EFFEKTGRÖSSEN IN VARIANZANALYTISCHEN DESIGNS (MIT FESTEN FAKTOREN)

Jan Seifert

User Interface Design GmbH, Ludwigsburg, Deutschland

Draft of 30. Mai 2008

Zusammenfassung

Effektgrößen werden seit Jahrzehnten von Methodikern aus den Sozialwissenschaften gefordert. Allerdings haben zahlreiche verschiedenen Vorschläge, wie die Größe eines Effektes optimal geschätzt werden kann, ergänzt durch mindestens genauso viele Unter-Varianten, sowie die häufig uneindeutige Namensgebung, einen klaren Einblick in die Thematik erschwert, und damit die Verbreitung von Effektgrößen behindert. Die vorliegende Arbeit stellt die geläufigsten Größen (d, $\eta^2, \phi^2, \omega^2, \epsilon^2$) im Überblick dar, zeigt Zusammenhänge auf, klärt über Widersprüche und Unklarheiten in der Literatur auf, und diskutiert die Stärken, Schwächen und Probleme beim Umgang mit Effektgrößen.

Einleitung

Es gibt keinen Grund Effektgrößen zu verschweigen, aber viele Gründe sie zu verwerten. Wie nützlich Effektgrößen sind, wissen wir aus dem Kontext der Regressionsanalysen. Der Regressionskoeffizient R ist ein Maß für die Stärke des Zusammenhanges zwischen den beteiligten Variablen. Niemand käme auf die Idee, die Ergebnisse einer Regression darzustel-

len und dabei R zu verschweigen, denn das R vermittelt eine wertvolle Information über die Art des Zusammenhanges: seine Größe. Umso überraschender ist es, dass dieselbe Größe bei Varianzanalysen ($\sqrt{\eta^2}$) so gut wie nie berichtet wird. Das η^2 hat dieselbe Bedeutung, trägt denselben Informationswert, und trotzdem wird diese Information konsequent ignoriert.

Eine Ursache für diese Geringschätzung mag die Effektgrößen-Literatur sein, die verwirrender kaum sein könnte. Wer sich aufrafft, seine knapp bemessene Zeit zu nutzen und sich über Effektgrößen zu informieren, wird bald enttäuscht und gibt spätestens dann frustriert auf, wenn andere Aufgaben sich auf seinem Schreibtisch zu stapeln beginnen. Denn die Darstellungen sind häufig unklar, manchmal sogar widersprüchlich. Allzu oft werden Formeln und Kennbuchstaben in der Literatur nicht eindeutig definiert werden. In vielen Arbeiten wird nicht unterschieden zwischen theoretischen Definitionen, empirischen Größen und Populationsschätzern (die den Populationseffekt aus empirischen Daten schätzen). Ein anderer Grund ist bestimmt, dass Autoren gelegentlich nicht darauf hinweisen, ob sie partielle Effektgrößen verwenden oder nicht. Derartige Verwechslungen sind von mehreren Autoren überliefert (Levine & Hullett, 2002; Cohen, 1973). Auf diese Weise wird zum Beispiel das totale Omega-Quadrat nach Hays (1973) leicht mit dem partiellen Omega-Quadrat (Keren & Lewis, 1979) verwechselt. Dass SPSS in älteren Versionen das *partielle* η_p^2 in den ANOVA-Tabellen lediglich mit "eta squared" überschreibt, trägt mit Sicherheit nicht zur Vereinfachung bei (Levine & Hullett, 2002). Man bemerkt den Irrtum nur dann, wenn man einen Blick ins Handbuch wirft.

Derartige Ungenauigkeiten erschweren die Lektüre enorm. Sie tragen darüber hinaus zu schweren Fehleinschätzungen der Daten bei (Pierce, Block, & Aguinis, 2004) und haben wesentlich dazu beigetragen, mich von der Notwendigkeit der hier vorliegenden Überblicksarbeit zu überzeugen. Der vorliegende Artikel verfolgt drei Ziele: Die gängigen Effektgrößen sollen (1) in einem vergleichenden Überblick dargestellt werden, um (2) anschließend ihre Vor– und Nachteile zu diskutieren, und schließlich (3) *jedem* genügend Einsicht in die Thematik zu verschaffen, so dass er in der Lage ist, selbständig vertiefende Literatur zu lesen und einzuordnen.

Bemerkungen

Um eine möglichst präzise, aber einleuchtende Darstellung zu ermöglichen, muss vorab folgendes festgestellt werden:

1. Die vorliegende Darstellung wird sich auf Varianzanalysen mit festen Faktoren beschränken, Zufallsfaktoren werden nicht berücksichtigt. Multivariate Ansätze werden ebenfalls nicht dargestellt. Die Erweiterung dieser Arbeit auf den multivariaten Fall und auf Zufallsfaktoren würde die Übersichtlichkeit zu sehr schmälern und dem gesteckten Ziel damit zuwiderlaufen. Statt alles auf einmal darzustellen, nehmen die Autoren lieber diese Einschränkung in Kauf und bieten stattdessen eine möglichst einfache und übersichtliche Darstellung, auf deren Grundlage sich die weiterführende Literatur besser verstehen lässt. Wer sich hierfür interessiert, kann sich bei Olejnik and Algina (2000) informieren.

2. Für Formeln, die sich nur auf bestimmte varianzanalytische Modelle beziehen (z.B. auf ANOVAs mit messwiederholten Faktoren) wird dies explizit angegeben. Ist kein derartiger Vermerk im Text vorhanden ist die Formel allgemeingültig.

3. Die vorliegende Darstellung wird sich streng an die Dreiteilung in (1) theoretische (in Großbuchstaben), (2) empirische Maße (Kleinbuchstaben) und (3) Populationsschätzer (Kleinbuchstaben mit einem $\widehat{Dach}$) halten. Verwechslungsgefahr besteht dabei hinsichtlich des großen griechischen Eta, das dem lateinischen H gleicht. Sie lassen sich dadurch unterscheiden, dass der griechische Buchstabe kursiv gesetzt ist (Eta: lateinisch H, griechisch H).

Die Effektgrößen

Man kann unterscheiden zwischen absoluten und relativen Effektgrößen. Absolute Effektgrößen sind v. a. Mittelwertsunterschiede. Sie geben an, wie sehr sich die Mitglieder zweier Gruppen im Durchschnitt voneinander unterscheiden. Relative Maße setzen diesen Mittelwertsunterschied ins Verhältnis zu den Unterschieden zwischen den Probanden; oder anders formuliert: die Effektvarianz wird relativiert anhand der Fehlervarianz oder der Gesamtvarianz. Beide Arten haben vor und Nachteile. Die relativen Effektgrößen sind abstrakter und daher (nicht nur für Erstsemester) manchmal nicht ganz einsichtig. Manchmal sind Mittelwertsunterschiede allerdings kaum verständlicher. Was bedeuten schon zwölf Millise-

kunden Unterschied in einem Reaktionszeitexperiment? Oft wird der Vorteil der relativen Maße hervorgehoben, dass Vergleiche zwischen Untersuchungen möglich werden. In der Praxis sind oftmals beiderlei Größen ratsam, um eine (oder mehrere) Studien erschöpfend zu interpretieren.

Bei den relativen Effektgrößen unterscheidet man weiterhin, woran der Effekt relativiert wird. Ist das die Fehlervarianz so spricht man vom *standardisierten Nonzentralitätsparametern*[1]. Wird der Effekt im Verhältnis zur Gesamtvarianz betrachtet, so bezeichnet man das als *Maß aufgeklärter Varianz.*

Das Signifikanzniveau p als Effektgröße?

Nicht selten fällt auf, dass der Fehler 1. Art (p) ähnlich einer Effektgröße verstanden wird. Wenn der Effekt die 5 %-Hürde unterschreitet, bekommt er ein Sternchen *; unter 1 % wird er besonders hervorgehoben mit zwei Sternchen **; und bei 0.1 % bekommt er sogar noch ein Sternchen mehr ***. Das wird dann als "hoch-signifikant" und damit als besonders bemerkenswert hervorgehoben.

Es ist natürlich richtig, dass ein gewisser Zusammenhang besteht zwischen dem Signifikanzniveau und der Größe eines Effekts. Dennoch ist es nicht ratsam Effekte anhand von p zu vergleichen, weder zwischen verschiedenen Studien, noch innerhalb ein und derselben Studie. Denn erstens (1) ist die Beziehung zwischen p und dem Effekt ist nicht linear, so dass $p = .01$ nicht auf einen doppelt so großen Effekt schließen lässt, als $p = .02$. Und zweitens (2) ist das beobachtete p abhängig von der Anzahl der Versuchspersonen. *Ein kleines p ist damit keinesfalls in Indikator für einen großen Effekt.* Das Zählen von Sternchen führt meist in die Irre. Die "echten" Effektgrößen bieten dagegen einen deutlich besseren Interpretationsrahmen.

Standardisierte Mittelwertsdifferenz

Die standardisierten Mittelwertsdifferenzen werden vor allem im Kontext von Metaanalysen diskutiert. Es gibt zahlreiche Alternativen, um eine Mittelwertsdifferenz zu standardisieren: das d nach Glass, das d nach Cohen und das g nach Hedges. Die absolute

[1] Nicht zu verwechseln mit *dem* Nonzentralitätsparameter. Das ist der spezifische Parameter der nonzentralen F-Verteilung. Hier im Text wird dieser mit λ abgekürzt.

