Evaluation und Metaanalyse

MSc Klinische Psychologie und Psychotherapie

SoSe 2024

Prof. Dr. Dirk Ostwald

(2) Metaanalyse

Motivation

Randomisierte kontrollierte Studien

Effektstärke

Selbstkontrollfragen

Motivation

Randomisierte kontrollierte Studien

Effektstärke

Selbstkontrollfragen

Definitionsversuch

Metaanalyse

... ist eine Gruppe quantitiver Methoden zur Kombination von Evidenz

... ist die Analyse der Resultate statistischer Analysen

... ist überwiegend Frequentistisch geprägt

Glass (1976) "Primary, Secondary, and Meta-Analysis of Research"

"My major interest currently is in what we have come to call — not for want of a less pretentious name — the **meta-analysis of research**. The term is a bit grand, but it is precise, and apt, and in the spirit of"meta-mathematics," "meta-psychology," and "meta-evaluation." Meta-analysis refers to the analysis of analyses. I use it to refer to the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings. It connotes a rigorous alternative to the casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding research literature."

Glass (1976)

Ethischer Imperativ

- Die Öffentlichkeit investiert substantielle Mittel in die (außer)universitäre Forschung.
- Neben wenigen Großforschungsprojekten gibt es viele kleine Forschungsprojekte.
- Die Kleinforschungsprojekte sind oft nicht zentralisiert und interessens- und zufallsgeleitet.
- Empirische Wissenschaft basiert auf der Replikation oder Nichtreplikation von Erkenntnissen.
- Die Analyse von Kleinforschungsergebnissen hat den potentiellen Mehrwert von Großforschung.
- Es ist intuitiv sinnvoll und plausibel nach der gesammelten Evidenz für etwas zu fragen.
- Metaanalysen wird das Potential zugeschrieben, aussagekräftiger als Einzelstudien zu sein.
- Entscheidungsrelevante Aussagen werden gerne aufgrund großer Datenmengen getroffen.
- Forschung hat die Pflicht, integrierte Evidenz dem Wohle der Gesellschaft bereitzustellen.

Borenstein (2009), Harrer et al. (2021)

Motivation

Narrative Reviews

- Standard der Akkumulation wissenschaftlicher Evidenz bis etwas 1990
- Subjektive und intransparente Auswahl von Studien durch Expert:innen
- Nicht-explizite und qualitative Gewichtung von Studien
- Relativ gut möglich bei eher geringer Anzahl von zu inkludierenden Studien

Systematische Reviews und Metaanalysen

- Standard der Akkumulation wissenschaftlicher Evidenz seit etwa 2000
- Transparente Auswahl von Studien anhand festgelegter Kriterien
- Explizite, quantitative Wichtung von Studien mithilfe metaanalytischer Statistik
- Hilfreich bei einer großen Anzahl von zu inkludierenden Studien durch Automatisierung

Borenstein (2009), Harrer et al. (2021)

Metaanalysen in der Psychotherapieforschung

Eysenck (1952)

"A survey was made of reports on the improvement of neurotic patients after psychotherapy, and the results compared with the best available estimates of recovery without benefit of such therapy. **The figures fail to support the hypothesis that psychotherapy facilitates recovery from neurotic disorder**. In view of the many difficulties attending such actuarial comparisons, no further conclusions could be derived from the data whose shortcomings highlight the necessity of properly planned and executed experimental studies into this important field."

Smith (1977)

"Results of nearly 400 controlled evaluations of psychotherapy and counseling were coded and integrated statistically. **The findings provide convincing evidence of the efficacy of psychotherapy**. On the average, the typical therapy client is better off than 75% of untreated individuals. Few important differences in effectiveness could be established among many quite different types of psychotherapy. More generally, virtually no difference in effectiveness was observed between the class of all behavioral therapies (systematic desensitization, behavior modification) and the nonbehavioral therapies (Rogerian, psychodynamic, rational-emotive, transactional analysis, etc.)."

