Meta-Analyse
Meta-Analysis

Wolfgang Viechtbauer

1 Einleitung

Das Ziel einer Meta-Analyse ist eine systematische Zusammenfassung und ein Vergleich der Ergebnisse von verwandten Studien. Im Vergleich zu einzelnen Studien, wo Versuchspersonen untersucht werden, bilden die in einer Meta-Analyse berücksichtigten Studien nun die Menge der Untersuchungseinheiten. Die Meta-Analyse grenzt sich von traditionellen narrativen Literaturübersichten („narrative literature reviews") dadurch ab, dass statistische Verfahren benutzt werden, um die Synthese der Literatur zu systematisieren und dadurch replizierbarer (sprich: wissenschaftlicher) zu machen. Um dies zu gewährleisten, ist es daher unerlässlich, alle Schritte, Entscheidungen und Beschlüsse in einer Meta-Analyse umfassend zu dokumentieren (z. B. die Suchstrategien um relevante Studien zu finden, die Einschlusskriterien für Studien, die Analyseschritte).

2 Notwendigkeit von Meta-Analysen

Dass die narrative Literaturübersicht in manchen Forschungsbereichen von der Meta-Analyse bereits abgelöst wurde, lässt sich zum einen mit der Informationsexplosion in der wissenschaftlichen Literatur (Adair & Vohra, 2003) und zum anderen mit den Unzulänglichkeiten traditioneller Literaturzusammenfassungen erklären (Light & Pillemer, 1984). Wegen des rapide wachsenden Literaturumfangs findet man bei einer systematischen Suche zu einem bestimmten Thema meist eine große Anzahl von relevanten Artikeln. In narrativen Literaturübersichten fallen daher die Beschreibung der Studien und deren Ergebnisse oft selektiv, unsystematisch und subjektiv aus. Zur Vereinfachung der Darstellung beziehen sich narrative Literaturübersichten meist auf die statistische Signifikanz der Ergebnisse, was leicht zu fehlerhaften Interpretationen führen kann.

Folgendes (zum Zwecke der vereinfachten Darstellung konstruiertes) Beispiel soll das Problem verdeutlichen. In drei Studien wurde die Kriteriumsvalidität von Vorstellungsgesprächen untersucht. Die auf Vorstellungsgesprächen basierenden Beurteilungen der Bewerber wurden dabei mit Messungen der Arbeitsproduktivität der Bewerber korreliert. Die Studien wurden mit übereinstimmenden Stichprobengrößen durchgeführt ($n = 72$) und lieferten Korrelationen von .07, .22 und .27. Nur die Korrelation in der dritten Studie ist statistisch signifikant (bei einem

zweiseitigen Test mit $\alpha = .05$). In zwei der drei Studien kann also die Nullhypothese $H_0: \rho = 0$ (d. h. Vorstellungsgespräche haben keine prädiktive Validität) nicht verworfen werden. Basierend auf einer Auszählung der signifikanten und nicht signifkanten Ergebnisse (einem sogenannten „vote count") müsste man entweder die Validität von Vorstellungsgesprächen in Frage stellen oder wenigsten von widersprüchlichen Ergebnissen sprechen.

Die Schlussfolgerung, dass die Ergebnisse von Studien sich widersprechen, ist aber oft falsch, da die Teststärke in vielen Fällen relativ klein ist. Wenn man z. B. davon ausgeht, dass die wirkliche Kriteriumsvalidität von Vorstellungsgesprächen .20 beträgt (vgl. McDaniel, Whetzel, Schmidt & Maurer, 1994), dann lag die Wahrscheinlichkeit, ein signifikantes Ergebnis zu erzielen, in jeder der drei Studien gerade einmal bei 39 % (Cohen, 1988, Tab. 3.3.5, S. 93; siehe auch → Effektgröße und Teststärke). Der Wurf einer Münze wäre also ein Test mit höherer Schärfe gewesen.

Im Vergleich ergibt eine Zusammenfassung der drei Ergebnisse mit Hilfe von meta-analytischen Verfahren eine geschätzte Korrelation von .19, die statistisch signifikant ist. Darüber hinaus finden sich keine Anzeichen von Diskrepanz in den Ergebnissen (d. h., die Unterschiede in den beobachteten Korrelationen lassen sich auf die zu erwartende Stichprobenvarianz zurückführen). Die Meta-Analyse liefert also nicht nur eine direkte Schätzung der Kriteriumsvalidität von Vorstellungsgesprächen, sondern zeigt in diesem Fall auch, dass die Ergebnisse widerspruchsfrei sind.

3 Schritte in einer Meta-Analyse

Nach Cooper (1998) kann man den Prozess der Durchführung einer Meta-Analyse in fünf Schritte unterteilen: (a) Formulierung der Fragestellung, (b) Suche nach relevanten Daten/Ergebnissen (d. h. Literaturrecherche), (c) Beurteilung und Kodierung der Daten, (d) Analyse und Interpretation und (e) Präsentation der Ergebnisse.