Mittelwertsdifferenz wird jeweils ins Verhältnis gesetzt mit einer Standardabweichung des Fehlers. Der Unterschied zwischen den verschiedenen Varianten besteht in der Standardabweichung, die zur Standardisierung verwendet wird (Tatsuoka, 1993).

Glass' d

Definition.

In einem Design mit einer Kontroll– und einer (oder auch mehreren) Experimentalgruppe(n) verwendet das Glass's D zur Standardisierung einer Mittelwertsdifferenz die Standardabweichung innerhalb der Kontrollgruppe (σ_{kont}). Der Mittelwert der Kontrollgruppe wird von dem der jeweiligen Experimentalgruppe abgezogen.

$$D_{Glass} = \frac{\bar{\mu}_{exp} - \bar{\mu}_{kont}}{\sigma_{kont}} \tag{1}$$

Empirische Größe.

$$d_{Glass} = \frac{\bar{x}_1 - \bar{x}_2}{s_{kont}} \tag{2}$$

Das Glass' d ist besonders dann geeignet, wenn die Binnenvarianzen der verschiedenen Probandengruppen nicht gleich sind, beispielsweise bei dem Vergleich klinischer Gruppen mit einer gesunden Kontrollgruppe.

Sind die Standardabweichungen aller Gruppen gleich, dann sind das Glass' d und das Hedges' g zahlenmäßig identisch (aber nur zahlenmäßig, denn das Hedges' g zeichnet sich aus durch ein engeres Konfidenzinterval).

Populationsschätzer.

Ein approximativer, aber ausreichender Populationsschätzer wird von Hedges (1981) wie folgt angegeben.

$$\hat{d}_{Glass} \approx d_{Glass} \cdot \left(1 - \frac{3}{4n_{kont} - 5}\right) \tag{3}$$

Wobei n_{kont} für die Anzahl der Probanden in der Kontrollgruppe steht. Die Approximation ist mit einem maximalen Fehler von .007 bei $n_{kont} = 3$ für die meisten Anwendungen ausreichend.

Cohen's d

Definition.

Cohen (1988) beschreibt die standardisierte Mittelwertsdifferenz so:

$$D_{Cohen} = \frac{\bar{\mu}_1 - \bar{\mu}_2}{\sigma_{Fehler}} \tag{4}$$

Das σ_{Fehler} ist die Standardabweichung innerhalb der beiden verglichenen Gruppen.

Empirische Größe.

$$d_{Cohen} = \frac{\bar{x}_1 - \bar{x}_2}{s_{Fehler}} = \frac{\bar{x}_1 - \bar{x}_2}{s_{Fehler}} \tag{5}$$

Gleichung 5 erwartet gleich große Varianzen. Sind gleichgroße Varianzen nicht gegeben, gibt es zwei Möglichkeiten. Die eine Möglichkeit besteht in dem Rückgriff auf das Glass's D, das grundsätzlich die Varianz innerhalb der Kontrollgruppe verwendet. Oder die Varianzen aus beiden Gruppen werden über $(s^2_{in1} + s^2_{in2})/2$ zusammengefasst.

Populationsschätzer.

Besonders bei kleinen Stichproben überschätzt das d_{Cohen} den "wahren" Effekt und muss nach unten korrigiert werden (Hedges, 1981; Cohen, 1988):

$$\hat{d}_{Cohen} \approx d_{Cohen} \cdot \left(1 - \frac{3}{4(n_1 + n_2 - 2) - 1}\right) \overset{n_1 = n_2}{=} d\left(1 - \frac{3}{8n - 9}\right) \tag{6}$$

Hedges's G

Während Cohen (1988) lediglich das d auf lediglich zweistufige Situationen anwendet und für mehrstufige Designs das Phi (s.u.) bevorzugt, favorisiert Hedges die standardisierte Mittelwertsdifferenz auch, wenn mehr als zwei Gruppen in einer Statistik betrachtet werden. Wenn es keinen Grund gibt zur Annahme gibt, dass sich die Varianzen innerhalb der Gruppen unterscheiden, dann schlägt Hedges (1986) vor, die "gepoolte" Standardabweichung innerhalb der Gruppen zu verwenden. Auf diese Weise wird der Standardfehler der Effektgröße reduziert, d.h. die Schätzung des Effekts wird besser.

Definition. Siehe Hedges (1986):

$$G = \frac{\bar{\mu}_1 - \bar{\mu}_2}{\sigma_{Fehler}} \tag{7}$$

Empirische Größe.

$$g = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2_{Fehler}}} = t \cdot \sqrt{\frac{n_1 + n_2}{n_1 n_2}} = d_{Cohen} \sqrt{\frac{df_{Fehler}}{N_{obs}}} \overset{n_1 = n_2}{=} \frac{2t}{\sqrt{N}} \tag{8}$$

Populationsschätzer.

Besonders bei kleinen Stichproben überschätzt das g den "wahren" Effekt und muss nach unten korrigiert werden (Hedges, 1981; Hedges & Olkin, 1985):

$$\hat{d} \approx d \left(1 - \frac{3}{4df_{Fehler} - 1}\right) \tag{9}$$

Verallgemeinerung auf Kontraste

Eine standardisierte Mittelwertsdifferenz ist einfach und intuitiv interpretierbar. Ein Nachteil besteht darin, dass diese Größen nicht auf Omnibus-Tests anwendbar sind. Es werden immer nur zwei Mittelwerte miteinander verglichen und somit sind sie nur für t-Tests interessant. Die Mittelwertsdifferenzen lassen sich auf Kontraste übertragen und erlauben damit auch eine Effektschätzung, wenn mehr als zwei Gruppen an einem Vergleich beteiligt sind.

Ein Kontrast Ψ ist eine gewichtete Summe aus mehreren Mittelwerten:

$$\Psi = w_1 \cdot \mu_1 + w_2 \cdot \mu_2 + \ldots + w_n \cdot \mu_n \tag{10}$$

Wobei für die Gewichte $0 \leq w_i \leq 1$ und $\sum w_i = 0$ gilt. Eine Mittelwertsdifferenz ist die einfachste Form eines solchen Kontrasts.

$$\Psi_{diff} = 1 \cdot \mu_1 + (-1) \cdot \mu_2 \; [+0 \cdot \mu_3 + \ldots + 0 \cdot \mu_n] \tag{11}$$

Das Prinzip der standardisierten Mittelwertsdifferenzen lässt sich also auch einfach auf Kontraste übertragen, wenn wir die Formeln einfach so schreiben:

$$g = \frac{\Psi}{\hat{\sigma}} = \frac{w_1 \cdot \mu_1 + w_2 \cdot \mu_2 + \ldots + w_n \cdot \mu_n}{\hat{\sigma}} \tag{12}$$

Die obige Formel ist ein Hedges g. Indem man die Größe im Nenner verändert, kann man entsprechend auch das Glass' d bestimmen.

Nonzentralitätsparameter

Das Phi

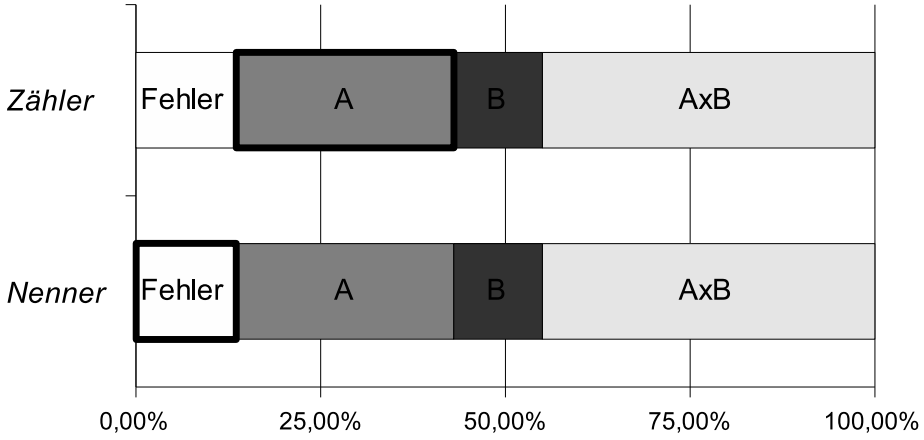

Abbildung 1. Die Abbildung veranschaulicht, welche Varianzanteile das Φ^2 miteinander vergleicht (jeweils dick umrandet).

Abbildung 1 veranschaulicht, welche Varianzanteile das Φ^2 miteinander vergleicht. Die dick umrandete Fläche oben ist der Varianzanteil des Effekts. Sie wird relativiert an der dick umrandeten Fläche unten (der Fehlervarianz). Das Φ^2 nimmt Werte größer 0 an und kann auch größer werden als 1.

Definition. Das Φ^2 ist definiert als Quotient der Varianzkomponenten des Effektes geteilt durch die des Fehlers (Cohen, 1988).

$$\Phi^2 = \frac{\theta^2_{Effekt}}{\theta^2_{Fehler}} \tag{13}$$

Empirische Größe.

$$\begin{aligned} \varphi^2 &= f^2 = \frac{\eta_p^2}{1-\eta_p^2} && (14) \\ &= \frac{\lambda}{df_{Fehler}} = \frac{QS_{Effekt}}{QS_{Fehler}} \\ &= \frac{df_{Effekt} \cdot F}{df_{Fehler}} && (15) \end{aligned}$$

Wie aus Formel 15 zusätzlich hervor geht lässt sich das Phi ebenfalls beschreiben als Nonzentralitätsparameter der nonzentralen F-Verteilung (λ), der an den Nenner-Freiheitsgraden relativiert wurde. Es handelt sich also um einen *standardisierten Nonzentralitätsparameter.*

Das φ^2 kann weiterhin als Verallgemeinerung einer standardisierten Mittelwertsdifferenz auf einen Omnibus-Test verstanden werden. Wenn nur zwei Stufen betrachtet werden, spiegelt sich diese enge Beziehung wider in diesem einfachen Zusammenhang zwischen φ^2 und dem Cohen's d (Cohen, 1988, S. 276ff):

$$\varphi^2 = (\frac{1}{2} d_{Cohen})^2 \tag{16}$$

Werden weitere Annahmen über die Struktur der Effekte gemacht, lassen sich auch in mehrstufigen Fällen eindeutige Beziehungen zwischen den standardisierten Mittelwertsdifferenzen und dem Phi herstellen (siehe Cohen, 1988, S. 277ff).

