**Systematisches Review (Master)**

**Was? Welche Fragen beantwortet das Modell?**

Welche Studien sind bisher zu meiner Fragestellung erschienen? Was ist der aktuelle Forschungsstand?

**Wann? Welche Arten von Daten werden analysiert?**

Stichprobe: Studien.

**Hinweise zur konkreten Umsetzung**

Häufig wird an ein systematisches Review eine Metaanalyse angeschlossen. Wie bei Forschungsvorhaben üblich ist auch bei einem systematischen Review eine konkrete Planung des Vorhabens für eine saubere Umsetzung notwendig. Guidelines, Checklisten und Flowcharts findet ihr unter PRISMA.

Bei einem systematischen Review ist es besonders wichtig, bereits zu Beginn klare Einschlusskriterien für die Auswahl relevanter Studien festzulegen. Diese Kriterien helfen nicht nur dabei, die Fragestellung gezielt zu beantworten, sondern dienen auch als Grundlage für die Auswahl passender Suchbegriffe bei der Literaturrecherche.

Ein typisches Beispiel: In der klinischen Psychologie werden häufig Therapiestudien betrachtet. Hier sind randomisierte kontrollierte Studien (RCTs) zu einem bestimmten therapeutischen Verfahren von besonderem Interesse.

**Von der Recherche zum Screening**

Sobald die Literaturrecherche abgeschlossen ist, werden zunächst doppelte Einträge entfernt. Anschließend erfolgt die Verteilung der Artikel auf mehrere Personen: Jeweils zwei Personen prüfen unabhängig voneinander dieselben Abstracts. Wenn ein Abstract die festgelegten Kriterien erfüllt – oder wenn die Entscheidung unklar ist – wird der Artikel für das Screening des Volltexts vorgemerkt.

Dabei gilt eine einfache Regel: Ein Artikel wird zum Volltext-Screening zugelassen, wenn mindestens eine der beiden prüfenden Personen ihn einschließt.

**IRR Berechnen**

**Warum ist die IRR wichtig?**

Um sicherzustellen, dass alle Beteiligten die Einschlusskriterien einheitlich verstehen und anwenden, ist es sinnvoll, die **Interrater-Reliabilität (IRR)** zu prüfen – also die Übereinstimmung zwischen den Bewertungen der verschiedenen Personen.

Besonders während des Abstract-Screenings lohnt es sich, nach einer gewissen Anzahl an geprüften Artikeln die IRR zu berechnen. So kann frühzeitig erkannt werden, ob es Missverständnisse bei der Anwendung der Kriterien gibt.

**IRR berechnen – so geht’s**

Zunächst ist es wichtig, sich das Skalenniveau der Übereinstimmung und die Anzahl der Beurteiler:innen anzuschauen. Je nachdem ist eine andere Methode für die Berechnung der IRR notwendig.

Bei einem systematischen Review bewerten z.B. häufig zwei Rater, über Einschluss bzw. Ausschluss einer Studie. Dementsprechend liegt ein dichotomes **nominales** Skalenniveau mit **zwei Beurteiler:innen** vor, wofür ***Cohen’s Kappa*** ein gängiges statistisches Maß ist. Auch wenn nominalskalierte Variablen mit mehr als zwei Kategorien vorliegen, ist Cohen’s Kappa geeignet. Neben nominalen Bewertungen treten manchmal auch ordinale oder intervallskalierte Bewertungen auf. Für **ordinalskalierte** Variablen eignet sich ein **gewichtetes Cohen’s Kappa (Weighted kappa)**. Bei der Berechnung der Übereinstimmung ist dabei auch die Größe der Abweichungen zwischen den Beurteiler:innen berücksichtigt. Je stärker die Abweichung, desto stärker wird diese bestraft.