Im ersten Schritt muss die Fragestellung (d. h. das Forschungsziel) klar definiert werden. Hier werden die zentralen Variablen konzeptuell und operativ definiert und in Bezug zueinander gesetzt. Im Allgemeinen geht es bei Meta-Analysen um drei Arten von Fragestellungen: *Um die Größe einer univariaten Statistik* (z. B. wie viel Prozent der Bevölkerung leiden an Depressionen?), *um den Zusammenhang von zwei Variablen* (z. B. wie stark korrelieren die Symptome von Depressionen und Angststörungen?) oder *um Unterschiede zwischen zwei natürlich oder experimentell definierten Gruppen* (z. B. wie effektiv ist Johanniskraut im Vergleich mit Placebos zur Behandlung von Depressionen?). In jedem Fall muss

genau bestimmt werden, welche Populationen, welche methodologischen und inhaltlichen Rahmenbedingungen und operativen Definitionen von abstrakten Konstrukten für die Meta-Analyse relevant und zulässig sind. Das Ziel der Meta-Analyse ist meistens nicht nur eine Zusammenfassung der einzelnen Ergebnisse in einen Punktschätzer, sondern eine Untersuchung der Variabilität in den Ergebnissen. Diese beruht oft auf systematischen Unterschieden zwischen den Studien. Somit kann der Einfluss von moderierenden Drittvariablen untersucht werden. Dabei ist es wichtig, mögliche Drittvariablen im Voraus und theoriegeleitet zu definieren. Umfangreiche Kenntnisse des gängigen Fachvokabulars, des jeweiligen Forschungsstandes, über bereits vorhandene Literaturübersichten und der aktuellen Theorien innerhalb des zu untersuchenden Forschungsbereichs sind bei diesem ersten Schritt unerlässlich.

Basierend auf den Ein- und Ausschlusskriterien geht es beim zweiten Arbeitsschritt darum, alle relevanten Ergebnisse zusammenzutragen. Man beginnt meist mit einer Suche innerhalb von elektronischen Datenbanken (z. B. PsycINFO, PSYNDEX, MedLine, ERIC). Weiterhin sollten die Literaturverzeichnisse von relevanten Publikationen durchsucht werden. Zitierungsregister (z. B. SCI, SSCI) sind ebenfalls hilfreich, wenn bereits relevante Studien ermittelt wurden. Informellere Informationskanäle (z. B. persönliche Kontakte, Mailing-Listen), Forschungsregister (z. B. solche, wie sie von der Cochrane und der Campbell Collaboration geführt werden), Diplom- und Dissertationsregister (z. B. Dissertation Abstracts International), Kongressprogramme und selbst eine direkte Internetsuche können weitere Treffer liefern. Gegebenenfalls ergiebig, aber sehr zeitaufwändig kann auch eine Handsuche von einschlägigen Zeitschriften sein.

Beim dritten Arbeitsschritt werden nun die relevanten Ergebnisse aus den einzelnen Studien extrahiert und in eine vergleichbare Metrik übersetzt, d. h. die Ergebnisse müssen auf eine vergleichbare Weise quantifiziert werden (s. u.). Außerdem werden potenzielle Drittvariablen systematisch kodiert. Die Interraterreliabilität der Ergebnisextraktion und der Kodierungen ist zu überprüfen. Ein gründliches Training der Kodierer, ein sorgfältig und detailliertes Kodierhandbuch und ein dazu passendes Kodierschema sind daher unerlässlich.

Die Art der zu kodierenden Drittvariablen ist vom Inhalt der Meta-Analyse abhängig. So interessieren z. B. bei einer Zusammenfassung von Studien zur Effektivität einer therapeutischen Behandlungsmethode Aspekte der Behandlung (z. B. Länge der Behandlung), die Art der Kontrollgruppe (z. B. Placebo- versus Wartelistegruppe), Eigenschaften der Studienteilnehmer (z. B. durchschnittliches Alter), das Studiensetting (z. B. stationäre versus ambulante Behandlung) oder die Art der Messinstrumente (z. B. Selbstbewertungen versus Bewertungen mittels eines klinischen Interviews). Bei einer Zusammenfassung von Studien, die den Zusammenhang zwischen zwei Variablen untersuchen, wird die Stärke der Assoziation

eventuell durch Merkmale der Studienteilnehmer oder die Art und Weise, wie die Variablen operationalisiert wurden, beeinflusst. Im Allgemeinen sollten methodologische Eigenschaften, welche Aufschluss über die Qualität der Studien liefern können, kodiert werden. Somit lässt sich deren Einfluss auf die Ergebnisse systematisch untersuchen. Dies gilt meist als die bevorzugte Alternative zu einem A-priori-Ausschluss von Studien auf Grund methodologischer Qualitätsmerkmale.

Das Ziel des dritten Arbeitsschritts ist also die Erstellung eines Datensatzes mit den Ergebnissen und Werten der potenziellen Drittvariablen für jede der einzelnen Studien. Dieser Datensatz bildet nun die Grundlage für den vierten Arbeitsschritt, die Analyse und Interpretation der Ergebnisse. Im Allgemeinen geht es dabei um drei Kernaspekte der Daten: (a) die mittlere Tendenz der Ergebnisse, (b) die Variabilität in den Ergebnissen und (c) der Einfluss von Drittvariablen auf die Ergebnisse. Die zur Untersuchung dieser Aspekte einschlägigen statistischen Verfahren werden im sechsten Abschnitt dieses Kapitels an Hand eines Beispieles verdeutlicht. Dabei werden auch einige Möglichkeiten für die Präsentation der Ergebnisse vorgestellt.