Populationsschätzer.

$$\begin{aligned} \hat{\varphi}^2 &= \hat{f}^2 = \frac{\hat{\omega}_p^2}{1 - \hat{\omega}_p^2} && (17) \\ &\approx \frac{df_{Effekt}(F-1)}{N_{obs}} \\ &\approx \frac{QS_{Effekt} - df_{Effekt} MQS_{Fehler}}{n_{obs} MQS_{Fehler}} && (18) \end{aligned}$$

Maße der aufgeklärten Varianz

Global aufgeklärte Varianz: $H^2_{gl}, \Omega^2_{gl}, R^2$

Das Eta/Omega (H^2/Ω^2) beschreibt das Verhältnis der Effektvarianz zur Gesamtvarianz: wie viel Varianz wird durch die experimentelle Variation hervorgerufen verglichen mit der gesamten Varianz. Abbildung 2 veranschaulicht dies genauer. Die dick umrandete Fläche oben ist der Varianzanteil des Effekts. Sie wird relativiert an der dick umrandeten Fläche unten (der gesamten Varianz). Daher variiert es zwischen Werten von 0 und 1.

Um Verwechslungen mit anderen Varianten dieser Effektgröße zu vermeiden, wird im weiteren Verlauf vom *globalen* Eta oder Omega die Rede sein.

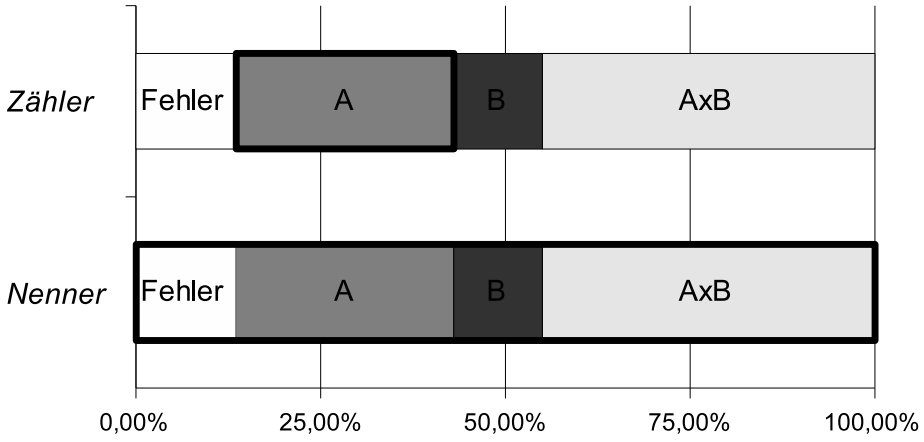

Abbildung 2. Die Abbildung veranschaulicht, welche Varianzanteile das globale Eta/Omega miteinander vergleicht (jeweils dick umrandet).

Definition.

$$H_{gl}^2 = \Omega_{gl}^2 = \rho^2 = \frac{\theta_{Effekt}}{\theta_{Total}} \tag{19}$$

Empirische Größe. Nach Richardson (1996) geht das (globale) Eta ($\eta_g^2 l$) unter dem Namen "correlation ratio" zurück auf Pearson (1905). Das empirische H^2 ist identisch mit dem unkorrigierten R^2 der multiplen Regression. Bei einer einfaktoriellen ANOVA mit zwei Stufen ist das H identisch mit der quadrierten punkt-biserialen Korrelation (r^2). Die hier dargestellten Zusammenhänge stehen bei Cohen (1973) und Haase (1983):

$$\eta_{gl}^2 = \omega_{gl}^2 = R^2 \tag{20}$$

$$= \frac{QS_{Effekt}}{QS_{Total}} = \frac{df_{Effekt}F}{\sum(df_{Effekt} \cdot F) + df_{Fehler}} \tag{21}$$

Wie man aus Abbildung 3 und Formel 19 schließen kann, sind die einzelnen η^2 für die verschiedenen Effekte voneinander abhängig. Je nach experimentellem Design wird sich die Gesamt-Quadratsumme QS_{total} verändern. Auch wenn die Effektquadratsumme immer dieselbe bleibt, wird sich das η^2 abhängig vom verwendeten varianzanalytischen Design verändern. Es kann also unter Umständen schwierig sein, verschiedene Experimente zu verglei-

chen. Eine Ausnahme hierfür sind einfaktorielle Varianzanalysen, da hier η^2 und partielles η^2 identisch sind. Die partielle Variante des Eta wird im nächsten Abschnitt beschrieben.

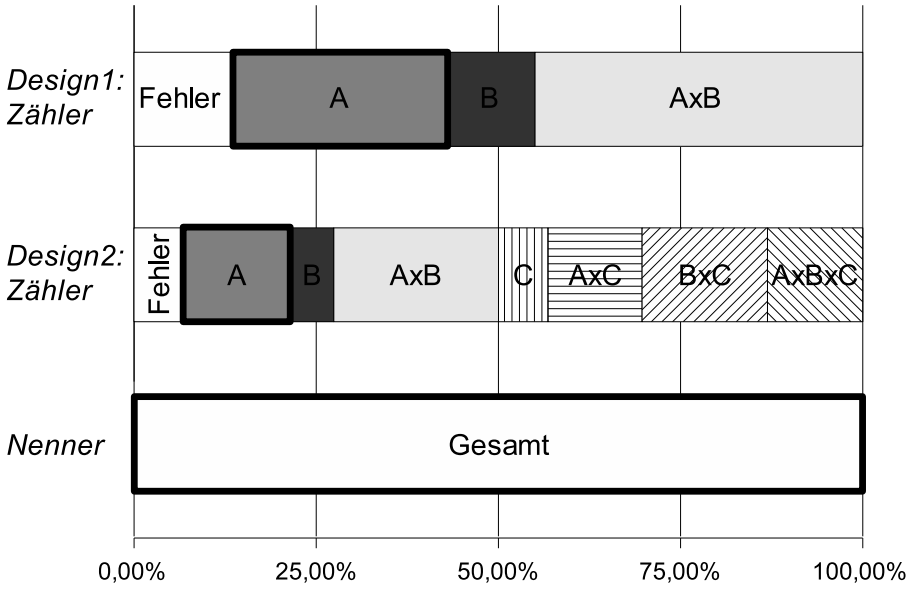

Abbildung 3. Diese Abbildung verdeutlicht, wie das Design der ANOVA die Effektgrößen beeinflussen kann. Der oberste Balken der Abbildung zeigt der Haupteffekte A und B, sowie der Wechselwirkung A×B deutlich größer als in der Mitte. Wenn ein Design um einen Faktor C erweitert wird, so werden manche Effektgrößen zwangsläufig kleiner.

Populationsschätzer Epsilon.

Da die empirische Größe die Eigenschaft besitzt, den "wahren Wert" systematisch zu überschätzen, wurde von Kelley (1935) das $\hat{\epsilon}^2$ (epsilon) vorgeschlagen[2]. Das $\hat{\epsilon}^2$ ist identisch mit dem schrumpfungskorrigierten R^2 nach Wherry (1931) (Maxwell, Camp, & Arvey, 1981). Bezüglich der Formeln, siehe Kelley (1935) und Cohen (1965).

$$\hat{\epsilon}^2_{gl} = \frac{QS_{Effekt} - df_{Effekt} MQS_{Fehler}}{QS_{Total}} \tag{22}$$

Im einfaktoriellen Fall lässt sich einfacher auch folgende Formel verwenden:

[2]Das $\hat{\epsilon}^2$ darf nicht verwechselt werden mit den Effektgrößen ε, die bei Bortz (1993) dargestellt sind. Das "Bortz'sche" ε ist auf S. 115 ein Cohen's d und wird später zum φ^2 verallgemeinert (S. 235).

$$\hat{\epsilon}^2_{gl} = \frac{df_{Effekt}(F-1)}{df_{Effekt}F + df_{Fehler}} \tag{23}$$

Populationsschätzer Omega.

Eine weitere Korrekturformel wurde von Hays (1973) vorgeschlagen. Diese Größe wird allgemein als Omega bezeichnet (hier $\hat{\omega}^2_{gl}$).

$$\hat{\omega}^2_{gl} = \hat{\eta}^2_{gl} \quad = \quad \frac{QS_{Effekt} - df_{Effekt}MQS_{Fehler}}{MQS_{Fehler} + QS_{Total}} \tag{24}$$

Bei einfaktoriellen ANOVAs vereinfacht sich das zu

$$\hat{\omega}^2_{gl} = \hat{\eta}^2_{gl} = \frac{df_{Effekt}(F-1)}{df_{Effekt}(F-1) + N_{obs}} \tag{25}$$

Für den Fall eines t-Tests gelten auch folgende Formeln (Hays, 1973):

$$\hat{\omega}^2_{gl} = \hat{\eta}^2_{gl} = \frac{t^2 - 1}{t^2 + n_1 + n_2 - 1} \tag{26}$$

Messwiederholung.