Cuijpers et al. (2019)

"In the 1950s, Eysenck suggested that psychotherapies may not be effective at all. Twenty-five years later, the first meta-analysis of randomised controlled trials showed that the effects of psychotherapies were considerable and that Eysenck was wrong. However, since that time methods have become available to assess biases in meta-analyses. We examined the influence of these biases on the effects of psychotherapies for adult depression, including risk of bias, publication bias and the exclusion of waiting list control groups. The unadjusted effect size of psychotherapies compared with control groups was g = 0.70 (limited to Western countries: g = 0.63) (...). Only 23% of the studies could be considered as a low risk of bias. **When adjusting for several sources of bias, the effect size across all types of therapies dropped to g = 0.31.**"

Borenstein (2009), Harrer et al. (2021)

Metaanalysen in der Medizin

- Inhaltliche Beiträge, z.B. Peto and Parish (1980)
- Methodische Beiträge, z.B. DerSimonian and Laird (1986)
- Richtlinien, z.B. Moher et al. (2009)

Cochrane Collaboration

- Stiftung zur Förderung der evidenzbasierten metaanalytischen Medizin seit 1993
- www.cochrane.org

Campbell Collaboration

- Stiftung zur Förderung der evidenzbasierten metaanalytischen Sozialwissenschaft seit 2000
- www.campbellcollaboration.org

Typische Charakteristika des Metaanalysen-Genres

- Spezifikation von Suchtermen für Studiendatenbanken wie Web Of Science
- Spezifikation des Studienscreenings
- Extraktion von Effektstärken (Cohen's d)
- Korrektur von Effektstärken (Hedges' g)
- Angabe eines "mittleren" Hedges' g
- Mehrebenenanalyse mithilfe eines Linear Mixed Models
- Versuche der Quantifizierung und Korrektur von Biases

Borenstein (2009), Harrer et al. (2021)

Typische Probleme des Metaanalysen-Genres

Apples and Oranges

- Metaanalysen integrieren Studien, die sich in experimentellen Details unterscheiden
- Die zulässige experimentelle Varianz hängt von der metaanalytischen Fragestellung ab

Garbage In, Garbage Out

- Die Qualität eines metaanalytischen Ergebnisses hängt von der Qualität der Primärstudien ab
- Bei geringer Primärstudienqualität kann eine Metaanalyse zumindest diese Erkenntnis bringen

The Filedrawer Problem

- Nicht alle themenrelevante Forschung wird tatsächlich auch publiziert
- Insbesondere sind meist hochwertig angefertigte Nullresultatstudien unterrepräsentiert

Researcher Agendas

- Jeder Studienreport, auch eine Metaanalyse, ist letztlich ein Kommunikationsakt
- Die Kommunikationssender sind bewusst oder unbewusst nicht frei von eigenen Motiven

Borenstein (2009), Harrer et al. (2021)