4 Gruppendifferenz- und Assoziationsmaße

In den meisten Meta-Analysen sind die relevanten Ergebnisse der zu integrierenden Studien entweder als Gruppendifferenzen oder als Assoziationsstärken zu quantifizieren (die Integration von univariaten Statistiken bildet eher die Ausnahme). Im Allgemeinen kann man in diesen Fällen von Effektmaßen sprechen, d. h. bei Gruppendifferenzstudien handelt es sich um den Effekt der (dichotomen) Gruppierungvariablen auf die abhängige Variable, bei Assoziationsstudien um den Effekt einer Variablen auf die andere (wobei in beiden Fällen nicht notwendigerweise ein kausaler Effekt impliziert ist). Die zwei in den Verhaltenswissenschaften üblichsten Effektmaße für die Meta-Analyse werden nun kurz vorgestellt. Weitere Maße zur einheitlichen Quantifizierung der Ergebnisse von Studien werden von Fleiss (1994) und Rosenthal (1994) beschrieben.

Bei einer kontinuierlichen abhängigen Variablen werden Gruppendifferenzen üblicherweise durch die mittlere Differenz oder die standardisierte mittlere Differenz quantifiziert. Bei der Erstereren wird einfach die Differenz der Mittelwerte gebildet. Wie schon erwähnt, müssen allerdings die Ergebnisse der verschiedenen Studien so quantifiziert werden, dass eine Vergleichbarkeit gewährleistet ist. Zum Beispiel ist die mittlere Differenz zwischen einer Behandlungs- und einer Kontrollgruppe, gemessen mit dem Beck Depression Inventory, wegen der unterschiedlichen Skalierung nicht direkt vergleichbar mit einer mittleren Differenz, gemessen mit der Hamilton Depression Rating Scale. Daher werden Gruppen-

unterschiede meistens mit der standardisierten mittleren Differenz quantifiziert, d. h. innerhalb von Studie i berechnen wir

$$y_i = \frac{\bar{x}_i^{G1} - \bar{x}_i^{G2}}{s_i},$$

wobei $\bar{x}_i^{G1}$ und $\bar{x}_i^{G2}$ für die Mittelwerte der Gruppen und s_i für die Wurzel aus der gepoolten (zusammengelegten) Varianz der beiden Gruppen stehen. Der beobachtete y_i-Wert schätzt die wahre standardisierte mittlere Differenz in Studie i, d. h.,

$$\theta_i = \frac{\mu_i^{G1} - \mu_i^{G2}}{\sigma_i}.$$

Die Stichprobenvarianz von y_i kann mit

$$v_i = \frac{n_i^{G1} + n_i^{G2}}{n_i^{G1} n_i^{G2}} + \frac{y_i^2}{2(n_i^{G1} + n_i^{G2})}$$

geschätzt werden, wobei n_i^{G1} und n_i^{G2} die Gruppengrößen angeben (Hedges & Olkin, 1985).

Bei der Integration von Assoziationsstudien werden die Ergebnisse üblicherweise mit dem Pearsonschen Korrelationskoeffizienten quantifiziert, d. h., $y_i = r_i$ ist ein Schätzer der wahren Korrelation $\theta_i = \rho_i$ in Studie i. Die Stichprobenvarianz von r_i kann mit

$$v_i = \frac{(1 - r_i^2)^2}{n_i - 1}$$

geschätzt werden. Alternativ kann man vor der weiteren Analyse mittels der Fisher-Transformation

$$y_i = \frac{1}{2} \ln\left(\frac{1 + r_i}{1 - r_i}\right)$$

die Stichprobenvarianzen stabilisieren (d. h., die Größe der Stichprobenvarianz von y_i ist nun unabhängig von ρ_i). Nach der Transformation ist y_i ein Schätzer von

$$\theta_i = \frac{1}{2} \ln\left(\frac{1 + \rho_i}{1 - \rho_i}\right)$$

mit Stichprobenvarianz gleich

$$v_i = \frac{1}{n_i - 3}.$$

5 Beispiel

Die statistische Analyse von meta-analytischen Daten wird nun an Hand eines Beispiels verdeutlicht. Die Daten dafür basieren auf einer Meta-Analyse von Studien, die den sogenannten „Pygmalioneffekt" untersucht haben (Raudenbush, 1984; Raudenbush & Bryk, 1985). Dabei wurden künstliche Erwartungen bei Lehrern erzeugt, indem ein rein zufällig ausgewählter Teil der Schüler, also unabhängig von deren wirklichen Fähigkeiten, als besonders begabt charakterisiert wurde. Tatsächliche Unterschiede zwischen diesen und den übrigen Schülern wurden nach einer längeren Unterrichtsphase mittels Intelligenz- oder Begabungstests ermittelt.