Bei messwiederholten Varianzanalysen sind die Zusammenhänge um einiges komplexer (Vaughan & Corballis, 1969; Dodd & Schultz, 1973). Hier ist zu unterscheiden, ob Interaktionen des Probandenfaktors mit den inhaltlichen Faktoren vorhanden sind. Gibt es derartige Interaktionen nicht gibt (sog. additives Modell), dann gilt entsprechend zum nicht-messwiederholten Fall (Olejnik & Algina, 2000):

$$\hat{\omega}^2_{gl} = \hat{\eta}^2_{gl} \quad = \quad \frac{QS_{Effekt} - df_{Effekt}MQS_{Fehler}}{MQS_{vpn} + QS_{total}} \tag{27}$$

Bei Effekten messwiederholter Faktoren (und Interaktionen mit messwiederholten Faktoren) entspricht die mittlere Quadratsumme des Fehlers (MQS_{Fehler}) der Interaktion des/der Faktoren mit dem Probandenfaktor ($MQS_{Effekt\times vpn}$).

Wenn Additivität nicht gegeben ist, dann muss der Nenner dieses Bruchs erweitert werden, um eine oder mehrere mittlere Quadratsummen (siehe Dodd & Schultz, 1973). Allerdings ergibt sich in diesem Fall, dass der wahre Effekt leicht unterschätzt wird. Susskind

and Howland (1980) setzen sich genauer mit dem Problem der Additivität auseinander. Gemäß ihrer Einschätzung kann man in vielen Fällen einfach von Additivität ausgehen, der damit verbundene Fehler sei vernachlässigbar. Je komplexer das Design (d.h. je größer die Anzahl Stufen der beteiligten Faktoren) und je größer der Effekt, umso kleiner wird der Schätzfehler, wenn man fälschlicherweise von Additivität ausgeht. Dennoch schließen die Autoren nicht aus, dass es unter verschiedenen (nicht näher spezifizierten) Umständen zu einem beträchtlichen Problem werden kann, wenn Additivität nicht gewährleistet ist.

Partiell aufgeklärte Varianz: H_p^2, Ω_p^2

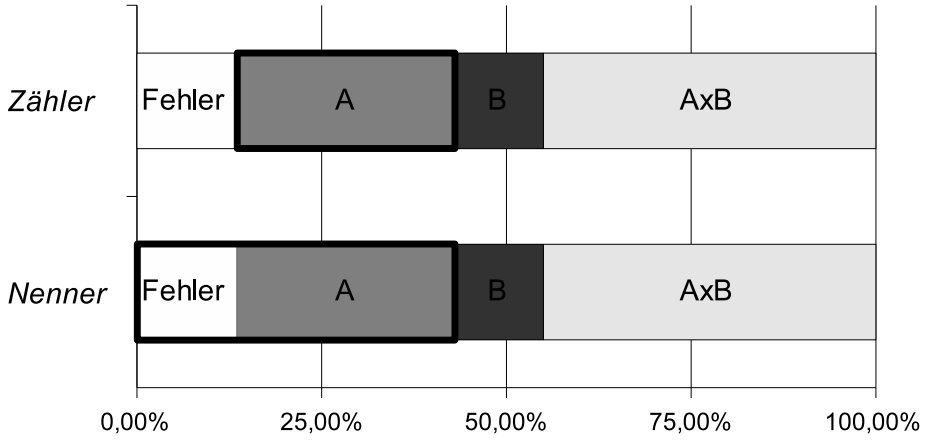

Abbildung 4. Die Abbildung veranschaulicht, welche Varianzanteile das Ω^2 miteinander vergleicht (jeweils dick umrandet).

Auch dieses "Ω^2" (Omega) lässt sich umschreiben als (prozentualer) Anteil der *aufgeklärten Varianz*. Es beschreibt die aufgeklärte Varianz eines Effekts, nachdem alle übrigen Effekte heraus partialisiert wurden. Abbildung 4 veranschaulicht dies genauer. Die dick umrandete Fläche oben ist der Varianzanteil des Effekts. Dieser wird relativiert an der dick umrandeten Fläche unten (der Summe aus Effekt– und Fehlervarianz). Als Prozentanteil variiert es zwischen 0 und 1.

Es wurde eingeführt mit der Absicht, ein Effektmaß zu schaffen, das bessere Vergleiche zwischen verschiedenen Studien zu ermöglichen. Das globale Omega verändert sich mit der Gesamtvarianz in einer Varianzanalyse. Die Gesamtvarianz wiederum verändert sich mit

der Anzahl der Faktoren. Dieselbe experimentelle Manipulation kann also zu einem anderen Omega führen, allein aufgrund einer Veränderung im varianzanalytischen Design und *nicht* aufgrund eines veränderten Effektes.

Definition. (Cohen, 1973)

$$\Omega_p^2 = H_p^2 = \frac{\theta_{innerhalb}}{\theta_{innerhalb} + \theta_{zwischen}} \tag{28}$$

Empirische Größe.

Die Formeln zur partiell aufgeklärten Varianz findet man bei Cohen (1973) und Haase (1983). In der Literatur dominiert für diesen Kennwert der Bezeichner η^2.

$$\eta_p^2 = \omega_p^2 = \frac{\varphi^2}{1+\varphi^2} = \frac{QS_{Effekt}}{QS_{Effekt} + QS_{Total}} = \frac{df_{Effekt}F_{Effekt}}{df_{Effekt}F_{Effekt} + df_{Fehler}} \tag{29}$$

Populationsschätzer Epsilon.

Das partielle Epsilon beschreiben Olejnik and Algina (2000):

$$\hat{\epsilon}_p^2 = \frac{QS_{Effekt} - df_{Zaehler}MQS_{Fehler}}{QS_{Effekt} + QS_{Fehler}} \tag{30}$$

Populationsschätzer Omega.

Der Populationsschätzer $\hat{\omega}_p^2$ wird bei Keren and Lewis (1979) oder Olejnik and Algina (2000) beschrieben:

$$\begin{aligned} \hat{\omega}_p^2 &= \hat{\eta}_p^2 = \frac{\hat{\varphi}^2}{1+\hat{\varphi}^2} \\ &= \frac{QS_{Effekt} - df_{Zaehler}MQS_{Fehler}}{QS_{Effekt} + (N_{obs} - df_{Zaehler})MQS_{Fehler}} \qquad (31) \\ &= \frac{df_{Zaehler}(F-1)}{df_{Zaehler}(F-1) + N_{obs}} \qquad (32) \end{aligned}$$

Hinsichtlich messwiederholter Designs sind zum partiellen Omega zwei Dinge festzuhalten. Erstens, besteht das Problem der Additivität in messwiederholten Analysen (siehe oben) auch für das partielle Omega. Wenn Additivität gegeben ist, kann das partielle Omega wie folgt berechnet werden:

$$\begin{aligned} \hat{\omega}_p^2 &= \hat{\eta}_p^2 = \frac{\hat{\varphi}^2}{1+\hat{\varphi}^2} \\ &= \frac{QS_{Effekt} - df_{Zaehler} MQS_{Fehler}}{QS_{Effekt} + QS_{Fehler} + N_{vpn} MQS_{Fehler}} \quad (33) \\ &= \frac{df_{Zaehler}(F-1)}{df_{Zaehler}(F-1) + N_{obs}} \quad (34) \end{aligned}$$

Und zum zweiten sollte darauf hingewiesen werden, dass eine partielle Effektgröße bei messwiederholten Faktoren immer eine ipsativierte Größe darstellt. Das bedeutet, dass die Varianz zwischen Versuchspersonen heraus partialisiert ist. Dies ist beim globalen Omega/Eta nicht der Fall, dieses muss durch eliminieren der entsprechenden Varianzkomponente gesondert berechnet werden (Gaebelein & Soderquist, 1978).

Das generalisierte Omega

Es gibt zahlreiche Möglichkeiten, um einer Studie die Fehlervarianz zu reduzieren. Zum Beispiel lassen sich verschiedene experimentelle Bedingungen an denselben Probanden realisieren (Messwiederholung). Man erhöht damit die Wahrscheinlichkeit, dass ein Effekt die Signifikanzschwelle überschreitet. Allerdings bedeutet das nicht, dass der Effekt dadurch größer wird. Und doch würden einige Effektgrößen dies suggerieren (so beispielsweise das Phi und das partielle Omega).

Ein ähnliches Problem ergibt sich, wenn eine nicht-randomisierte, beobachtete Variable in eine Analyse aufgenommen wird (z.B. das Geschlecht, eine Persönlichkeitseigenschaft). In nahezu jeder Stichprobe sind Probanden beiderlei Geschlechts vertreten. Fügt man also beispielsweise einen Faktor "Geschlecht" in die Varianzanalyse ein, werden die Quadratsummen innerhalb der Zellen (QS_{Fehler}) reduziert.

