,37	2,31
	"Die Lern– und Leistungsmoral der Gymnasiasten ist signifikant (p<.01) niedriger als die der Hauptschüler." (S. 91 f.)	Leistungsmoral	16	0,65	4,06
		Lernmoral	16	0,57	3,56
	"Gesamtschüler befürworten eindeutig stärker eine Selektion nach Leistung in der Schule" (S. 214)		16	0,34	2,13

	"Lehrer an Gesamtschulen sind ... progressiver als die Lehrer an jeder Schulform des herkömmlichen Schulsystems." (S. 216)	GS/HS	16	2,04	12,75
		GS/Gy	16	2,36	14,75
	"[Eigen-Orientierungen] nehmen mit dem Herkunftsstatus *ab*." (S. 289)	US/MS[5]	16	0,54	3,38
	Gymnasiasten haben ein höheres Selbstbewusstsein als Hauptschüler (S. 400)		16	1,74	10,88
Müller-Fohrbrodt, 1973	"Während des Studiums werden ... Studenten ... progressiver." (S. 108)	Abi/StF-[6]männlich	90	7	7,78
	Studenten werden zwischen Abitur und 2. Studienhälfte "konformistischer" (S. 109)	Abi/StF männlich	30	4,5	15,00
Oser, 1981	"Die ... Stufen 1 und 2.5 lassen keine inhaltliche Integrierung zweier Treatments zu." (S. 401)	ohne/mit Regel	6	0,7	11,67
	"... hochsignifikanter Regelhaupteffekt... Der größte Unterschied liegt bei Stufe 3" (S. 402)	ohne/mit Regel	6	0,51	8,50
Gesamtskalenbreite / Gesamteffekte = relativer Effekt (in %):			276,00	21,60	**7,83**

Zur theoriegeleiteten Einschätzung von absoluten Effektstärken bei gut erforschten Skalen

In einer fortgeschrittenen Phase der Forschung auf einem bestimmten Gebiet fragen wir aber oft danach, ob eine bestimmte Therapie oder Lehrmethode einen größeren Effekt hat als bisher erfolgreich eingesetzte Therapien oder Lehrmethoden. Oder wir fragen danach, welche praktischen Auswirkungen bestimmte Differenzen auf einer Messskala haben können.

In diesem Fall sollte das Evaluationsdesign eine Vergleichsgruppe enthalten, bei der die bisherigen Methoden angewendet werden. Der Effekt der neuen Methode ergibt sich dann aus dem Vergleich zweier Vorher-Nachher-Differenzen:

$$aES = ([M_{exp,2} - M_{exp,1}] - [M_{kon,2} - M_{kon,1}])$$

Beispiel: In der Interventionsstudie von Glasstetter (2005) bei jugendlichen Straftätern haben sich folgende Mittelwerte (moralische Urteilsfähigkeit, C-Wert, MUT) in der Experimental- und der Kontrollgruppe ergeben:

[5] Abkürzungen: Unterschicht, Mittelschicht.

[6] Abiturienten versus fortgeschrittene Studierende..

	Vorher	Nachher	Effekt
Experimentalgruppe (Dilemmadiskussionen u.a.)	$M_{exp,1} = 19,1$	$M_{exp,2} = 18,3$	**-0,8**
Kontrollgruppe (traditionelle Erziehungs-maßnahmen)	$M_{kon,1} = 16,7$	$M_{exp,2} = 12,3$	**-4,4**
$aES = ([M_{exp,2} - M_{exp,1}] - [M_{kon,2} - M_{kon,1}]) = -0,8 - (-4,4) =$ **+3,6**			

Glasstetter (2005, p. 194 ff.)

Wie man sieht, ergibt sich ein scheinbar paradoxes Ergebnis: In jeder der beiden Gruppen nahm die moralische Urteilsfähigkeit über die Zeit ab, beide Maßnahmen hatten also einen negativen Effekt. Aber die Inventionsmaßnahme von Glasstetter (Experimentalgruppe) konnte offenbar den deutlich negativen Effekt des Aufenthaltes in der Institution (Kontrollgruppe) fast neutralisieren. Sie hatte also einen Effekt. Die aES von 3,6 bemerkenswert. So stark ist ungefähr auch der Effekt eines ganzen Schuljahres auf der Sekundarstufe (Lind, 2002, s. 159 ff.).