Bei **mehr als zwei Beurteiler:innen** und kategorialen Variablen ist ein andere Methode notwendig. Die Forschung berechnet die Übereinstimmung meist mit ***Fleiss Kappa*.** Fleiss’ Kappa setzt voraus, dass die **Rater zufällig auf die Objekte verteilt** sind.  
Das macht es besonders nützlich in Studien, wo **nicht alle Rater jedes Objekt** beurteilen, sondern unterschiedliche Subsets von Ratern beteiligt sind.

Für **intervallskalierte** Variablen ist der Intraclass Correlation Coefficient gängig. Der **Intraclass Correlation Coefficient (ICC)** ist ein Maß für die Zuverlässigkeit oder Übereinstimmung von metrischen Bewertungen, die von **zwei oder mehr Ratern** oder Messungen stammen.

Ein Beispiel zur Berechnung der verschiedenen Maße mit **R** findet ihr im Lehrportal.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. [doi:10.2307/2529310](https://doi.org/10.2307/2529310)

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*(3), 613–619. [https://doi.org/10.1177/001316447303300309](https://psycnet.apa.org/doi/10.1177/001316447303300309)

**Qualitäts Assessment**

Nachdem das Screening der Volltexte abgeschlossen ist, sind die gefundenen Studien hinsichtlich ihrer Qualität zu prüfen. Dafür gibt es verschiedene Instrumente, die in der Forschung etabliert sind. Für die Ermittlung der Qualität von RCTs eignet sich z.B. die Psychotherapy Quality Rating Scale (PQRS) mit insgesamt 25 Items. Die Methode ist von der American Psychiatric Association mit dem Ziel eine standardisierte Methode der Qualitätsüberprüfung einzuführen (Kocsis et al., 2010). Die Skala ist explizit für Psychotherapie Studien entwickelt und hat gute psychometrische Eigenschaft in Bezug auf die interne Konsistenz und externe Validität gezeigt (Trotta et al., 2024).

Des Weiteren ist bei RCTs zu überprüfen, ob ein Verzerrungsrisiko vorliegt. Dafür eignet sich das weit verbreitete „Cochrane risk of Bias Tool“ (Sterne et al., 2019), welches mögliche Verzerrungen durch den Prozess der Randomisierung, die Zuweisung zur Intervention, fehlende Ergebnisdaten, die Messmethode und eine selektive Ergebnisdarstellung anschaut. Die Studien werden für jede Domäne und insgesamt anhand der Skala „low risk of bias“, „some concerns“ und „high risk of bias“ bewertet. Das Instrument gibt konkrete Vorgaben für das Einsortieren.

Bei naturalistischen Studien muss ein anderes Instrument angewendet werden. Das National Institute for Health Research hat verschiedene Instrumente entwickelt, um die interne Validität einer Studie zu prüfen. Das „Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies” eignet sich für die Bewertung von naturalistischen Studien. Dabei erfolgt die Bewertung anhand von 14 Kriterien z.B. inwiefern die Forschungsfrage verständlich ist, ob die Messzeitpunkte sinnvoll gewählt wurden oder ob die Ergebnismessung valide ist (National Heart, Lung and Blood Institute, o. J.).

Kocsis, J. H., Gerber, A. J., Milrod, B., Roose, S. P., Barber, J., Thase, M. E., ... & Leon, A. C. (2010). A new scale for assessing the quality of randomized clinical trials of psychotherapy. Comprehensive psychiatry, 51(3), 319-324.

National Heart, Lung and Blood Institute, N. (o. J.). Study Quality Assessment Tools | NHLBI, NIH. Abgerufen 18. Mai 2025, von https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools.

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., … Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. BMJ, 366, l4898. <https://doi.org/10.1136/bmj.l4898>

Trotta A, Gerber AJ, Rost F, Robertson S, Shmueli A and Perelberg RJ (2024) The efficacy of psychodynamic psychotherapy for young adults: a systematic review and meta-analysis. Front. Psychol. 15:1366032. doi: 10.3389/fpsyg.2024.1366032