Um solchen Problemen aus dem Weg zu gehen, schlagen Olejnik and Algina (2003) eine weitere Variation des Ω^2 vor, das generalisierte Ω^2_{gen}

Definition.

$$\Omega^2_{gen} = H^2_{gen} = \frac{\theta^2_{Effekt}}{\Upsilon \cdot \theta^2_{Effekt} + \sum \theta^2_{beobachtet}} \quad (35)$$

Die Varianzkomponenten $\theta^2_{beobachtet}$ umfassen alle Effekte, an denen einer oder mehrere beobachtete Faktoren beteiligt sind. Wenn an dem untersuchten Effekt eine beobachtete Va-

riable beteiligt ist (z.B. Geschlecht, Geschlecht × Manipulierter Faktor), dann ist $\Upsilon = 0$. In diesem Fall ist die Varianzkomponente des Effekts (θ^2_{Effekt}) auch in $\theta^2_{beobachtet}$ enthalten und darf daher kein zweites Mal in den Nenner aufgenommen werden. Wenn nur experimentell manipulierte Faktoren an dem untersuchten Effekt beteiligt sind, ist $\Upsilon = 1$.

Empirische Größe.

Analog zur seiner Definition wird das generalisierte Stichproben-Omega wie folgt berechnet:

$$\eta^2_{gen}(=\omega^2_{gen}) = \frac{SS_{Effekt}}{\Upsilon \cdot SS_{Effekt} + SS_{Fehler} + \sum SS_{beobachtet}} \tag{36}$$

Populationsschätzer.

Das Formelwerk zu den Populationsschätzern hier darzustellen würde zu viel Raum einnehmen. Bei Olejnik and Algina (2003) sind Formeln für das $\hat{\omega}^2$ zu zahlreichen unterschiedlichen Versuchspläne dargestellt.

Die "drei Omega's" im Vergleich.

Theoretisch reichen die Wertebereiche der verschiedenen Maße aufgeklärter Varianz alle von Null bis Eins. Aufgrund ihrer Funktion als Schätzer können $\hat{\omega}^2$ und $\hat{\epsilon}^2$ im praktischen Fall auch negative Werte annehmen. Faktisch entspricht dies einem Effekt von Null, und nach Hays (1973) sollte auch Null berichtet werden. Damit würde allerdings würden die Effektschätzungen jedoch ins Positive verzerrt und Meta-Analysten sähen sich vor dem Problem einer systematischen Vergrößerung der Effekte, weshalb Vaughan and Corballis (1969) dafür plädieren, auch negative Werte zu berichten.

Davon abgesehen kann es im konkreten Fall zu beträchtlichen Größenunterschieden kommen, die umso beträchtlicher sind, je kleiner die Stichprobe ist. Grundsätzlich gilt: $\Omega^2_{gl} \leq \Omega^2_{gen} \leq \Omega^2_p$. Nur bei der einfaktoriellen ANOVA sind das generalisierte, das partielle und das globale Omega identisch.

Bei allen dreien Varianten des Omega wird in der Literatur niemals das Kürzel ω^2 verwendet, wenn von dem Stichprobenkennwert die Rede ist; statt dessen spricht man dann vom η^2. Als "ω^2" wird immer der entsprechende Populationsschätzer bezeichnet. Das "ω^2" ist also immer ein $\hat{\omega}^2$ und das η^2 ist kein "$\hat{\eta}^2$" (Ausnahmen wie Kotrlik & Williams, 2003 bestätigen leider die Regel). Wenn in einer Arbeit ein Maß aufgeklärter Varianz berichtet

wird, ohne näher zu spezifizieren, ob nun ein partielles gemeint ist, oder nicht, dann lässt sich das bestimmen, indem man alle berichteten Maße addiert. Die Summe aller globalen Omegas ist immer 1!

Beide Populationsschätzer $\hat{\omega}^2$ und $\hat{\epsilon}^2$ sind beide nur approximative Größen (Glass & Hakstian, 1969; Winkler & Hays, 1975). In einer Monte-Carlo-Studie konnten Carroll and Nordholm (1975) nur unwesentliche Abweichungen vom wahren Effekt nachweisen, die Approximation ist für praktische Zwecke ausreichend präzise. Der Unterschied zwischen den Populationsschätzern Omega und Epsilon ist zahlenmäßig nur unwesentlich (Maxwell et al., 1981; Carroll & Nordholm, 1975). Allerdings ist das Omega das mathematisch korrektere Maß (Carroll & Nordholm, 1975), und dessen mathematischen Zusammenhänge sind deutlich besser verstanden, wie die Ausführungen zum Addititvitätsproblem nahe legen[3]. Darum wäre dem Omega der Vorzug zu geben.

Varianzverhältnisse bei Kontrasten

Alle vier dargestellten Maße lassen sich auch verwenden, um die Größe von Kontrasteffekten anzugeben. Allerdings reduzieren sich die Varianten der aufgeklärten Varianz zu einem einzelnen Maß. Es ergeben sich folgenden Formeln (Olejnik & Algina, 2000):

$$\eta^2 = \frac{SS_{Kontrast}}{SS_{Total}} \tag{37}$$

$$\hat{\epsilon}^2 = \frac{SS_{Kontrast} - MS_{Fehler}}{SS_{Total}} \tag{38}$$

$$\hat{\omega}^2 = \frac{SS_{Kontrast} - MS_{Fehler}}{SS_{Total} + MS_{Fehler}} \tag{39}$$

$$\varphi^2 = \frac{SS_{Kontrast}}{SS_{Fehler}} \tag{40}$$

$$\hat{\varphi}^2 = \frac{SS_{Kontrast} - MS_{Fehler}}{SS_{Fehler}} \tag{41}$$

[3]Das Additivitätsproblem bei messwiederholten ANOVA's wird in der Literatur zum $\hat{\epsilon}^2$ nicht thematisiert.

Vergleichende Diskussion

In Sachen Effektgröße hat jeder Autor seine eigene Meinung. Es gibt keine eindeutige Empfehlung darüber, welche Größe die Beste ist. Man kann einige wünschenswerte Eigenschaften für Effektgrößen festlegen. Leider gibt es keine einzige Größe, die sie alle erfüllen könnte. Folgende Aspekte werden im Folgenden diskutiert:

1. Unabhängigkeit vom varianzanalytischen Design, um die *Vergleichbarkeit* zwischen verschiedenen Studien zu ermöglich.

2. Welche Vor– und Nachteile sind an die Verwendung von *Populationsschätzern* geknüpft?

3. Welche Vor– und Nachteile bietet die Verwendung partialisierter Effektgrößen?

4. Wie präzise ist die Schätzung? Die Effektgrößen sollten ein kleines *Konfidenzintervall*, bzw. geringen Standardfehler aufweisen.

5. *Einfache Berechenbarkeit*/ Verfügbarkeit. (a) Aus Gründen der Praktikabilität sollte ein Effektmaß ohne großen Aufwand zu bestimmen sein. (b) Berechenbarkeit post hoc. Es sollte ein Populationsschätzer zur Verfügung stehen, der sich auch post hoc aus vorliegenden Publikationen heraus sehr einfach ermitteln lässt.

3. Die Verwendung von Effektgrößen bei Omnibus-Tests gegenüber Einzelvergleichen.

4. Die statistischen Voraussetzungen an die Effektgrößen.

5. Einfache Interpretierbarkeit: (a) *Ein definierter Wertebereich.* Manche Autoren vertreten die Ansicht, dass dadurch die Interpretation anschaulicher wird. (b) Wie nützlich sind *Klassifikationssysteme* in diesem Zusammenhang?

Vergleichbarkeit

Die Vergleichbarkeit von Effekten erscheint mir die bedeutendste Anforderung, die an eine Effektgröße gestellt werden muss. Erst aus dem Vergleich heraus entwickelt sich eine Informationsgrundlage, die eine Interpretation des Effekts ermöglicht.

Die Maßnahme, die alle Effektgrößenschätzer einsetzen, um die Vergleichbarkeit zu verbessern, ist die Standardisierung. Der Effekt wird an der Fehler– oder der Gesamtvarianz standardisiert. Damit haben sie den Vorteil, dass sie von der Skalierung der abhängigen Variable nicht beeinflusst werden. Nur der "Reaktionszeitforscher" mag diesen Vorteil mög-

licherweise nicht besonders hoch schätzen. Denn Reaktionszeiten weisen eine feste Skalierung auf und sind somit jederzeit vergleichbar. Nicht unbedingt, muss man hier einwenden. MacDonald, Joordens, and Seergobin (1999) berichteten Reaktionszeiteffekte, die größer seien "als eine Brotbüchse". Die Reaktionszeitdifferenzen seien in ihrer Studie mehr als viermal so groß, als sonst in der Literatur üblich. Ein standardisiertes Effektmaß (hier ein d_{Cohen}) aber zeichnet ein anderes Bild (siehe Tabelle 1). In Experiment 1A sind die Effekte nahezu identisch, in Experiment 2A hingegen gelingt es den Autoren durch eine sorgfältigere Versuchsanordnung den Effekt zu verdoppeln. Da die größere versuchsplanerische Sorgfalt nur die Brotbüchsen-Bedingung betraf, ist es allerdings nicht weiter verwunderlich, dass sich auch nur dieser Effekt vergrößert hat, denn die Reliabilität der Messung beeinflusst die Effektgrößen (siehe unten).

Tabelle 1: Absolute und standardisierte Effekte nach MacDonald, Joordens und Seergobin (1999).

	Experiment 1A		Experiment 2A	
	$rt_1 - rt_2$	d_{Cohen}	$rt_1 - rt_2$	d_{Cohen}
Standard-Effekt	35	0.46	19	0.33
Brotbüchsen-Effekt	103	0.43	91	0.60
Verhältnis	2.94	0.93	4.79	1.81

Vergleichbarkeit zwischen verschiedenen Designs bzw. zwischen Experimenten ist nur beim generalisierten Omega und den standardisierten Mittelwertsdifferenzen gewährleistet (Olejnik & Algina, 2000, 2003). Die partiellen Varianzverhältnisse (inklusive dem Phi) können dies bereits in vielen Fällen gewährleisten, nur die globalen Varianzaufklärer sind hier besonders problematisch.