Die Verwendung absoluter Effektstärken setzt also voraus, dass wir mit dem Messinstrument bzw. der Messskala gut vertraut sind, was in einem neuen Forschungsgebiet nicht der Fall ist. In den Sozialwissenschaften liegen kaum Messskalen vor, mit denen wir so gut vertraut sind wie mit physikalischen Größen wie Temperatur, Länge oder Zeit. Die Differenz von 3,6 C-Punkten in unserem obigen Beispiel sagt – außerhalb eines kleinen Expertenkreises – kaum jemandem etwas, ob dies viel oder wenig, praktisch relevant oder theoretisch signifikant ist. Die Beurteilung dieser Differenz hängt zunächst von der Frage ab, wie stark dieser Wert überhaupt schwanken kann. Wenn die Skala unendlich lang ist, hilft diese Information wenig; aber in den Sozialwissenschaften haben wir es meistens mit endlichen Skalen zu tun. Der C-Wert kann zum Beispiel nur zwischen 0 und 100 liegen. Somit stellt eine Differenz von 3,6 immerhin einen Anstieg um 3,6 Prozent der Gesamtskala dar. Wenn die gesamte Länger der Skala nur 20 Punkte betragen würde, wäre diese Differenz noch eindrucksvoller.

Dann hängt die Beurteilung der Bedeutung dieser Differenz von der Frage ab, wie stark der C-Wert in verschiedenen Gruppen üblicherweise voneinander abweicht. Da er für Gruppen von Personen je nach Entwicklungsstand zwischen 10 und 40 liegt, bedeutet ein Effekt von durchschnittlich 3,6 C-Punkten,

dass er ungefähr 10 % der üblichen Spannbreite der Moralentwicklung bedeutet, was wir als ziemlich viel empfinden.

Zudem müssen wir die gefundene Differenz in Relation sehen zu der Zeit, innerhalb der diese Veränderung erreicht wurde, gemessen daran, wie viel Zeit sonst dafür benötigt wird. Ich habe einmal alle Studien zusammen getragen, in denen der C-Wert in Abhängigkeit von der Zeit bzw. von Bildungsprozessen untersucht wurde. Dabei habe ich festgestellt, dass der C-Wert bei Schülern an einer allgemeinbildenden Schule pro Jahr um ca. 3,5 C-Punkte zunimmt. Ein Effekt von 6,2 C-Punkten in einem pädagogischen Interventionsexperiment, das nur drei Monate dauerte, ist daher als groß zu bezeichnen (auch wenn er, wie im ersten Fall unseres hypothetischen Beispiels oben, statistisch nicht signifikant war).

Schließlich müssen wir eine festgestellte Differenz in Relation zu dem Verhalten sehen, das es auslösen oder nicht auslösen kann. Eine bestimmte Mittelwertdifferenz ist weder praktisch noch theoretisch signifikant, wenn sie zwar statistisch "hoch signifikant", aber für sonst irrelevant ist. Es wäre also noch zu zeigen, dass ein Unterschied von 6,2 C-Punkten bei Menschen zu unterschiedlichen Verhaltensweisen führt. Aufgrund vieler Studien kann eine solche Relevanz heute angenommen werden (Lind, 2002).

Schlussfolgerung

Statistische Signifikanztests sollten nur dafür eingesetzt werden, wofür sie gemacht wurden, nämlich zur Abschätzung der Genauigkeit eines Messvorgangs. Der Schluss von (statistischer) Signifikanz auf praktische Bedeutsamkeit unzulässig ist. Die Bezeichnung "Signifikanz" (was übersetzt ja Bedeutsamkeit heißt) ist irreführend und sollte geändert werden.

Aus konventionellen Gründen wird man den Signifikanzbegriff wohl noch eine zeitlang auch dann verwenden müssen, wenn er unangebracht ist, da selbst die Herausgeber renommierter Fachzeitschriften noch nicht die Problematik erkannt haben und Veröffentlichungen ablehnen, die statt der Signifikanz Effektstärkemaße berichten. Aber der Autor/die Autorin sollten immer darauf hinweisen, dass er/sie sich der Begrenztheit dieses Konzepts bewusst ist, und sollten auch immer Effektstärken berichten, wenn es das ist, was sie berichten wollen.