Tabelle 1 zeigt jeweils die mittlere standardisierte Differenz zwischen den beiden Schülergruppen (positive Werte entsprechen einer gesteigerten Intelligenz inner-

Tabelle 1: Ergebnisse von 19 Studien zur Untersuchung des Pygmalioneffektes (Raudenbush, 1984; Raudenbush & Bryk, 1985)

i	Studie	y_i	v_i	Wochen	Blind
1	Rosenthal et al., 1974	0.03	.0156	2	0
2	Conn et al., 1968	0.12	.0216	3	0
3	Jose & Cody, 1971	–0.14	.0279	3	0
4	Pellegrini & Hicks, 1972	1.18	.1391	0	0
5	Pellegrini & Hicks, 1972	0.26	.1362	0	1
6	Evans & Rosenthal, 1968	–0.06	.0106	3	0
7	Fielder et al., 1971	–0.02	.0106	3	1
8	Claiborn, 1969	–0.32	.0484	3	0
9	Kester, 1969	0.27	.0269	0	0
10	Maxwell, 1970	0.80	.0630	1	1
11	Carter, 1970	0.54	.0912	0	1
12	Flowers, 1966	0.18	.0497	0	1
13	Keshock, 1970	–0.02	.0835	1	1
14	Henrikson, 1970	0.23	.0841	2	1
15	Fine, 1972	–0.18	.0253	3	0
16	Grieger, 1970	–0.06	.0279	3	1
17	Rosenthal & Jacobson, 1968	0.30	.0193	1	0
18	Fleming & Anttonen, 1971	0.07	.0088	2	1
19	Ginsburg, 1970	–0.07	.0303	3	0

halb der angeblich begabten Gruppe), die dazugehörige geschätzte Stichprobenvarianz und die Werte von zwei potenziellen Moderatorvariablen für 19 Studien dieser Art. Die erste Moderatorvariable spezifiziert die Anzahl der Wochen vorausgehenden Kontakts zwischen Lehrern und Schülern vor Beginn der Studien (kleinere Gruppenunterschiede sind möglicherweise zu erwarten, wenn Lehrer sich im Voraus einen eigenen Eindruck von den Fähigkeiten der Schüler machen konnten). Zur Vereinfachung wurden alle Zeiträume von drei oder mehr Wochen mit einer 3 kodiert. Die zweite Moderatorvariable gibt an, ob die abhängige Variable blind gemessen wurde (0 = nein, 1 = ja; da bei Studien ohne blinde Messung die Gefahr besteht, dass Lehrer bewusst oder unbewusst den besonders begabten Schülern beim Testen Hilfestellung leisten, könnten bei diesen Studien möglicherweise größere Gruppenunterschiede entstehen).

6 Analyse und Interpretation

Die statistische Auswertung der Ergebnisse innerhalb einer Meta-Analyse beginnt also mit den beobachteten Schätzwerten $(y_1, \ldots, y_k)$ der entsprechenden (unbekannten) wahren Parameterwerte $(\theta_1, \ldots, \theta_k)$ und den dazugehörigen (geschätzten) Stichprobenvarianzen $(v_1, \ldots, v_k)$ von $i = 1, \ldots, k$ unabhängigen Studien. Wir gehen nun von der Modellgleichung $y_i = \theta_i + \varepsilon_i$ aus, wobei wir annehmen, dass die ε_i-Werte, also die Differenzen zwischen den beobachteten und den wahren Werten, (annähernd) normalverteilt sind. Entsprechend kann ein 95 %iges Konfidenzintervall für θ_i mit $y_i \pm 1.96\sqrt{v_i}$ berechnet werden. Die beobachteten Schätzwerte mit entsprechenden Konfidenzintervallen können in einem Fehlerbalkendiagramm dargestellt werden (vgl. Abb. 1). In nur 3 der 19 Studien schließt das Konfidenzintervall den Wert Null aus, was bei diesen drei Studien ein statistisch signifikantes Ergebnis impliziert (d. h., die Nullhypothese H_0: $\theta_i = 0$ kann in diesen Fällen bei $\alpha = .05$ verworfen werden).

6.1 Das Modell fester Effekte

Im einfachsten Fall sind die wahren Effekt- oder Assoziationswerte homogen ($\theta \equiv \theta_1 = \ldots \theta_k$). Dieser Fall könnte zum Beispiel dann eintreten, wenn sich die zu integrierenden Studien kaum in ihrer Methodik und den Charakteristiken der Stichproben unterscheiden. Unter diesem sogenannten Modell fester Effekte ist damit jeder der y_i-Werte ein Schätzwert dieses homogenen Parameters. Das Ziel besteht nun darin, so gut wie möglich die wahre Effektgröße θ zu ermitteln. Da Schätzwerte mit einer kleineren Stichprobenvarianz durchschnittlich gesehen näher an θ liegen, wird diesen Werten auch mehr Bedeutung (sprich: Gewicht) in der Analyse beigemessen. Unter der Annahme, dass die Stichprobenvarianzen exakt bekannt sind, ist der effizienteste Schätzer von θ (d. h. der Schätzer mit der kleinsten Streuung) gleich

