Freie Universität Berlin

Diskussionsbeiträge
des Fachbereichs Wirtschaftswissenschaft
der
Freien Universität Berlin

Nr. 2004/8
Betriebswirtschaftliche Reihe

Metaanalyse –
Einführung und kritische Diskussion

Martin Eisend

Institut für Marketing

www.ls-kuss.de

März 2004

ISBN 3-935058-77-2
Freie Universität Berlin
Garystr. 21
D-14195 Berlin

Zusammenfassung

Metaanalyen integrieren empirische Befunde mehrerer Untersuchungen zu einer bestimmten Problemstellung und untersuchen die Variabilität dieser Befunde. Damit helfen sie Wissenschaftlern bei der Informationsintegration und –bewertung im Rahmen ihrer Arbeit und unterstützen Praktiker bei der Entscheidungsfindung. Der vorliegende Beitrag gibt eine erste Einführung in die Methode der Metaanalyse und geht auf problematische Aspekte der Metaanalyse ein und zeigt dabei auf, wie diesen Problemen auf dem heutigen Stand der methodischen Entwicklung der Metaanalyse sinnvoll entgegnet werden kann. Schließlich wird untersucht, welchen eigenständigen Beitrag die Metaanalyse zum Erkenntnisfortschritt einer Disziplin leisten kann.

Summary

Meta-analysis is a means of combining numerical results of multiple studies on a particular research question and also a means of explaining the variability of these results. Meta-analysis assists scientists with the information integration and evaluation and assists practitioners with their decision making. The following article gives a short introduction into the method of meta-analysis, addresses the problems of meta-analysis and points out how these problems can be resolved according to the state-of-the-art of meta-analytic techniques. Finally, the peculiar contribution of meta-analysis to the knowledge development of science is discussed.

Inhalt

1 Einleitung

Gottfried Willhelm Leibniz wird gerne als der letzte Universalgelehrte der Neuzeit bezeichnet. Ihm war es im siebzehnten Jahrhundert noch möglich, einen umfassenden Überblick über den gesamten Wissensbestand der wissenschaftlichen Disziplinen zu erlangen. Seit dieser Zeit, die auch den Ursprung der neuzeitlichen Wissenschaft kennzeichnet, unterliegt die Produktion wissenschaftlichen Wissens einem exponentiellen Wachstum, wie das nicht zuletzt auch die wachsende Zahl wissenschaftlicher Zeitschriften verdeutlicht. Gab es 1750 etwa zehn wissenschaftliche Zeitschriften, so hat sich diese Zahl bis zum Ende des zwanzigsten Jahrhunderts mit großer Exaktheit alle fünfzig Jahre verzehnfacht, während sich vergleichsweise im gleichen Zeitraum die Weltbevölkerung etwa einmal verdoppelt (Price 1976, 1986). Angesichts der zunehmenden Zahl von publizierten wissenschaftlichen Untersuchungen ist es für heutige Wissenschaftlerinnen und Wissenschaftler kaum noch möglich, einen Überblick über alle Forschungsergebnisse selbst in einem klar abgegrenzten Forschungsgebiet zu bekommen und zu behalten. Hinzu kommt, dass zu einer Fragestellung oftmals mehrere Untersuchungen vorliegen, die uneinheitliche und manchmal sogar widersprüchliche Befunde ausweisen. Diese Entwicklung erklärt den zunehmenden Bedarf an Möglichkeiten der Informationsverdichtung und –bewertung von wissenschaftlichen Forschungsergebnissen, der schließlich zur Entwicklung unterschiedlicher Methoden der Ergebniszusammenfassung führte. Neben den traditionellen Formen wie den Reviews haben seit Mitte der siebziger Jahre auch quantitative Ergebniszusammenfassungen, so genannte Metaanalysen, immer mehr an Bedeutung in den verschiedensten Disziplinen mit empirischer Ausrichtung gewonnen. Dabei sind Metaanalysen nicht vorbehaltlos in das Methodenarsenal der Wissenschaft übernommen worden. Vielmehr gab es zunächst eine sehr kritische und zum Teil auch ablehnende Diskussion, aus der sich schließlich eine Reihe von wichtigen Kritikpunkten an der Metaanalyse herauskristallisierten, denen aber durch die Weiterentwicklung der Methode innerhalb der letzten drei Jahrzehnte sehr konstruktiv entgegnet werden konnte.