Solange nicht genau als Norm festgelegt ist, wie groß die Stichproben sein müssen und sein dürfen, sind die Ergebnisse solcher rein statistischen Verfahren zur Bestimmung von "Signifikanz" undurchsichtig (weil sich darin nicht nur Unterschiede in den Mittelwerten, sondern auch verschiedene Stichprobengrößen und Varianzen spiegeln) und offen für willkürliche Eingriffe des Experimentators. Und wer den Wissenschaftsbetrieb kennt, weiß, dass von dieser Möglichkeit, den Ausgang einer Studie zu bestimmen, auch reichlich Gebrauch gemacht wird, ohne dass die Wissenschaftsgemeinde darin einen Verstoß gegen ihr Ethos sieht.

Die Verwendung von Maßen der *relativen Effektstärke (rES) wie "r" und "d"* stellt einen deutlichen Fortschritt gegenüber den statistischen Signifikanztests dar, sind jedoch aus mehreren Gründen mehrdeutig und erlauben keine einfachen Schluss auf den 'wirklichen' Effekt einer Behandlung oder pädagogischen Maßnahme (siehe oben). Es sollte auch immer gesagt werden, ob r oder d berichtet wird. Leider fehlt dieser Hinweis oft, so dass der Leser aufs Raten angewiesen ist. Der Nachteil der rES besteht darin, dass sie zwar unabhängig von der Größe der Stichprobe sind, aber, wie der Name sagt, "relativ" sind zu der Streuung der Werte in den untersuchten Daten. Da diese Streuung (Varianz, Standardabweichung) sich von Studie zu Studie stark unterscheiden kann, ist ein direkter Vergleich der relativen Effektstärken oft nicht möglich.

Maß der *absoluten Effektstärke (aES)*, die aus einfachen Differenzberechnungen gewonnen werden, heben diesen Nachteil auf. Aber auch sie erlauben keine mechanische Interpretation. Eine gefundene Differenz zwischen Vortest- und Nachtestwerten, oder zwischen Experimental- und Kontrollgruppe erlaubt keine simple Schlussfolgerung auf die Wirksamkeit einer Behandlung oder einer pädagogischen Maßnahme.

Zur Interpretation der Bedeutsamkeit einer aES sollte man immer den aktuellen Stand der Forschung heranziehen und ggf. Meta-analysen durchführen. Koeffizienten für Effektstärke können die fachwissenschaftliche Beurteilung von Befunden nicht ersetzen. Die praktische Bedeutung eines Interventionseffekts hängt u.a. von den Kosten der Intervention, der Wichtigkeit des Zieles und der Relevanz des Effekts für bestimmte Verhaltensweisen ab. Beispiel: Wenn mit der Methode A derselbe Effekt mit weniger Aufwand erzielt werden kann als mit Methode B, dann ist Methode A 'effizienter'. Wenn mit einem Medikament nur wenigen Patienten das Leben gerettet werden kann, kann dieser zahlenmäßig geringe Effekt bedeutsamer sein als wenn mit einem anderen Medikament sehr vielen Menschen bei der Heilung ihres Schnupfens geholfen werden kann. Wenn bereits eine geringe Zunahme der angezielten Fähigkeit sich langfristig stark auswirkt, kann auch dieser geringe Effekt

hoch bedeutsam sein. Die theoretische Bedeutung eines Effekts hängt ergibt sich aus der bisherigen Forschungslage. In einem frühen Stadium eines neuen Forschungsfelds kann bereits ein geringer Effekt einen wichtigen Hinweis geben; wenn bereits große Effekte vorliegen, sollten neue Ansätze diese Effekte übertreffen können.

Damit sind aber die Möglichkeiten einer Effektanalyse bei weitem nicht ausgeschöpft. So kann beispielsweise gefragt werden:

- Wie groß sind die Effekte einer neuen psychologischen Therapie oder einer neuen pädagogischen Maßnahme im Vergleich zu "natürlichen" Veränderungen oder im Vergleich zu den Wirkungen bisheriger Therapien und Maßnahmen?
- Ab welcher Effektgröße ist damit zu rechnen, dass die Effekte nachhaltig sind, also über längere Zeit stabil sind oder gar größer werden? So kann der gewünschte Effekt einer neuen Unterrichtsmethode darin liegen, dass sie nicht mehr Wissen vermittelt, sondern den Lernenden in die Lage versetzt, sein Wissen selbst weiter zu vermehren.
- Welche Effektstärke muss mindestens erreicht werden, damit das (zukünftige) Verhalten von Menschen davon merklich beeinflusst wird? So kann gefragt werden, ab welchem Unterschied in Schulleistungstests Schüler später einmal größere Chancen haben, einen Beruf zu finden oder ein bestimmtes Einkommen zu erzielen. Diesen Zusammenhang nennt man "prognostische Validität".
- Welche anderen Bedingungen müssen erfüllt sein, damit ein bestimmter Effekt auftritt? Ist zum Beispiel eine bestimmte Unterrichtsmethode immer effektiv oder nur dann, wenn sie Lehrern angewandt wird, die in dieser Methode ausreichend ausgebildet sind?
- Schließlich ist für die Beurteilung der Bedeutung eines Ergebnisses wichtig zu wissen, in welchem Verhältnis (therapeutischer oder pädagogischer) Aufwand einerseits und der Nutzen für das Individuum und die Gesellschaft andererseits stehen. Dies nennt man die *Effizienz* der überprüften Maßnahme.

Für den Fortschritt der sozialwissenschaftlichen Forschung ist es also unabdingbar, dass man sich um inhaltlich sinnvolle Hypothesen bemüht und praktisch und theoretisch bestimmt, was z.B. *psychologisch bedeutsame* Unterschiede sind (Meehl, 1978). Erst dann kann sich die Frage stellen, wie eine Studie geplant werden muss, um mit ausreichender Präzision und Eindeutigkeit eine Hypothese zu bestätigen oder zu widerlegen oder – was noch besser wäre – von zwei alternativen Hypothesen die richtige zu wählen. Im Rahmen einer solchen Fragestellung würde sich auch schnell zeigen, welche statistischen Kennzahlen adäquat sind und für die wissenschaftliche Erkenntnisgewinnung hilfreich sind, und welche es nicht sind. Die mechanische Anwendung von statistischen Konzepten bringt uns

weder wissenschaftlich noch praktisch weiter (Gigerenzer, 1998; Meehl, 1958; Hoffrage, 1998; Sedlmeier, 1998; Lind, 2002; Haller & Krauss, 2002; Thompson, 2006; Bracey, 2006).

Hinweise auf Fehler und Ergänzungen sind willkommen: Georg.Lind@uni-konstanz.de

Literatur

Aaron, B., Kromrey, J.D., & Ferron, J.M. (1998). *Equating r-based and d-based effect size indices: problems with a commonly recommended formula*. Paper presented at the meeting of the Florida Educational Research Association, Orlando, FL, (ERIC Document No. ED 433 353).

American Psychological Association, APA (1994). *Publication guidlines*. Washington, D.C., 4th edition.

Bortz, J. (1994). *Statistik*. Berlin: Springer-Verlag.

Bracey, G. W. (2006). *Reading Educational Research: How to Avoid Getting Statistically Snookered*. Heinemann.

Bredenkamp, J. (1970). Über Maße der praktischen Signifikanz. *Zeitschrift für Psychologie, 177,* 310-318.

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The Earth Is Round (p < .05). *American Psychologist, 49* (12), 997-1003.

Cooper, H. & Hedges, L. V., Hg. (1994). *The Handbook of Research Synthesis*. New York: Russell Sage.

Fend, H., Knörzer, W., Nagl, W., Specht, W. & Väth-Szusdziara, R. (1976). *Sozialisationseffekte der Schule*. Weinheim: Beltz.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review, 103*, 592-96.

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences, 21*, 199-200.

Glass, G. V. & Stanley, J. C. (1970). Statistical methods in education and psychology. Englewood-Cliffs, NJ: Prentice Hall.

Glass, G.V., McGaw, B., & Smith, M.L. (1978). Meta-analysis in social research. London: Sage Publications.

Glasstetter, S. (2005). *Moralerziehung nach Lawrence Kohlberg - Die Auswirkungen der Just-Community in einem geschlossenen Heim für delinquente Jugendliche*. Diplomarbeit, FB Psychologie der Universität Landau.

Haller, H. & Krauss, S. (2002). Misinterpretations of Significance: A Problem Students Share with Their Teachers? Methods of Psychological Research Online 2002, Vol.7, No.1. Bezogen von http://www.uni-landau.de/~agmunde/mpr/issue16/art1/haller.pdf am 22.3.2004

Hays, W. (1963). *Statistics for Psychologists*. New York: Holt.