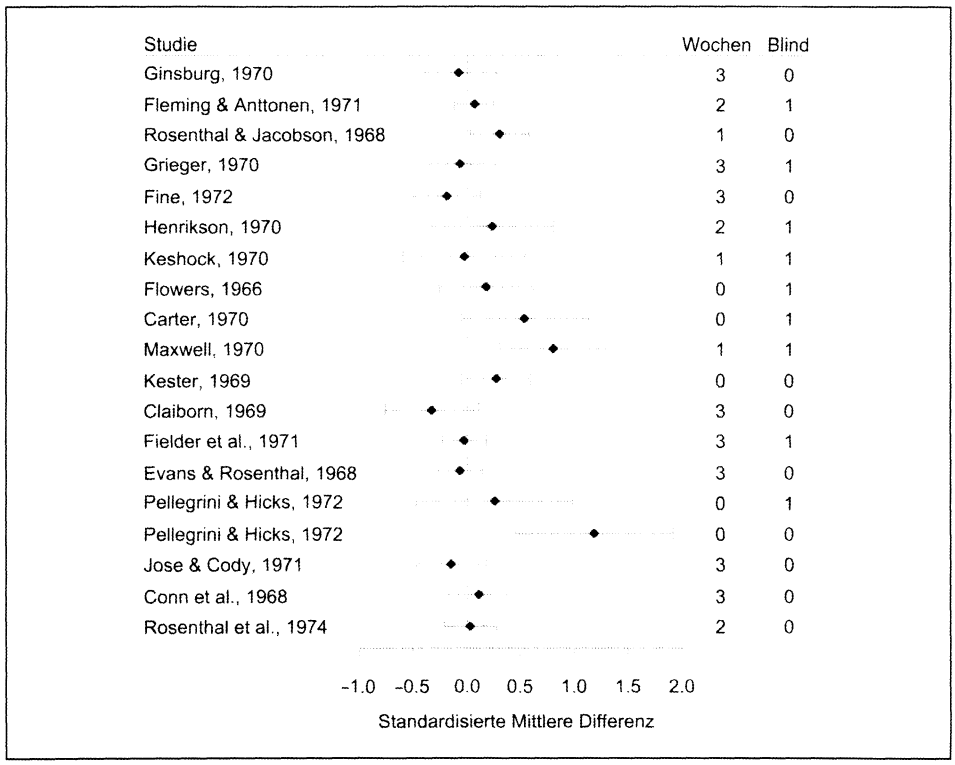

Abbildung 1: Standardisierte mittlere Differenz und entsprechendes Konfidenzintervall in 19 Studien zum Pygmalioneffekt

$$\hat{\theta} = \frac{\sum w_i y_i}{\sum w_i}$$

wobei $w_i = 1/v_i$. Der Standardfehler von $\hat{\theta}$ kann mit

$$SE(\hat{\theta}) = \sqrt{\frac{1}{\sum w_i}}$$

berechnet werden. Ein 95 %iges Konfidenzintervall für θ lässt sich anschließend mit $\hat{\theta} \pm 1.96\ SE(\hat{\theta})$ ermitteln (Hedges & Olkin, 1985).

Für die Beispieldaten finden wir $\hat{\theta} = 0.060$ und $SE(\hat{\theta}) = .0365$. Die Grenzen des 95 %igen Konfidenzintervalls liegen also bei –0.011 und 0.132, womit man die Nullhypothese H_0: $\theta = 0$ (d. h. der Effekt der Erwartungsinduktion ist gleich null) bei $\alpha = .05$ nicht verwerfen kann (der Wert 0 liegt innerhalb des Konfidenzintervalls). Die weiteren Ergebnisse zeigen jedoch, dass diese Schlussfolgerung voreilig wäre.

6.2 Heterogenität in den Ergebnissen

Es ist natürlich möglich und in vielen Fällen plausibel, dass die wahren Effekt- oder Assoziationswerte heterogen sind. Um die Nullhypothese H_0: $\theta_1 = \ldots \theta_k$ zu testen, ermitteln wir die Teststatistik $Q = \sum w_i (y_i - \hat{\theta})^2$ und vergleichen sie mit dem einseitigen kritischen Wert einer Chi-Quadrat-Verteilung mit $k-1$ Freiheitsgraden (Hedges & Olkin, 1985). Wird der kritische Wert überschritten, so muss die Nullhypothese verworfen werden und es ist von heterogenen Ergebnissen auszugehen.

Für die Beispieldaten finden wir $Q = 35.83$, was bei 18 Freiheitsgraden und einem kritischen Chi-Quadrat-Wert von 28.87 bei $\alpha = .05$ zu einem signifikanten Test führt. Wir schließen also daraus, dass die θ_i-Werte heterogen sind, d. h., der Effekt der Erwartungsinduktion ist innerhalb der Studien unterschiedlich groß.

Bei heterogenen Ergebnissen stellt sich die Frage, wie diese Heterogenität zu erklären ist. Zwei Ansätze sind üblich: Entweder betrachtet man die Heterogenität als das Ergebnis eines zufälligen Prozesses, oder man versucht, die Heterogenität mit systematischen Unterschieden zwischen den Studien (d. h. basierend auf den Moderatorvariablen) weitestmöglich zu erklären.