Der vorliegende Beitrag gibt eine erste Einführung in die Methode der Metaanalyse, geht auf kritische Aspekte der Metaanalyse ein und zeigt auf, welchen eigenen Beitrag die Metaanalyse für den wissenschaftlichen Erkenntnisprozess leisten kann. Da-

zu werden zunächst das Wesen der Metaanalyse und ihre historische Entwicklung kurz gekennzeichnet. Aufbauend auf den prototypischen Ablaufschritten einer Metaanalyse erfolgt dann eine anwendungsorientierte Darstellung der Durchführung einer Metaanalyse. Daran schließt sich eine Darstellung und Diskussion der wichtigsten Pro- und Contraargumente der Metaanalyse an und es wird aufgezeigt, wie auf dem heutigen Entwicklungsstand diesen Problemen sinnvoll entgegnet werden kann. Schließlich wird versucht, die Metaanalyse wissenschaftstheoretisch und methodologisch zu verorten um somit den eigenständigen Beitrag der Metaanalyse für den Erkenntnisfortschritt der Wissenschaft aufzuzeigen.

Im Rahmen des vorliegenden Beitrags ist sicherlich nur eine sehr vereinfachte Darstellung der Durchführung einer Metaanalyse möglich. Daher wird gerade bei der Darstellung des Ablaufs der Metaanalyse auch immer wieder auf weiterführende Literatur verwiesen. Detaillierter finden sich Ausführungen auch in einer Reihe von Lehrbüchern mit zum Teil beträchtlichem Umfang. Interessierte Leserinnen und Leser seien hier auf die klassischen Lehrbücher von Glass et al. (1981), Hedges & Olkin (1985) und Hunter & Schmidt (1990), sowie auf das umfangreiche Herausgeberband von Cooper & Hedges (1994c) verwiesen. Ein Lehrbuch in deutscher Sprache stammt von Fricke & Treinies (1985), eines der aktuellsten Lehrbücher mit hohem Anwendungsbezug ist wohl das Buch von Lipsey & Wilson (2001).

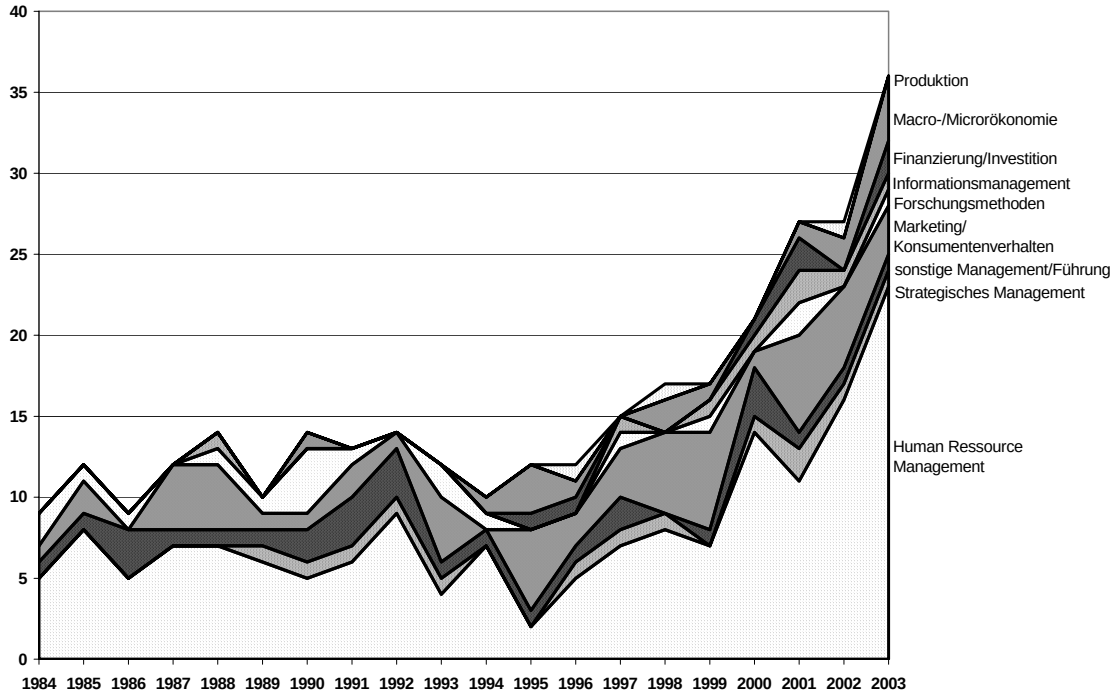
2 Zur Entwicklung und zum Wesen der Metaanalyse