Hepach, R. (2007). *Interventionsstudien zur Förderung der moralischen Urteilsfähigkeit; eine Meta-Analyse von Studien aus den Jahren 1985 bis 2006* [Interventions studies for fostering moral judgment competence. Meta-analysis of studies in the years1985 to 2006]. Bachelor thesis, Department of Psychology, University of Konstanz.

Hoffrage, (1998). Statistik verstehen. BerliNews. Bezogen von http://www.berlinews.de/archiv/1580.shtml (22.3.2004)

Jacobs, B. (o.J.). Einige Berechnungsmöglichkeiten von Effektstärken. http://www.phil.uni-sb.de/~jakobs/seminar/vpl/bedeutung/effektstaerketool.htm (20.8.2007).

Journal of Experimental Education. (1993). Special Issue "*The role of statistical significance testing in contemporary analytic practice: Alternatives with comments from journal editors.*" Washington, DC: Heldref Publications. (Available from ERIC/AE).

Kendall, M. G. & Stuart, A. (1973). *The advance theory of statistics*. Vol. 2. London: Griffin.

Law, K. S. (1995). The use of Fisher´s Z in Schmidt-Hunter-type meta-analyses. *Journal of Educational and Behavioral Statistics, 20,* 287-306.

Lind, G. (2002). *Ist Moral lehrbar? Ergebnisse der modernen moralpsychologischen Forschung*. Berlin: Logos-Verlag.

Lind, G. (2004). Jenseits von PISA — Für eine neue Evaluationskultur, S. 1 - 7. In: Institut für Schulentwicklung PH Schwäbisch Gmünd, Hg., *Standards, Evaluation und neue Methoden. Reaktionen auf die PISA-Studie*. Baltmannsweiler: Schneider Verlag Hohengehren.

Lind, G. (2008). The meaning and measurement of moral judgment competence revisited – A dual-aspect model. In: D. Fasko & W. Willis, Hg., *Contemporary Philosophical and Psychological Perspectives on Moral Development and Education,* S. 185 - 220. Cresskill. NJ: Hampton Press.

Loftus, G.R. (1993). A picture is more worth than thousands of p-values: On the irrelevance of hypothesis testing in the computer age. *Behavior Research Methods, Instrumentation and Computers, 25*, 250-256.

Meehl, P. (1958). When to use your head instead of the formula? In: H. Feigl, M. Scriven & G. Maxwell, Hg., *Minnesota studies in the philosophy of science,* S. 498-506.

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft Psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

Müller-Fohrbrodt, G. (1973). *Wie sind Lehrer wirklich? Ideal, Vorteile, Fakten*. Stuttgart: Klett.

Oser, F. (1981). *Moralisches Urteil in Gruppen. Soziales Handeln*. Frankfurt: Suhrkamp.

Popper, K. (1968). The logic of scientific discovery. London: Hutchinson (Original 1934).

Rosenthal, R. & Rubin, D.B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Education Psychology, 74,* 166-169.

Rosenthal, R. & Rosnow, R.L. (1984). *Essentials of Behavioral Research*. New York: McGraw-Hill

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L.V. Hedges, Hg., *The handbook of research synthesis,* S. 231-244. New York: Russell Sage Foundation.

Schmidt, F.L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47*, 1173-1181.

Schmidt, F. L. & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529-540.

Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. Methods of *Psychological Research Online 1996,* Vol.1, No.4.

Sedlmeier, P. (1998). Was sind gute Gründe für Signifikanztests? *Methods of Psychological Research Online,* Vol. 3, No. 1; bezogen von http://www.uni-landau.de/~agmunde/mpr/issue4/art4/&e=747 (22.3.2004).

Shaver, J.P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education 61*, 293-316.

Statistical significance testing in contemporary practice (1993). *The Journal of Experimental Education, 61(4),* September 1993.

Thompson, B. (1993). The use of statistical significance tests in research: bootstrap and other alternatives. *Journal of Experimental Education* 6(4), 361-377.

Thompson (1994): http://ericae.net/pare/getvn.asp?v=4&n=5

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (2006). *Foundations of Behavioral Statistics. An Insight-Based Approach*. Guilford Publications.

Wuttke, J. (2007). Die Insignifikanz signifikanter Unterschiede. In: T. Jahnke & W. Meyerhöfer, Hg., *Pisa & Co. Kritik eines Programm*. 2., erweitere Auflage. S. 99-246. Hildesheim: Franzbecker.