Motivation

Randomisierte kontrollierte Studien

Effektstärke

Selbstkontrollfragen

Randomisierte kontrollierte Studie

⇒ Randomized controlled trial (RCT) ⇒ Goldstandard der Interventionsforschung

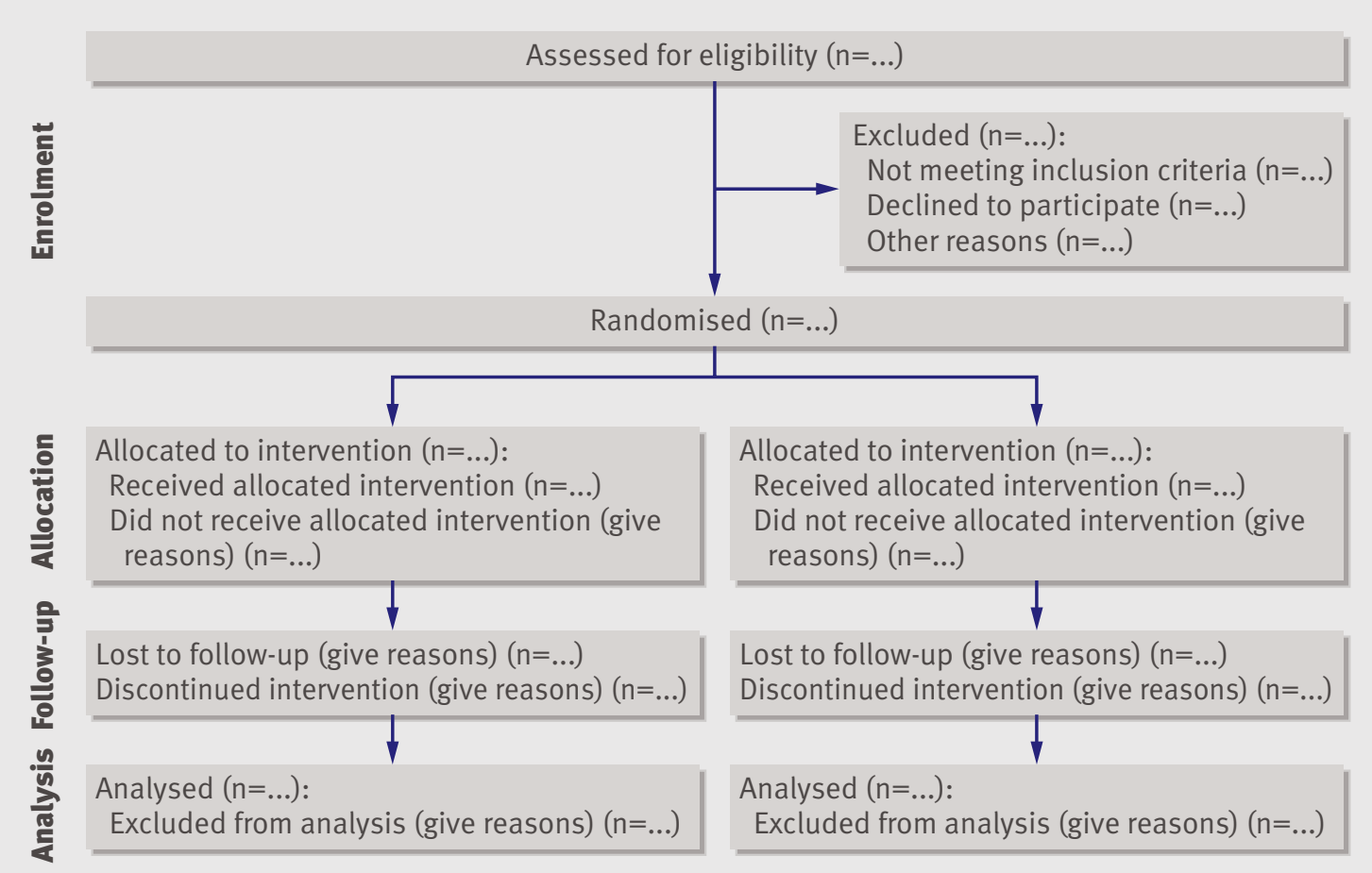
Randomisierte Studie

- Zuordnung zu einer Behandlungsgruppe nach dem Zufallsprinzip
- Gewährleistung der gleichmäßigen Verteilung von Einflüssen auf alle Gruppen

Kontrollierte Studie

- Ergebnisse in der Studiengruppe werden mit denen einer Kontrollgruppe verglichen
- Kontrollgruppe z.B. keine Intervention, Placebo, beste verfügbare Behandlung

Randomisierte kontrollierte Studien



Flow diagram of the progress through the phases of a parallel randomised trial of two groups (that is, enrolment, intervention allocation, follow-up, and data analysis)

Schulz et al. (2010)

Allgemeine Systematik von Studiendesigns

Randomisierte kontrollierte Studie (Experiment)

- Die experimentellen Einheiten werden den Studienbedingungen zufällig zugeordnet
- Beispiel: Online Psychotherapie vs. Face-To-Face Psychotherapie bei Depression

Nicht-randomisierte kontrollierte Studie (Quasiexperiment)

- Untersuchung natürlich bzw. bereits bestehender Gruppen
- Beispiel: Online Psychotherapie bei Depression vs. Schizophrenie

Analyse eines bestehenden Datensatzes (Korrelationstudie)

- Nicht-randomisierte, nicht kontrollierte Studie
- Beobachtungsstudie ohne Intervention
- Beispiel: Analyse von Paneldaten

Charakteristika randomisierter kontrollierter Studien

- Vorhandensein einer kausaltheoretischen Hypothese vor Studienbeginn
- Gute Manipulierbarkeit von unabhängigen Variablen
- Explizite Operationalisierung der untersuchten Konstrukte
- Kontrollierbarkeit möglichst vieler Studienbedingungen
- Typisch für bereits gut erschlossene Gegenstandsbereiche

Faktorielle Studiendesigns

- Kategoriale unabhängige Variable, die Faktor genannt wird
- Die Werte der unabhängigen Variablen werden Level genannt
- Einfaktorielle oder mehrfaktoriell

Parametrische Studiendesigns

- Kontinuierliche unabhängige Variable
- Die Werte der unabhängigen Variablen werden oft Level genannt
- Meist einfaktoriell

Designschemata

- R: Randomisierung
- O: Observation (Test, Messung)
- X: Exposition experimenteller Bedingung
- Experimentelle Bedingungen von oben nach unten
- Zeitliche Abfolge von links nach rechts