Die erste bekannte quantitative Ergebniszusammenfassung, die nach heutigem Verständnis als Metaanalyse zu bezeichnen ist, wurde vor hundert Jahren von dem britischen Mathematiker Karl Pearson durchgeführt (vgl. Pearson 1904)[1]. Zu dieser Zeit war man sich über den Erfolg von Impfungen gegen Typhus nicht im Klaren, zumal die durchgeführten Untersuchungen – typischerweise bei sehr kleinen Stichproben – zu unterschiedlichen Ergebnissen kamen. Pearsons Idee war nun, die in den Studien ermittelten Korrelationen zwischen Impfung und Todeswahrscheinlichkeit zu mitteln, um so eine Verbesserung der Parameterschätzung auf der Basis einer größeren Stichprobe zu erhalten. Methodische Weiterentwicklungen dieser zunächst sehr simplen Methode der Zusammenfassung etwa durch die Berücksichtigung von Stichprobengrößen oder aber durch die alternative Zusammenfassung von Signifikanzniveaus sind in den folgenden Jahrzehnten vor allem im Bereich der Agrarforschung und der Biostatistik zu beobachten (Hunt 1997, S. 10f.). Ab den fünfziger Jahren finden sich dann auch erste quantitative Ergebnisintegrationen zu Fragestellungen aus der Psychologie und den Erziehungswissenschaften (vgl. Glass et al. 1981, S. 24ff.). Die eigentliche Geburtsstunde der heutigen Metaanalyse wird auf Mitte der siebziger Jahre datiert als Gene V. Glass in einer Ansprache auf der Jahreskonferenz der *American Educational Research Association* die von ihm entwickelte Methode zur quantitativen Ergebnisintegration vorstellt, der er zum ersten Mal die Bezeichnung Metaanalyse verleiht (vgl. Hunt 1997, S. 12). Danach lässt sich eine deutliche Zunahme der Anzahl der durchgeführten quantitativen Ergebniszusammenfassungen ebenso wie die systematische Auseinandersetzung mit metaanalytischen Methoden beobachten, die auch recht schnell in anderen Disziplinen übernommen wurden. Die nachfolgende Grafik verdeutlicht die Ausbreitung der Metaanalyse in der betriebswirtschaftlichen Forschung. Durchsucht wurden dazu in einer Datenbank für betriebswirtschaftliche Literatur (Business Source Elite) an Hand des Suchbegriffs "meta-analy*" (stellvertretend für "meta-analysis", "meta-analyses" oder "meta-analytic") alle Zeitschriftenartikel innerhalb des Veröffentlichungszeitraums von 1984 bis Ende 2003. Aus den Fundstellen der Recherche, die neben Metaanalyen

[1] Weitere Beschreibungen der Geschichte der Metaanalyse finden sich bei Bangert-Drowns (1986), Cooper & Hedges (1994b) und sehr detailliert bei Hunt (1997).

unter anderem auch Arbeiten methodologischer Art oder Kommentare und Repliken zu durchgeführten Metaanalysen beinhalteten, wurden diejenigen Studien ausgewählt, die Metaanalysen im eigentlichen Sinne darstellten, also quantitative Methoden zur Ergebnisintegration angewandt haben. Auf diese Weise wurden insgesamt 316 Metaanalysen ermittelt, die entsprechend ihres Erscheinungsjahres unterschiedlichen betriebswirtschaftlichen Teildisziplinen zugeordnet wurden (vgl. Abb. 1).