Appendix: Indices for Relative Effect Size: Conversion Formulas

Last revision: Aug. 2012

Conventions for using symbols (if not otherwise stated):

N_i is the size of the *i*-th sample, whereby *I* may be a number between 1and *k*, the total number of samples. For example, if the first sample has 6 members then $N_1 = 6$.

x_i is a variable, which can represent a set of numbers, e.g., the subjects´ scores in a math test: x_i= {72, 64, 45, 34, 95, 93}; specifically $x_2 = 64$.

Σ is the summation symbol.

k is the number of groups which are compared.

Combining coefficients of correlation[7]	$M_r = \frac{\sum_{i=1}^{k} N_i * r_i}{\sum_{i=1}^{k} N_i}$	**Example:** In two studies, these correlations were found between moral development scores and level of education (the Ns are in parentheses): r = 0.45 (50) and 0.65 (230). The estimated mean of correlations is M_r = (0.45 * 50 + 0.65 * 230) / 280 = 0,625.
X² (Chi-square)[8]	$r_{xy} \approx C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$	**Example:** If χ^2 = 14.34 and N = 80 then r_{xy} = sqrt (14.34 / (14.34 + 80)) = sqrt (0.152) = **0.39**.
Effect size measure *D*[9]	$r_{xy} = \frac{d}{\sqrt{d^2 + 4}}$	with $d = \frac{M_1 - M_2}{s}$
***t*-statistic**[10]	$r_{pb} = \sqrt{\frac{t^2}{t^2 + df}}$	**Example**: If t = 3.5 and $n_1 + n_2 = N$ = 250 then df = 249 and r_{pb} = 0.12. **Note**: If you use *t* to compare repeated measurements as in follow-up studies, make sure that you use the *t*-statistics for *dependent* (or paired) groups.
Point-biserial correlation[11]	$r_{xy} = r_{pb} * 1.25$ This formula can be used only if the ration n_1 / n_2 is bigger than 0.2 and smaller than 0.8.	r_{pb} denotes the coefficient of point-biserial correlation and df the degrees of freedom: $df = n_1 + n_2$ - 1. n_1 and n_2 are the number of subjects in each of the two groups that are compared. **Example:** If r_{pb} = 0.12, n_1 = 125, and n_2 = 125, then r_{xy} = 0.12 * 1.25 = 0.15
***F*-statistic**[12]	$r_{xy} = \sqrt{\frac{df_j * F}{df_j * F + df_i}}$	with df_i being the degrees of freedom *within* groups, and df_j the degrees of freedom *between* groups (number comparisons - 1). **Example**: Three groups (k=3) are compared with n=50 subjects. F = 23.45, df_j = k-1 =1, and df_i = n - k = 47. Then r_{xy} = 0,70.

[7] Law (1995) showed that prior Z-transformation are not needed for averaging *r* because results are almost identical.

[8] Kendall & Stuart (1967, p. 557)

[9] Glass et al. (1978)

[10] Glass & Stanley (1970, p. 318)

[11] Magnusson (1966, p. 205)

[12] Rosenthal & Rosnow (1984, p. 249)

Variance-model	$r = \sqrt{\frac{S^2_{between}}{S^2_{between} + S^2_{error}}}$	with $S^2_{between}$ being the variance due to the treatment or effect-variable. The variance ration r^2 is normally called the *coefficient of determination*. Its square root is the coefficient of correlation, however of total correlation, not only of linear correlation.
Binomial effect size display, BESD[13]	$r = 2 * BESD - 1$ $BESD = .50 + \frac{r}{2}$	**Example:** If r = .30 then BESD = 0,65 If BESD is .75, then *r* is 0.50
Mann-Whiteney U[14]	$r_{pb} = 1 - 2\frac{U}{n_1 * n_2}$	whereby r_{pb} is the point-biserial correlation coefficient.

Weitere Umrechungsformeln finden sich in Cooper & Hedges, 1994.

Umrechungs-Tools für *d* (Effektstärke) finden sich im Internet, von Bernd Jacobs, Uni Saarbrücken: http://www.phil.uni-sb.de/~jakobs/seminar/vpl/bedeutung/effektstaerketool.htm

[13] Rosenthal & Rubin (1982)

[14] Wilson (1976)