6.3 Das Modell zufallsvariabler Effekte

Das Modell zufallsvariabler Effekte nimmt an, dass die Heterogenität in den θ_i-Werten rein zufallsbedingt ist; d. h., man macht die Annahme, dass ein stochastischer Prozess den θ_i-Wert für eine Studie aus einer Population von möglichen θ_i-Werten generiert hat (Hedges & Olkin, 1985). Üblicherweise geht man davon aus, dass die Population von θ_i-Werten normalverteilt ist. Der Erwartungswert und die Varianz dieser Verteilung wird im Weiteren mit μ und τ_T^2 gekennzeichnet. Ziel der Analyse ist es nun, diese beiden Parameter zu schätzen. Das Modell zufallsvariabler Effekte berücksichtigt also zwei Varianzkomponenten: Zum einen die Stichprobenvarianz v_i (wodurch der beobachtete Effekt einer Studie y_i von θ_i abweicht) und zum anderen τ_T^2, also die Heterogenität in den wahren Effekten (wodurch θ_i von μ abweicht).

Im ersten Schritt schätzen wir τ_T^2, wobei eine ganze Reihe von Verfahren dazu existieren (Viechtbauer, 2005; Sidik & Jonkman, 2007). Meist verbreitet ist allerdings der Schätzer von DerSimonian und Laird (1986):

$$\hat{\tau}_T^2 = \frac{Q - (k-1)}{\sum w_i - \frac{\sum w_i^2}{\sum w_i}}.$$

Sollte der geschätzte Wert von τ_T^2 negativ sein, dann wird $\hat{\tau}_T^2$ auf Null gesetzt. Im nächsten Schritt werden die Gewichte mit $w_i = 1/(\hat{\tau}_T^2 + v_i)$ neu berechnet. Nun kann μ mit

$$\hat{\mu} = \frac{\sum w_i y_i}{\sum w_i}$$

und der Standardfehler von $\hat{\mu}$ mit

$$SE(\hat{\mu}) = \sqrt{\frac{1}{\sum w_i}}$$

geschätzt werden. Ein 95 %iges Konfidenzintervall für μ erhalten wir mit $\hat{\mu} \pm 1.96 \; SE(\hat{\mu})$. Da $\hat{\tau}_T^2$, also die geschätzte Varianz in den θ_i-Werten, schwer zu interpretieren ist, empfiehlt es sich zusätzlich $I^2 = (1-(k-1)/\mathrm{Q}) \times 100\,\%$ zu berechnen. Der I^2-Wert gibt an, wie viel der Gesamtvarianz in den beobachteten Ergebnissen (also in den y_i-Werten) durch Heterogenität in den θ_i-Werten erzeugt wurde (Higgins, Thompson, Deeks & Altman, 2003).

Für die Beispieldaten finden wir $\hat{\tau}_T^2 = 0.0259$, $\hat{\mu} = 0.089$ und $SE(\hat{\mu}) = 0.0558$. Das Konfidenzintervall umfasst also die Werte –0.020 bis 0.199, womit wir die Nullhypothese $H_0: \mu = 0$ (d. h. der durchschnittliche Effekt der Erwartungsinduktion ist gleich null) nicht verwerfen können. Der I^2-Wert beträgt 50 %; wir schätzen also, dass die Hälfte der gesamten Varianz in den Ergebnissen auf Heterogenität zurückzuführen ist.

6.4 Moderatorenanalyse

In vielen Fällen ist die Heterogenität in den Ergebnissen nicht zufallsbedingt, sondern wird (wenigstens zu einem Teil) durch systematische Unterschiede zwischen den Studien generiert. Mittels einer Moderatorenanalyse lassen sich solche Einflüsse untersuchen (Raudenbush, 1994). Wir gehen dabei von der Annahme aus, dass der Erwartungswert von θ_i eine lineare Funktion von p Moderatorvariablen ist, was wir mit der Regressionsgleichung

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

wiedergeben können, wobei x_{ij} den Wert von Moderatorvariable j in Studie i bezeichnet. Also ist β_0 der Erwartungswert von θ_i für $x_{i1} = \ldots = x_{ip} = 0$, und β_j gibt an, wie sich μ_i verändert, wenn x_{ij} um eine Einheit steigt. Zusätzliche Restheterogenität in den θ_i-Werten, die von den Moderatorvariablen nicht erklärt wird, gilt wiederum als rein zufällig und normalverteilt. Die auf Restheterogenität basierende Varianz wird im Weiteren mit τ_R^2 gekennzeichnet.