Unabhängig von der Diskussion darüber, welche Effektgröße nun die bestmögliche Vergleichbarkeit gewährleistet, wird die Vergleichbarkeit durch einige Aspekte grundsätzlich eingeschränkt, und diese Einschränkungen lassen sich durch keine mathematische Formel ausgleichen. Prinzipiell ist jede Maßnahme zur Maximierung der Power des statistischen Tests ein Problem bezüglich der Vergleichbarkeit von Effektgrößen (McClelland, 2000, 1997):

- Die *Reliabilität der Messung* bestimmt der Obergrenze, die eine Effektgröße errei-

chen kann. Je reliabler, desto besser die Schätzung. Unterschiedliche Reliabilitäten in verschiedenen Studien können daher die Vergleichbarkeit der Effektgrößen einschränken. Wenn die Reliabilität bekannt ist, dann kann die Effektgröße entsprechend korrigiert werden. Bei Baugh (2002) wird erläutert, wie diese Korrektur für das Cohen's d durchgeführt werden kann.

- *Stichprobengröße*. Je größer die Stichprobe, desto besser die Schätzung der Effektgröße.
- *Anzahl der unabhängigen Variablen* und deren Stufen. Je größer die Anzahl der Probanden pro Zelle (im Design), desto besser ist die Schätzung.
- *Homogenität der Stichprobe*. Die Schätzung eines Effekts verbessert sich mit zunehmender Homogenität.
- Stärke der experimentellen Manipulation.
- Die Fragestellung, die der Versuchsanordnung zugrunde liegt. Selektion von Extremgruppen bei vorgefundenen Merkmalen erhöht einen (linearen) Effekt.

Somit ist es schwierig Studien mit und ohne Extremgruppenselektion zu vergleichen.

Derartige Unterschiede in der Anlage von Studien müssen bei der Interpretation von Effektgrößen berücksichtigt werden. Andernfalls kann es leicht zu Fehlschlüssen kommen (Sechrechst & Yeaton, 1982; O'Grady, 1982).

Stichprobenmaß oder Populationsschätzer?

Prinzipiell sind die Größen aus den empirisch Daten zu optimistisch, sie *überschätzen* den wahren Effekt. Der Grund dafür liegt in dem mathematischen Prinzip der Varianzmaximierung, das der Varianzanalyse zugrunde liegt. Die varianzanalytische Schätzung minimiert Fehlervarianz und maximiert daher Effektvarianz. Da die Effektgröße aufgrund von Stichprobenfehlern praktisch niemals Null wird, kommt es zu einer Überschätzung des Effekts (Stevens, 1992).

Welchen Unterschied macht es, ob man sich für einen Populationsschätzer entscheidet? "The answer is, it depends" (Snyder & Lawson, 1993). In welchem Ausmaß der empirische Effekt den Populationseffekt überschätzt hängt von vielen Faktoren ab.

- *Stichprobengröße*. Je größer die Stichprobe, desto besser die Schätzung der Effekt-

größe.

• *Anzahl der unabhängigen Variablen* und deren Stufen. Je größer die Anzahl der Probanden pro Zelle (im Design), desto besser ist die Schätzung.

• *Größe des Effekts.* Je größer der Effekt, desto kleiner ist der Fehler eines Stichprobenmaßes. Maßnahmen zur Maximierung von Effekten, reduzieren zwar den Schätzfehler von Stichprobenmaßen, sie sind aber (wie oben dargestellt) ein Problem für die Vergleichbarkeit von Effektgrößen.

Wer an einer möglichst präzisen Schätzung eines Effekts interessiert ist, sollte daher größere Stichproben verwenden und sich weniger auf die Ergebnisse der empfohlenen Power-Analyse stützen.

Globale oder partialisierte Effekte?

Wenn alle partiellen Varianzverhältnisse (hier ist auch das generalisierte Omega dazu zu zählen) aus einer Varianzanalyse zusammengezählt werden resultiert in der Summe immer mehr als 1. Das bedeutet nicht, dass mehr Varianz aufgeklärt wurde, als tatsächlich beobachtbar ist. Die partiellen Größen berechnen den Effekt lediglich derart, als wäre eine ANOVA gerechnet worden, die ausschließlich den interessierenden Effekt beinhaltet. Sechrechst and Yeaton (1982) kritisieren die partiellen Größen wegen dieser Eigenschaft und favorisieren globalen Varianzaufklärer. Doch außer der "Befriedigung durch die ästhetische und geistige Vollendung" (wie Susskind & Howland, 1980 das ausdrücken) hat diese Eigenschaft der globalen Effektgößen keinerlei nutzbringende Bedeutung. Stattdessen sind globale Maße eher problematisch hinsichtlich ihrer Vergleichbarkeit zwischen verschiedenen Studien (wie oben bereits ausführlich dargestellt).

Konfidenzintervalle

Ein Populationsschätzer erlaubt es also, den wahren Effekt besser vorher zu sagen, sie sollten damit auf jeden Fall den Vorzug erhalten. Doch ganz so eindeutig ist dei Sachlage auch hier nicht. Auch Effektgrößenschätzer sind von Stichprobenfehlern betroffen, das heißt, es lassen sich Konfidenzintervalle bestimmen. Wenn das Verhältnis Stichprobengröße zur Anzahl der Zellen sehr ungünstig ist, ist dann das Konfidenzintervall nicht auch bei Populationsschätzern zu groß, um eine sinnvolle Aussage zu ermöglichen. Wäre es möglicherweise

von Vorteil, wenn auf Konfidenzintervalle zurückgegriffen würde? Wenn der Null-Effekt innerhalb des Konfidenzintervalls liegt, dann ist das Ergebnis nicht signifikant. Somit würde ein Konfidenzintervall einer Effektgröße den Signifikanztest einschließen. Darüber hinaus würde es nicht nur eine Einschätzung der Größe des Effekts liefern, sondern auch eine Einschätzung der Präzision dieser Schätzung.

Formeln zu Bestimmung von Konfidenzintervallen zum $\hat{\varphi}^2$ und dem partiellen $\hat{\omega}_p^2$ bietet Fowler (1985). Hedges und Olkin (Hedges, 1981; Hedges & Olkin, 1985) stellen Prozeduren zur Schätzung von Konfidenzintervallen bei standardisierten Mittelwertsdifferenzen dar.

Berechenbarkeit/ Verfügbarkeit

Da Effektgrößen in den Statistikpaketen immer noch weitgehend ein Schattendasein führen ist dieser Punkt – trotz schneller Computer – nach wie vor von Bedeutung. Am einfachsten verfügbar sind denn auch vor allem diejenigen Effektgrößen, die von den Statistikpaketen selbst ausgegeben werden (wie z.B. das partielle Eta, das auf Wunsch von SPSS ausgedruckt wird).

Davon abgesehen sind die Populationsschätzer aufgrund der komplizierteren Formel grundsätzlich schwieriger zu bestimmen. Zusätzlich muss man beachten, ob Messwiederholung vorliegt oder nicht, ob in dem messwiederholten Modell Additivität gegeben ist oder nicht (d.h. keine Interaktionen mit dem Probandenfaktor).

Eine hohe Vergleichbarkeit eines Effektmaßes nützt wenig, wenn es nichts zu vergleichen gibt. Effektgrößen sollten auch aus bereits vorliegenden Publikationen post-hoc berechenbar sein, denn bisher werden sie nur selten berichtet. Abgesehen von dem generalisierten Varianzaufklärern ist es für alle übrigen hier beschriebenen Größen möglich, den Effekt aus dem F-Wert,den Freiheitsgraden und den Angaben zum Design zu errechnen. Wiederum problematisch sind Populationsschätzer in messwiederholten Analysen.

In den meisten Zeitschriften werden die mittleren Quadratsummen des Fehlers (MS_{Fehler}) nicht berichtet, was die Berechnung eines Hedges' g erschwert. Seifert (1991) beschreibt Techniken, mit deren Hilfe die MS_{Fehler} rekonstruiert werden kann.

Omnibus–Effekte vs. Einzelvergleiche

In der Diskussion um Effektgrößen wird gelegentlich bemängelt, dass diese Größen besonders bei Omnibus-Effekten in die Irre leiten können (Levin, 1967). Schließlich ist bei einer aufgeklärten Varianz von 18% eines Faktors mit 5 Stufen nicht klar, welche dieser Stufen in welchem Ausmaß an den 18% beteiligt ist. Dieses kritische Argument ist nun keines, dass spezifisch wäre für die Verwendung von Effekgrößen, es ist ein allgemeines Problem des Omnibus-Tests. Darum sollte auch bei Folgetests nicht auf Effektgrößen verzichtet werden, wenigstens für diejenigen Effekte, die für die Aussage einer Arbeit besonders relevant sind.

Nun stellt sich die Frage, ob es sinnvoll ist, für Einzelvergleiche andere Effektgrößen zu verwenden, als für Omnibus-Tests? Prinzipiell lässt sich jede (der hier dargestellten) Effektgrößen auch auf den Einzelvergleich anwenden. Die standardisierten Mittelwertsdifferenzen haben den Vorteil, dass bei Gruppenvergleichen die Vorzeichen erhalten bleiben und man der Effektgröße ansieht, welche Gruppe die größeren Werte hat. Auch insgesamt stehen sie in ihrer Logik dem Einzelvergleich näher, während die Varianzverhältnisse (Phi, Eta/Omega/Epsilon) der Logik des F-Tests folgen.