Beispiel

R	X	O
R		O

- Bedingungszuweisung erfolgt durch Randomisierung
- Nur eine Gruppe erhält das Treatment
- Beide Gruppen absolvieren die Messung

Randomisierte einfaktorielle Studiendesigns

- Gesamtgruppe wird zufällig auf experimentelle Bedingungen aufgeteilt
- Eine unabhängige Variable mit zwei oder mehr Leveln
- Populäres Designs in der klinischen Forschung
- Varianten
    - No-Treatment Kontrollgruppe
    - Placebo Kontrollgruppe
    - Vergleich zweier Treatments
    - Zwei-Treatment Vergleich mit Placebo-Kontrollgruppe
    - Pre-Posttest Designs

No-Treatment Kontrollgruppe

R	X	O
R		O

- Vergleich eines Treatments zu keinem Treatment

Placebo Kontrollgruppe

R	X	O
R	X_P	O

- Placebo = Scheintreatment
- Vergleich eines Treatments zu keinem Treatment
- Kontrolle studieninduzierter Effekte (Placeboeffekte)

Vergleich zweier Treatments

R	X_A	O
R	X_B	O

- Vergleich Standardtreatment A und neues Treatment B
- Keine Aussage über Effektivität des Standardtreatments

Zwei-Treatment Vergleich mit Placebo-Kontrollgruppe

R	X_A	O
R	X_B	O
R	X_P	O

- Vergleich Standardtreatment A und neues Treatment B
- Aussage über Effektivität des Standardtreatments möglich
- Placebotreatment kann ethisch nicht vertretbar sein

Beispiel: Einfluss von Psychotherapie auf Depressionssymptomatik

- Face-To-Face Psychotherapie (A)
- Online Psychotherapie (B)
- Seelsorge (P)

$\rightarrow$ Keine Aussagen über Pre-Treatment Gruppenunterschiede möglich

$\rightarrow$ Keine Aussage über Dropout Charakteristika möglich

Pre-Posttest Designs

R	O	X_A	O
R	O	X_B	O
R	O		O

- Fokus auf Treatment-induzierte Verbesserungen/Verschlechterungen
- Subtraktion von Pre-Test-Gruppenunterschieden möglich
- Untersuchung von Dropout Charakteristika möglich
- Mögliches Auftreten von Testeffekten (Lernen, Gewöhnung, Ermüdung)
- Höherer Zeit- und Kostenaufwand

Beispiel: Evaluation von Psychotherapiesettings bei Depression

Setting

Face-to-Face PreBDI  PostBDI

Online PreBDI  PostBDI

⇒ Randomisiertes einfaktorielles Preposttest Design ohne Kontrollgruppe

Motivation

Randomisierte kontrollierte Studien

Effektstärke

Selbstkontrollfragen

Primary outcome measure

- Ein zumeist quantitatives Maß, das den Effekt einer Intervention messen soll
- Festlegung im Rahmen von RCTs üblicherweise vor Studienbeginn
- Bei RCTs oft Grundlage von Poweranalysen

Beispiele bei RCTs im Bereich Psychotherapieforschung zur Depression

- Mittlere BDI-II Reduktion
- Anzahl an Treatment-Erfolgen

Primary outcome measure Extraktion für Metaanalysen $\Rightarrow$ Cohen's d

Cohen's d in den Worten von Jacob Cohen

"Thus, we see that the absence of the phenomenon under study is expressed by a null hypothesis which specifies an exact value for a population parameter, one which is appropriate to the way the phenomenon under study is manifested. Without intending any necessary implication of causality, **it is convenient to use the phrase "effect size" to mean "the degree to which the phenomenon is present in the population," or "the degree to which the null hypothesis is false."**

Whatever the manner of representation of a phenomenon in a particular research in the present treatment, the null hypothesis always means that the effect size is zero. By the above route, it can now readily be made clear that when the null hypothesis is false, it is false to some specific degree, i.e., the effect size (ES) is some specific nonzero value in the population. The larger this value, the greater the degree to which the phenomenon under study is manifested. Thus, in terms of the previous illustrations:

1. If the percentage of males in the population of psychiatric patients bearing a diagnosis of paranoid schizophrenia is 52%. and the effect is measured as a departure from the hypothesized 50%. the ES is 2%; if it is 60%, the ES is 10%, a larger ES.
2. If children of multiple births have a population mean IQ of 96, the ES is 4 IQ units (or- 4, depending on directionality of significance criterion); if it is 92, the ES is 8 (or - 8) IQ units, i.e., a larger ES."
3. If the population product moment r between neurophysiological and questionnaire measures of introversion-extroversion is .30, the ES is .30; if r is .60, so is t he ES, a larger value and a larger departure from the null hypothesis, which here is r $= 0$.
4. If the population of consumers preferring brand A has a median annual income \$700 higher than that of brand B, the ES is \$700. If the population median difference and hence the ES is \$1000, the effect of income on brand preference would be larger.