Die Gleichungen zum Schätzen von $\beta_0, \ldots, \beta_p$ und τ_R^2 lassen sich auf kompakte Weise mit Matrixnotation wiedergeben. Mit $\mathbf{y} = (y_1 \ldots y_k)'$ bezeichnen wir den Spaltenvektor der beobachteten Schätzwerte, mit $\mathbf{X}$ die $(k \times (p+1))$-Matrix mit Einsen in der ersten Spalte und den Werten der Moderatorvariablen in den übrigen Spalten und mit $\mathbf{W}$ die $(k \times k)$-Diagonalmatrix mit den $w_i = 1/v_i$ Gewichten auf der Diagonale. Nun lässt sich τ_R^2 mit

$$\hat{\tau}_R^2 = \frac{\mathbf{y'Py} - (k - p - 1)}{Spur[\mathbf{P}]}$$

schätzen, wobei $\mathbf{P} = \mathbf{W} - \mathbf{WX}(\mathbf{X'WX})^{-1}\mathbf{X'W}$ und $Spur[\mathbf{P}]$ die Summe der Diagonale in $\mathbf{P}$ bezeichnet (Raudenbush, 1994). Sollte $\hat{\tau}_R^2$ negativ sein, so wird der Wert auf Null gesetzt. Nun werden die Gewichte mit $w_i = 1/(\hat{\tau}_R^2 + v_i)$ und die Diagonalmatrix $\mathbf{W}$ neu berechnet. Die Parameter $\beta_0, \ldots, \beta_p$ werden dann mit

$$\mathbf{b} = (\mathbf{X'WX})^{-1}\mathbf{X'Wy}$$

und die entsprechende Kovarianzmatrix von $\mathbf{b} = (b_0 \ldots b_p)'$ mit

$$\hat{\mathbf{\Sigma}} = (\mathbf{X'WX})^{-1}$$

geschätzt. Die Wurzeln der Diagonalelemente von $\hat{\mathbf{\Sigma}}$ geben die Standardfehler von $b_0, \ldots, b_p$ an $(SE(b_0), \ldots, SE(b_p))$. Um zu testen, ob eine Moderatorvariable einen signifikanten Einfluss auf den Erwartungswert von θ_i hat (d. h. $H_0: \beta_j = 0$), vergleicht man $z_j = b_j / SE(b_j)$ mit den kritischen Werten einer Normalverteilung (also ± 1.96 bei $\alpha = .05$). Ein 95 %iges Konfidenzintervall für β_j lässt sich mit $b_j \pm 1.96 SE(b)$ berechnen. Der geschätzte Erwartungswert von θ_i für Moderatorvariablenwerte $\mathbf{x_i} = (1 \; x_{i1} \ldots x_{ip})$ ist $\mathbf{x_i b}$. Ein entsprechendes 95 %iges Konfidenzintervall für μ_i kann mit $\mathbf{x_i b} \pm 1.96\sqrt{\mathbf{x_i}\hat{\mathbf{\Sigma}}\mathbf{x_i}'}$ ermittelt werden. Bei der Interpretation der Ergebnisse ist es hilfreich, mittels $(\hat{\tau}_T^2 - \hat{\tau}_R^2)/\hat{\tau}_T^2 \times 100\,\%$ zu berechnen, wie viel der gesamten Heterogenität durch die Moderatorvariablen erklärt wird.

Für die Beispieldaten untersuchen wir nun das Modell $\mu_i = \beta_0 + \beta_1 Wochen_i + \beta_2 Blind_i$. Für dieses Modell finden wir $\hat{\tau}_R^2 = 0.0001$. Nach Neuberechnung der sich kaum ändernden Gewichte finden wir $\hat{\mu}_i = 0.394 - 0.156\, Wochen_i + 0.026\, Blind_i$. Die Standardfehler der Schätzwerte sind $SE(b_0) = 0.0974$, $SE(b_1) = 0.0366$ und $SE(b_2) = 0.0760$. Nur die Anzahl der Kontaktwochen vor der Erwartungsinduktion hat einen signifikanten Einfluss auf μ_i (da $z_1 = -4.28$ und $z_2 = 0.35$, kann nur $H_0: \beta_1 = 0$ verworfen werden). Für $Wochen_i = 0$ und $Blind_i = 0$ beträgt $\hat{\mu}_i = 0.394$. Das entsprechende 95 %ige Konfidenzintervall (0.204 bis 0.585) weist auf einen signifikanten Effekt der Erwartungsinduktion hin. Bei steigender Anzahl der Kontaktwochen sinkt allerdings der Effekt und nach zwei Kontaktwochen liegt der geschätzte Wert von μ_i nur noch bei 0.81. Da das Konfidenzintervall (–0.017