Voraussetzungen der Effektgrößenschätzer

Damit Effektgrößen aus unterschiedlichen Studien vergleichbare Ergebnisse wiedergeben, sollten die Variablen normalverteilt sein. Abbildung 5 verdeutlicht dies. Unterschiedliche Verteilungsformen führen zu unterschiedlichen Standardabweichungen, und dadurch verändern sich standardisierte Effektgrößen, obwohl die Effekte als gleichwertig zu betrachten wären.

Die Populationsschätzer derjenigen Effektgrößen, die den Effekt als Varianzverhältnisse ausdrücken ($\hat{\omega}^2$, $\hat{\epsilon}^2$, $\hat{\varphi}^2$) sind an zwei maßgebliche Voraussetzungen gebunden: die Homogenität der Varianzen und ein balanciertes Design (Vaughan & Corballis, 1969). Carroll and Nordholm (1975) haben Computersimulationen anhand eines einfaktoriellen dreistufigen Designs durchgeführt. Bezüglich Varianzheterogenität konnten lediglich vernachlässigbare Effekte auf diese beiden Schätzer nachgewiesen werden. Die kombinierte Verletzung beider Voraussetzungen führte allerdings zur systematischen Überschätzung der vorgegebenen Effekte. Diese Verzerrung war umso gravierender, je kleiner die Stichproben waren.

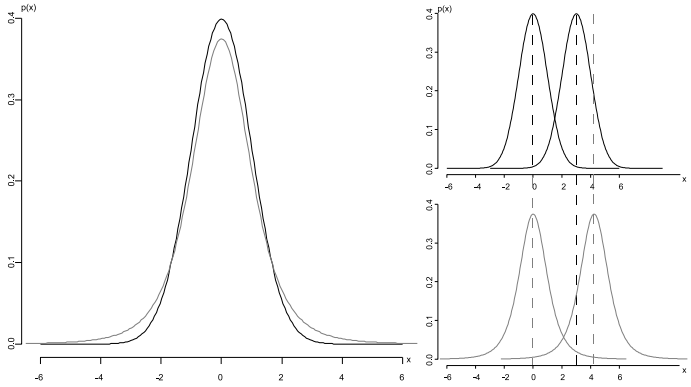

Abbildung 5. Links: eine "echte" Normalverteilung (dunkel) und eine veränderte mit stärkeren Ausläufern (hell). Rechts: Die beiden unterschiedlichen Verteilungsformen führen trotz unterschiedlich großer Effekte zu denselben Effektgrößen.

Standardisierte Mittelwertsdifferenzen sind einfacher zu handhaben. Wenn die Standardabweichungen innerhalb der Gruppen verschieden sind, kann man sich für eine Referenzgruppe entscheiden und die Standardabweichung innerhalb dieser Gruppe verwenden.

Bewertung von Effekten

Feste Wertebereiche

Die Maße aufgeklärter Varianz haben einen festen Wertebereich zwischen Null und Eins. Ob dies tatsächlich ein Vorteil ist, darüber herrscht nicht unbedingt Einigkeit. Auf den ersten Blick scheint ein fester Wertebereich eine intuitives Verständnis zu fördern, denn eine Größe zwischen Null und Eins lässt sich als Prozentanteil auffassen. Doch hier besteht durchaus die Gefahr, dass die praktische Bedeutsamkeit falsch eingeschätzt wird, insbesondere, wenn das Varianzverhältnis sehr klein ist. Bei einer aufgeklärten Varianz von 5% bleiben schließlich noch 95% ohne Erklärung (Rosenthal & Rubin, 1979)! Der Wissenschaftler mag ein derartig ungünstiges Verhältnis zwischen erklärter und nicht-erklärter Varianz leicht gering schätzen. Dennoch, in einem Präventivprogramm zur Krebsvorsorge würde ein 5%-Effekt zu einer milliardenschweren Entlastung der Gesundheitssysteme führen.

Ein Effekt ist nicht interpretierbar, wenn es keinen empirischen Rahmen gibt, in den ein Forschungsergebnis eingeordnet werden könnte. Seine *Bedeutung* erhält ein Effekt durch den Vergleich mit anderen Effekten aus anderen Arbeiten zu ähnlichen Fragestellungen.

"$\hat{\varphi}^2 = 1.05$" ist unmöglich interpretierbar, solange nicht bekannt ist, ob der Effekt in anderen Studien ähnlich groß war, oder deutlich kleiner, oder deutlich größer.

Klassifikationen

Die verschiedenen Effektgrößen sind ohne eine Art Bezugssystem nicht interpretierbar. Bei der Stichprobenumfangsplanung existiert darüber hinaus oftmals keine genaue Vorstellung, wie groß der Effekt sein sollte. Wenn keine Vorbild-Studien vorhanden sind, gibt es in der Tat keine Anhaltspunkte für den zu erwartenden Effekt. Cohen (1962) hat deshalb vorgeschlagen, Effektgrößen einzuteilen in drei Kategorien: große, mittlere und kleine Effekte. Er ermittelte die Werte empirisch, indem er den 60er Jahrgang der Zeitschrift "Journal of Abnormal Psychology" durchgesehen und alle Effekte berechnet hat. Diese Einteilung soll dem Psychologen nun als Anhaltspunkt dienen für die Bewertung eines Effektes.

So praktisch und sinnvoll dies in manchen Situationen aber auch sein kann, so problematisch und irreführend kann eine derartige Klassifikation manchmal sein. Nach Cohen gab es eine Reihe weiterer Autoren, die solche Klassifikationen für Zeitschriften aus anderen Teilgebieten der Psychologie aufgestellt haben, und es zeigen sich teilweise doch gehörige Diskrepanzen (Überblick in Sedlmeier & Gigerenzer, 1989). Was in dem einen Forschungszweig ein sinnlos kleiner Effekt ist, kann anderswo für revolutionäre Entdeckungen sorgen. Es gilt also, vertraue einer Klassifikation nur mit bedacht. Dennoch kann sie in vielen Fällen eine Orientierungshilfe sein.

Tabelle 2: Einteilung der Effektgrößen nach Cohen (1962)

	η^2	r	φ^2	$\hat{\omega}^2$	d	ε
Klein	0,010	0,100	0,020	0,010	0,200	0,100
Mittel	0,059	0,300	0,150	0,100	0,500	0,250
Groß	0,138	0,500	0,350	0,250	0,800	0,400

Zusammenfassung

Je idealer ein Effektmaß nach methodischen Kriterien ist, desto unhandlicher ist es leider auch. Das generalisierte Omega hat zwei Vorteile, die es beinahe ideal erscheinen lassen. Zum einen ist es von allen diskutierten Effektgrößen dasjenige, bei dem die Vergleichbarkeit zwischen verschiedenen Experimenten am größten ist. Und zum zweiten bewegt sich die Größe des Omega viel näher an den von Cohen (1988) vorgeschlagenen Richtlinien zur Bewertung von Effektgrößen. Es hat aber auch entscheidende Nachteile. Die Berechnung des entsprechenden Populationsschätzers ($\hat{\omega}^2_{gen}$) ist sehr kompliziert. Darüber hinaus ist es post hoc praktisch unmöglich zu bestimmen, weil keine Publikation hierfür die notwendigen Daten bereit hält. Ein dritter Punkt ist, dass es keine Software gibt, die eine Poweranalyse auf der Basis eines generalisierten Omega durchführen würde.

Dem generalisierten Omega steht das partielle Omega am Nächsten. Es ist von einigen Designmerkmalen immer noch unabhängig und es lässt sich relativ einfach schätzen. Darüber hinaus steht es in einer einfachen mathematischen Beziehung zum Phi. Das Phi wiederum wird in Programmen zur Poweranalyse (wie etwa G-Power, Erdfelder, Faul, & Buchner, 1996) verwendet.

Tabelle 3: Eigenschaften der beschriebenen Effektgrößen.

	Populations–schätzer	Designun–abhängigkeit	Berechen–barkeit	Post hoc	Werte–bereich
Φ^2	+	±	+	+	$0 \leq \Phi^2$
Ω^2_{gl}	+	−	±	+	$0 \leq \Omega^2 \leq 1$
Ω^2_{p}	+	±	+	+	$0 \leq \Omega^2_p \leq 1$
Ω^2_{gen}	+	+	−	−	$0 \leq \Omega^2_{gen} \leq 1$
D_{Cohen}	+	+	+	−	$0 \leq D_{(C)}$
D_{Hedges}	+	+	+	−	$0 \leq G_{(H)}$

Schließlich...

Die vorliegende Arbeit hatte es sich zum Ziel gemacht, eine möglichst "leicht verdauliche" Einführung über das Thema der Effektgrößenschätzer zu bieten. Darüber hinaus sollte

eine möglichst unvoreingenommene Diskussion der Vor– und Nachteile der einzelnen Größen, den Leser in die Lage versetzen, sich seine eigene Meinung zu bilden.

Effektgrößen werden in der wissenschaftlichen Praxis noch selten eingesetzt. In der Einleitung habe ich die Hypothese aufgestellt, dass diese Vernachlässigung zurückzuführen sei auf die komplizierte Literatur zu diesem Thema. Die Diskussion mag dem ein weiteres Argument hinzufügen. Es gibt keinen optimalen Effektgrößenschätzer. Es gibt zu viele Meinungen über deren jeweilige Eigenschaften und zahlreiche Schwierigkeiten bei der Interpretation.