Thus, whether measured in one unit or another, whether expressed as a difference between two population parameters or the departure of a population parameter from a constant or in any other suitable way, the ES can itself be treated as a parameter which takes the value zero when the null hypothesis is true and some other specific nonzero value when the null hypothesis is false, and in this way the ES serves as an index of degree of departure from the null hypothesis."

Cohen (1988) Statistical power analysis for the behavioral sciences

Cohen's d in den Worten von Jacob Cohen

2.2 The Effect Size Index: d

As noted above (Section 1.4), we need a "pure" number, one free of our original measurement unit, with which to index what can be alternately called the degree of departure from the null hypothesis of the alternate hypothesis, or the ES (effect size) we wish to detect. This is accomplished by standardizing the raw effect size as expressed in the measurement unit of the dependent variable by dividing it by the (common) standard deviation of the measures in their respective populations, the latter also in the original measurement unit. For the two independent samples case, this is simply

$$\mathbf{d} = \frac{\mathbf{m_A} - \mathbf{m_B}}{\sigma} \tag{2.2.1}$$

for the directional (one-tailed) case, and

$$\mathbf{d} = \frac{|\mathbf{m_A} - \mathbf{m_B}|}{\sigma} \tag{2.2.2}$$

for the nondirectional (two-tailed) case,

where $\mathbf{d}$ = ES index for $\mathbf{t}$ tests of means in standard unit,
$\mathbf{m_A}, \mathbf{m_B}$ = population means expressed in raw (original measurement) unit, and
σ = the standard deviation of either population (since they are assumed equal).

Cohen (1988) Statistical power analysis for the behavioral sciences

Cohen's d in den Worten der metaanalytischen Community

Computing *d* and *g* from studies that use independent groups

We can estimate the standardized mean difference (δ) from studies that used two independent groups as

$$d = \frac{\overline{X}_1 - \overline{X}_2}{S_{within}}. \tag{4.18}$$

In the numerator, $\overline{X}_1$ and $\overline{X}_2$ are the sample means in the two groups. In the denominator S_{within} is the within-groups standard deviation, pooled across groups,

$$S_{within} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \tag{4.19}$$

where n_1 and n_2 are the sample sizes in the two groups, and S_1 and S_2 are the standard deviations in the two groups. The reason that we pool the two sample

estimates of the standard deviation is that even if we assume that the underlying population standard deviations are the same (that is $\sigma_1 = \sigma_2 = \sigma$), it is unlikely that the sample estimates S_1 and S_2 will be identical. By pooling the two estimates of the standard deviation, we obtain a more accurate estimate of their common value.

The sample estimate of the standardized mean difference is often called Cohen's *d* in research synthesis. Some confusion about the terminology has resulted from the fact that the index δ, originally proposed by Cohen as a *population parameter* for describing the size of effects for statistical power analysis is also sometimes called *d*. In this volume we use the symbol δ to denote the effect size parameter and *d* for the sample estimate of that parameter.

Borenstein (2009) Introduction to Meta-Analysis

Cohen's d im Rest der Welt

(Theorem) T-Teststatistik des Zweistichproben-T-Tests

Gegeben sei das Zweistichproben-T-Tests Szenario des Allgemeinen Linearen Modells und es sei μ_0 der Nullhypothesenparameter. Dann ergibt sich für die T-Teststatistik, dass

$$T = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\frac{\bar{y}_1 - \bar{y}_2 - \mu_0}{s_{12}} \right) \tag{1}$$

und es gilt

$$T \sim t(\delta, n_1 + n_2 - 2) \text{ mit } \delta = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\frac{\mu_1 - \mu_2 - \mu_0}{\sigma} \right). \tag{2}$$

∘

Für eine ausführliche Diskussion, siehe Probabilistische Datenwissenschaft für die Psychologie.