Abb. 1: Veröffentlichte Metaanalysen in verschiedenen betriebswirtschaftlichen Teildisziplinen 1984-2003 (n = 316, Datenbankrecherche in Business Source Elite)



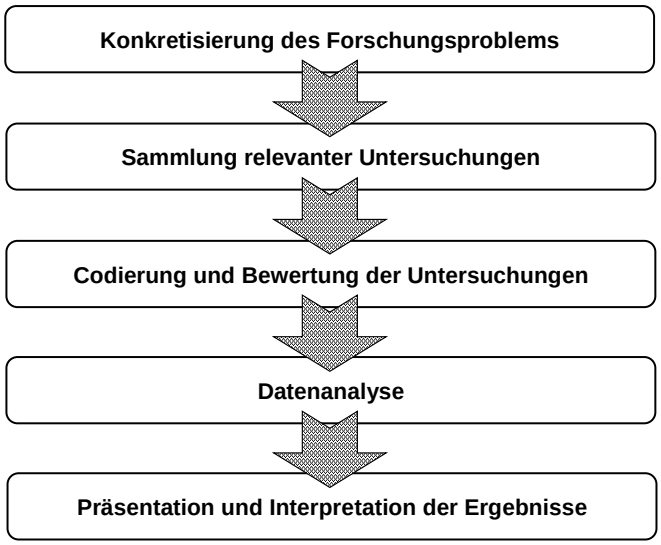
Offensichtlich hat die Anzahl der durchgeführten Metaanalysen in der betriebswirtschaftlichen Forschung vor allem in den letzten zehn Jahren stark zugenommen. Dabei sind es vor allem stark verhaltenswissenschaftlich orientierte und psychologienahe Teildisziplinen (Personal, Management, Marketing), in denen Metaanalysen durchgeführt wurden, was sicherlich auf die Vorreiterrolle der Psychologie bei der Anwendung der Metaanalyse zurückzuführen ist (vgl. auch die Daten bei White 1994).

Was genau ist nun eigentlich unter einer Metaanalyse zu verstehen? Als Glass den Begriff der Metaanalyse einführt, grenzt er die Methode als eine Art Tertiäranalyse von der Primär- und Sekundäranalyse ab: "*Primary analysis* is the original analysis of data in a research study. (…) *Secondary analysis* is the re-analysis of data for the purpose of answering the original research question with better statistical techniques, or answering new questions with old data. (…) *Meta-analysis* refers to the analysis of analyses (…) the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (Glass 1976, S. 3). Die weitere Auseinandersetzung mit und die zunehmende Anwendung von Metaanalysen hat dann im Laufe der Zeit auch zu einer Ausdifferenzierung von unterschiedlichen metaanalytischen Schulen geführt und damit auch zu unterschiedlichen Auffassungen davon, was eine Metaanalyse eigentlich genau ist und welche Intensionen sie verfolgt. Um zu einem übergreifenden Verständnis und einer allgemeingültigen Definition zu kommen, hat Drinkmann daher verschiedene Definitionen ausgewertet (Drinkmann 1990, S. 11). Als Ergebnis dieser Auswertung hebt er die varianzaufklärende Funktion und die quantitative Orientierung als die entscheidenden definitorischen Merkmale hervor und definiert die Metaanalyse als "eine an den Kriterien empirischer Forschung orientierte Methode zur quantitativen Integration der Ergebnisse empirischer Untersuchungen sowie zur Analyse der Variabilität dieser Ergebnisse" (ebd.). Neben die anfangs entscheidende Funktion der Integration von Ergebnissen ist also die (mittlerweile wohl fast wichtigere) Funktion der Erklärung der Variabilität der Ergebnisse getreten. Die Definition zeigt aber nicht nur auf, was eine Metaanalyse ist, sondern gleichzeitig auch, welche Anwendungsgrenzen sie hat (vgl. Lipsey & Wilson 2001, S. 2): Metaanalysen beruhen nämlich immer auf empirischen Untersuchungen und können daher auch keine Integration theoretischer oder konzeptioneller Arbeiten leisten. Sie benötigen außerdem quantitative empirische Ergebnisse, so dass auch Ergebnisse qualitativer Untersuchungsformen wie beispielsweise aus Fallstudien nicht metaanalysierbar sind. Dabei sollten auch tatsächlich nur Ergebnisse von Untersuchungen vorliegen, ist nämlich ein Zugriff auf Originaldaten der Primäruntersuchungen möglich, dann bietet die Anwendung einer Sekundäranalyse weit mehr Auswertungsmöglichkeiten, eine Metaanalyse würde dagegen Informationen "verschenken".