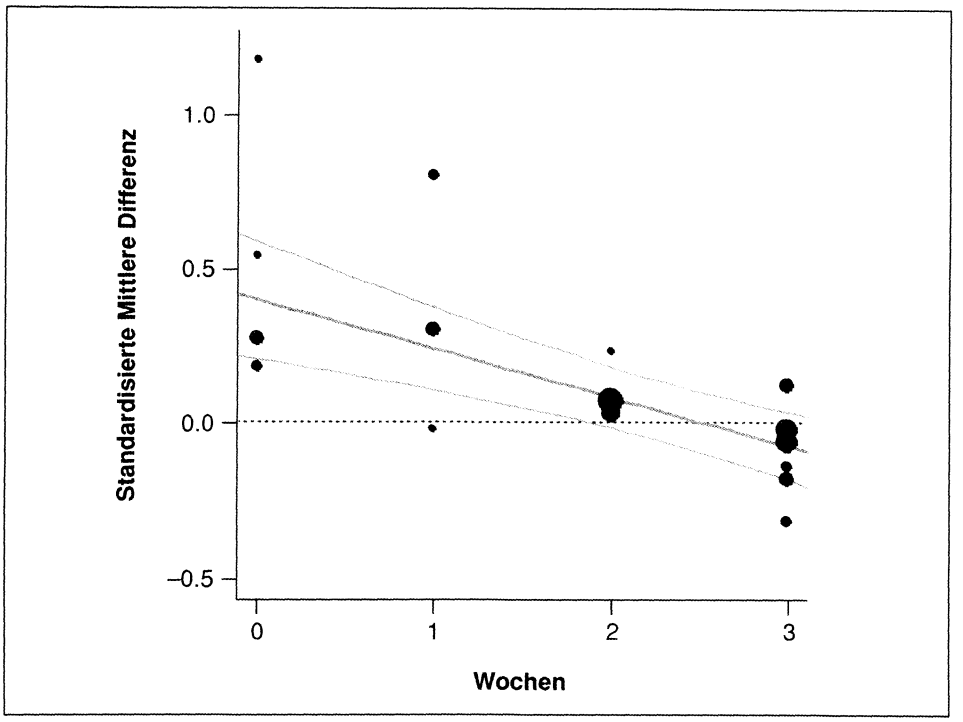

Abbildung 2: Anzahl der Kontaktwochen vor der Erwartungsinduktion und beobachteter Effekt in 19 Studien zum Pygmalioneffekt (die Linien zeigen den geschätzten Erwartungswert der standardisierten mittleren Differenz und das entsprechende 95 %ige Konfidenzintervall; die Punktgröße ist invers proportional zur Stichprobenvarianz und zeigt somit die Gewichtung der Daten bei der Analyse an).

bis 0.180) nun den Wert Null beinhaltet, kann schon nach zwei vorausgehenden Kontaktwochen die Erwartungsinduktion keinen signifikanten Effekt mehr erzielen. Abbildung 2 bietet eine grafische Darstellung der Ergebnisse. Zuletzt sei noch erwähnt, dass ein Vergleich mit den Ergebnissen des Modells zufallsvariabler Effekte andeutet, dass praktisch die gesamte Heterogenität durch die Moderatorvariablen erklärt wird.

7 Fazit

Das Beispiel verdeutlicht, wie mittels einer Meta-Analyse die Integration von Studienergebnissen systematisch durchgeführt werden kann. Vor allem der Einfluss von Drittvariablen lässt sich auf diese Weise gezielt untersuchen, was sich bei rein

narrativen Literaturübersichten als äußerst schwierig erweisen kann. Bei der Durchführung einer Meta-Analyse stößt man allerdings häufig auf praktische Schwierigkeiten, wie zum Beispiel bei fehlenden Informationen über Drittvariablen (z. B. wenn manche Autoren keine Angaben über die Anzahl der vorausgehenden Kontaktwochen machen) und bei statistischer Abhängigkeit, sobald mehrere relevante Ergebnisse von einer Stichprobe berechnet werden können (z. B. wenn zwei verschiedene Intelligenz- oder Begabungstests innerhalb einer Studie benutzt wurden). Zu diesen häufig auftretenden Problemen findet man Lösungsansätze bei Pigott (2001) und bei Kalaian und Raudenbush (1996).

Jede Integrationsstudie (unabhängig von den dabei verwendeten Methoden) muss sich auch der Frage stellen, ob die ermittelten Ergebnisse überhaupt repräsentativ für den wahren Wissensstand innerhalb eines Forschungsgebiets sind. Zum einen kann die Darstellung der Ergebnisse innerhalb einer Studie selektiv ausfallen. Zum anderen stammt meist der Großteil der zu integrierenden Studien aus der publizierten Literatur, und verschiedene Selektionsmechanismen beeinflussen, welche Ergebnisse dort zu finden sind (z. B. eher Studien mit signifikanten Resultaten). Diese Mechanismen können zu einer Verzerrung der Ergebnisse führen. Rothstein, Sutton und Borenstein (2005) beschreiben eingehend dieses Problem, bieten empirische Ergebnisse zu dessen Häufigkeit an und beschreiben statistische Lösungsverfahren.

Literatur

Adair, J. G. & Vohra, N. (2003). The explosion of knowledge, references, and citations. *American Psychologist, 58,* 15–23.

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

Cooper, H. M. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

DerSimonian, R. & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7,* 177–188.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245–260). New York: Russell Sage.

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

Higgins, J. P. T., Thompson, S. G., Deeks, J. J. & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327,* 557–560.

Kalaian, H. A. & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods, 1,* 227–235.

Light, R. J. & Pillemer, D. B. (1984). *Summing up: The science of reviewing research.* Cambridge, MA: Harvard University Press.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L. & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79,* 599–616.

Pigott, T. D. (2001). Missing predictors in models of effect size. *Evaluation and the Health Professions, 24,* 277–307.

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology, 76,* 85–97.

Raudenbush, S. W. (1994). Random effects models. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russell Sage.

Raudenbush, S. W. & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10,* 75–98.

Rosenthal, R. (1994). Parametric measures of effect size. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage.

Rothstein, H. R., Sutton, A. J. & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustments.* Chichester, England: Wiley.

Sidik, K. & Jonkman, J. N. (2007). A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine, 26,* 1964–1981.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics, 30,* 261–293.