Definition (Cohen's d des Zwei-Gruppen-Designs)

Gegeben seien ein Zwei-Gruppen-Design mit einer Treatmentgruppe und einer Kontrollgruppe und es seien $y_{11}, ..., y_{1n_1}$ und $y_{21}, ..., y_{2n_2}$ die skalaren reellen Werte des Primary Outcome Measures jeder Gruppe. Es seien weiterhin

$$\bar{y}_1 := \frac{1}{n}\sum_{i=1}^{n_1} y_{1i}, \quad \bar{y}_2 := \frac{1}{n}\sum_{i=1}^{n_2} y_{2i}, \quad s_1^2 := \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2 \text{ und } s_2^2 := \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2 \tag{3}$$

die Mittelwerte und die empirischen Varianzen beider Gruppen, respektive. Schließlich sei

$$s := \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{4}$$

die *gepoolte Standardabweichung* der beiden Gruppen. Dann ist *Cohen's d für das des Zwei-Gruppen-Design* definiert als

$$d := \frac{\bar{y}_1 - \bar{y}_2}{s}. \tag{5}$$

Bemerkungen

- d misst den Unterschied der Mittelwerte in Einheiten der gepoolten Standardabweichung, z.B.

$$d = 0 \Leftrightarrow \bar{y}_1 - \bar{y}_2 = 0, \quad d = 1 \Leftrightarrow \bar{y}_1 - \bar{y}_2 = s, \quad d = 2 \Leftrightarrow \bar{y}_1 - \bar{y}_2 = 2s. \tag{6}$$

- Für $\mu_0 = 0$ gilt im Sinne der Zweistischproben-T-Teststatistik T, dass

$$T = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\, d \Leftrightarrow d = \sqrt{\frac{n_1 + n_2}{n_1 n_2}}\, T. \tag{7}$$

Motivation

Randomisierte kontrollierte Studien

Effektstärke

Selbstkontrollfragen

Selbstkontrollfragen

1. Erläutern Sie, warum Metaanalysen aus ethischer Sicht sinnvoll sind.
2. Erläutern Sie den Unterschied zwischen Narrativen Reviews und Systematischen Reviews/Metaanalysen.
3. Skizzieren Sie die Geschichte der Metaanalyse zur Wirksamkeit der Psychotherapie anhand von Eysenck (1952), Smith (1977), Cuijpers et al. (2019).
4. Nennen und erläutern Sie vier typische Probleme des Metaanalyse-Genres.
5. Erläutern Sie den Begriff der randomisierten kontrollierten Studie.
6. Erläutern Sie den Begriff des randomisierten kontrollierten Pre-Posttest Design anhand eines Beispiels.
7. Erläutern Sie den Begriff des Primary Outcome Measures.
8. Geben Sie die Definition von Cohen's d für das Zwei-Gruppen-Design wieder.

Referenzen

Borenstein, Michael, ed. 2009. *Introduction to Meta-Analysis*. Chichester, U.K: John Wiley & Sons.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, N.J: L. Erlbaum Associates.

Cuijpers, P., E. Karyotaki, M. Reijnders, and D. D. Ebert. 2019. "Was Eysenck Right After All? A Reassessment of the Effects of Psychotherapy for Adult Depression." *Epidemiology and Psychiatric Sciences* 28 (1): 21–30. https://doi.org/10.1017/S2045796018000057.

DerSimonian, Rebecca, and Nan Laird. 1986. "Meta-Analysis in Clinical Trials." *Controlled Clinical Trials* 7 (3): 177–88. https://doi.org/10.1016/0197-2456(86)90046-2.

Eysenck, H J. 1952. "The Effects of Psychotherapy: An Evaluation." *Journal of Consulting Psychology* 16 (5): 319–24.

Glass, Gene V. 1976. "Primary, Secondary, and Meta-Analysis of Research," 7.

Harrer, Mathias, Pim Cuijpers, Toshi A. Furukawa, and David D. Ebert. 2021. *Doing Meta-Analysis with R: A Hands-On Guide*. 1st ed. Boca Raton: Chapman and Hall/CRC. https://doi.org/10.1201/9781003107347.

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, and The PRISMA Group. 2009. "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement." *PLoS Medicine* 6 (7): e1000097. https://doi.org/10.1371/journal.pmed.1000097.

Peto, R, and S Parish. 1980. "Aspirin After Myocardial Infarction." *The Lancet* 315 (8179): 1172–73. https://doi.org/10.1016/S0140-6736(80)91626-8.

Schulz, K. F, D. G Altman, D. Moher, and for the CONSORT Group. 2010. "CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials." *BMJ* 340 (mar23 1): c332–32. https://doi.org/10.1136/bmj.c332.

Smith, Mary Lee. 1977. "Meta-Analysis of Psychotherapy Outcome Studies." *American Psychologist* 32 (9): 752–60.