Doch trotz aller Schwierigkeiten, muss diese Arbeit dennoch mit einem kurzen Plädoyer enden: Viele der angesprochenen Probleme sind nicht spezifisch für Effektgrößen, sondern gelten generell für die Varianzanalyse (Notwendigkeit von Einzelvergleichen, Erfüllung varianzanalytischer Voraussetzungen, breite Konfidenzintervalle bei kleinen Stichproben).

Relativiert am Nutzen, den Effektgrößen für das Verständnis experimenteller Daten haben können, ist diese Unsicherheit ein eher geringes Problem. Es ist nicht ganz einfach und unproblematisch, Effektgrößen zu interpretieren, und es kann im Extremfall sogar zu Fehlschlüssen kommen. Aber es auf jeden Fall besser, als sich lediglich auf den Signifikanzkoeffizienten zu verlassen, denn Effektmaße verschaffen ein tieferen Einblick in die Struktur der Daten – in derselben Weise, wie Korrelationen dies seit Jahrzehnten tun. Und Korrelationskoeffzienten sind seit jeher ein nützliches Instrument, wer wollte das bezweifeln...?

Literatur

Baugh, F. (2002, April). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement*, *62*(2), 254-263.

Bortz, J. (1993). *Statistik für Sozialwissenschaftler*. Berlin: Springer Verlag.

Carroll, R. M., & Nordholm, L. A. (1975). Sampling characteristics of Kelley's ϵ^2 and Hays' $\hat{\omega}^2$. *Educational and Psychological Measurement*, *35*, 541-554.

Cohen, J. (1962). The statistical power of abnormal-social psychological research. *Journal of Abnormal and Social Psychology*, *65*(3), 145–153.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), (p. 95-121). New York: McGraw-Hill.

Cohen, J. (1973). Eta–Squared and Partial Eta–Squared in fixed factor ANOVA Designs. *Educational and Psychological Measurement*, *33*, 107–112.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. ed.). New York: Academic Press.

Dodd, D. H., & Schultz, R. F. (1973). Computational procedures for estimating magnitude of effect for some analysis of variance designs. *Psychological Bulletin*, *79*(6), 391-395.

Erdfelder, E., Faul, F., & Buchner, A. (1996). Gpower: A general power analysis program. *Behavior Research Methods, Instruments and Computers*, *28*, 1-11.

Fowler, R. L. (1985). Point estimates and confidence intervals in measures of association. *Psychological Bulletin*, *98*(1), 160-165.

Gaebelein, J. W., & Soderquist, D. R. (1978). The utility of within-subjects variables: Estimates of strength. *Educational and Psychological Measurement*, *28*, 351-360.

Glass, G., & Hakstian, A. R. (1969). Measures of association in comparative experiments: Their development and interpretation. *American Education Research Journal*, *6*(3), 403-414.

Haase, R. F. (1983). Classical and partial eta squared in multifactor anova designs. *Educational and Psychological Measurement*, *43*, 35-39.

Hays, W. (1973). *Statistics for the social sciences*. New York: Holt, Rinehart and Winston.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107-128.

Hedges, L. V. (1986). Issues in meta–analysis. *Review of Research in Education*, *13*, 353–398.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. London: Academic Press.

Kelley, T. L. (1935). An unbiased Correlation Ratio Measure. *Proceedings of the National Academy of Science*, *21*, 554–559.

Keren, G., & Lewis, C. (1979). Partial omega squared for ANOVA designs. *Educational and Psychological Measurement*, *39*(1), 119-128.

Kotrlik, J. W., & Williams, H. A. (2003). The Incorporation of Effect Size in Information Technology, Learning, and Performance Research. *Information Technology, Learning, and Performance Journal*, *21*(1), 1–7.

Levin, J. R. (1967). Comment: Misinterpreting the significance of "explained variation". *American Psychologist*, *22*, 675–676.

Levine, T. R., & Hullett, C. R. (2002). Eta Squared, Partial Eta Squared and Misreporting of Effect Size in Communication Research. *Human Communication Research*, *28*(4), 612–625.

MacDonald, P. A., Joordens, S., & Seergobin, K. N. (1999). Negative priming effects that are bigger than a breadbox: Attention to distractors does not eliminate negative priming, it enhances it.

Memory and Cognition, *27*(2), 197-207.

Maxwell, S. E., Camp, C. J., & Arvey, R. D. (1981). Measures of strength of association: A comparative examination. *Journal of Applied Psychology*, *66*(5), 525-534.

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods*, *2*(1), 3-19.

McClelland, G. H. (2000). Increasing statistical power without increasing sample size. *American Psychologist*, *55*(8), 963-964.

O'Grady, K. E. (1982). Measures of explained variance: Cautions and limitatios. *Psychological Bulletin*, *92*(3), 766-777.

Olejnik, S., & Algina, J. (2000). Measures of Effect Size for Comparative Studies: Applications, Interpretations, Limitations. *Contemporary Educational Psychology*, *25*, 241-286.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434-447.

Pearson, K. (1905). *Mathematical contributions to the theory of evolution: XIV. On the general theory of skew correlation and nonlinear regression.* London: Dulau.

Pierce, C. A., Block, R. A., & Aguinis, H. (2004, December). Cautionary note on reporting eta-squared values from multifactor anova design. *Educational and Psychological Measurement*, *64*(6), 916-924.

Richardson, J. T. E. (1996). Measures of effects size. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 12-22.

Rosenthal, R., & Rubin, D. B. (1979). A note on percent variance explained as a measure of the importance of effects. *Journal of Applied Social Psychology*, *9*(5), 395-396.

Sechrechst, L., & Yeaton, W. H. (1982). Magnitude of experimental effects in social science research. *Evaluation Review*, *6*(5), 579-600.

Sedlmeier, P., & Gigerenzer, G. (1989). Do Studies of Statistical Power Have an Effect on the Power of Studies? *Psychological Bulletin*, *105*(2), 309–316.

Seifert, T. L. (1991). Determining effects size in various experimental designs. *Educational and Psychological Measurement*, *51*, 341-347.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education*, *61*(4), 334–349.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Susskind, E., & Howland, E. (1980, Nov). Measuring effect magnitude in repeated measures ANOVA designs: implications for gerontological research. *Journals of Gerontology*, *35*(6), 867-876.

Tatsuoka, M. (1993). Effect Size. In G. Keren & C. Lewis (Eds.), *A Handbook for data analysis in the behavioural sciences: methodological issues* (p. 461-479). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Vaughan, G., & Corballis, M. (1969). Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin*, *72*, 204-213.

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient for multiple correlation. *Annals of Mathematical Statistics*, *2*, 440-457.

Winkler, R. L., & Hays, W. L. (1975). *Statistics: Probability, inference, and decision* (2nd ed.). New York: Holt, Rinehart and Winston.

Anhang

Anhang der verwendeten Zeichen

Effektgrößen in griechischen Großbuchstaben (Ω^2, H^2) stellen immer die Populationsgrößen dar, griechische Kleinbuchstaben (ω^2, η^2) die entsprechende Stichprobengröße und wird der Kleinbuchstabe mit einem Dach versehen ($\hat{\omega}^2, \hat{\eta}^2$), so handelt es sich um einen Schätzer für den Populationseffekt. Zeichen, die hier nicht erläutert sind, werden im Text spezifiziert.

df – Freiheitsgrade einer Wahrscheinlichkeitsverteilung

df_{Effekt} – Freiheitsgrade des Zählers im F-Bruch

df_{Fehler} – Freiheitsgrade des Nenners im F-Bruch

F – Wert einer F-Verteilung; in der ANOVA: $F = MQS_{Effekt}/MQS_{Fehler}$

MQS – **M**ittlere **Q**uadrat**s**umme ($MQS = \hat{s}^2 = QS/df$)

MQS_{Effekt} – Mittlere Quadratsumme im Zähler des F-Bruchs

MQS_{Fehler} – Die dem Effekt entsprechende mittlere Quadratsumme des Fehlers; in der ANOVA ist dies die Prüfvarianz im Nenner des F-Bruchs

MQS_{Total} – Gesamte mittlere Quadratsumme

N_{obs} – Anzahl der Beobachtungen (Zellen × Probanden pro Zelle)

N_{vpn} – Anzahl der Versuchspersonen

QS – Summe der quadrierten Abweichung (Quadratsumme)

QS_{Effekt} – Quadratsumme zwischen den Gruppen

QS_{Fehler} – Quadratsumme innerhalb der Gruppen

QS_{Total} – Gesamt-Quadratsumme

s – empirisch bestimmte Standardabweichung

s^2 – Stichprobenvarianz

$\hat{s}^2$ – Schätzer der Populationsvarianz aus den Stichprobendaten

t_{df} – Wert einer Student-t-Verteilung (mit df Freiheitsgraden)

α – Fehler 1. Art

λ – (empirischer) Nonzentralitätsparameter einer nonzentralen Verteilung

μ – Mittelwert in einer Population

θ^2 – Varianzkomponente

σ^2 – Populationsvarianz

Φ^2, ϕ^2 – Standardisierter Nonzentralitätsparameter

$\Omega^2_{gl}, \omega^2_{gl}$

H^2_{gl}, η^2_{gl} – Globales Maß der aufgeklärten Varianz

$\Omega^2_{gen}, \omega^2_{gen}$

H^2_{gen}, η^2_{gen} – Partialisiertes Maß aufgeklärter Varianz (nach Olejnik & Algina, 2003)

Ψ – Ein Kontrastvektor $w_1 \cdot \mu_1 + w_2 \cdot \mu_2 + \ldots + w_n \cdot \mu_n$