3 Ablaufschritte einer Metaanalyse

Die Durchführung einer Metaanalyse lässt sich vereinfacht auch mit anderen bekannten Untersuchungsformen der empirischen Forschung vergleichen, wie beispielsweise mit der Befragung von Personen, nur dass bei der Metaanalyse eine Studie bzw. die Untersuchungsergebnisse in dieser Studie die Untersuchungsobjekte darstellen und diese Untersuchungsobjekte durch einen Forscher oder eine Forscherin bzw. durch instruierte Codierer hinsichtlich relevanter Eigenschaften "interviewt" werden und die ermittelten Ergebnisse dann anhand statistischer Methoden analysiert werden. Eine metaanalytische Untersuchung operiert also nach einem ähnlichen Prinzip wie Primäruntersuchungen und daher sind auch die Vorgehensweise und der Ablauf einer Metaanalyse mit der Vorgehensweise von Einzeluntersuchungen vergleichbar. Auch hier werden ein Problem formuliert, Daten gesammelt, codiert und bewertet, analysiert und schließlich präsentiert und interpretiert (Cooper & Hedges 1994b; Cooper 1982; Durlak & Lipsey 1991). Abb. 2 verdeutlicht die Vorgehensweise, die auch der folgenden Beschreibung zugrunde liegt.

Abb. 2: Prototypischer Ablauf einer Metaanalyse



3.1 Konkretisierung des Forschungsproblems

Wie jede andere Untersuchung auch beginnt eine Metaanalyse mit der Konkretisierung des Forschungsproblems, der Fragestellung der Metaanalyse also, die umfassend formuliert sein sollte und bereits eine grobe Spezifizierung der zu untersuchenden (abhängigen und unabhängigen) Variablen sowie der relevanten Primärstudien enthalten sollte.

Beispiel:
Brown & Peterson (1993) untersuchen in einer Metaanalyse Einflussfaktoren und Konsequenzen der Arbeitszufriedenheit von Verkäufern (Titel der Veröffentlichung: "Antecedents and Consequences of Salesperson Job Satisfaction"). Grob spezifiziert wird bei dieser Formulierung des Forschungsproblems die zentrale Variable Arbeitszufriedenheit, die sowohl die abhängige als auch die unabhängige Variable in den einzubeziehenden Primärstudien darstellen kann.

3.2 Sammlung relevanter Untersuchungen

Die Sammlung relevanter Untersuchungen (Primärstudien) stellt die Datenerhebung der Metaanalyse dar. Die Grundgesamtheit dieser Studien muss im Hinblick auf die zu untersuchenden Variablen genau abgegrenzt werden, d.h. es muss genau festgelegt werden, welche Beziehung von welchen Variablen in der Primärstudie untersucht werden muss, damit die Studie in die Metaanalyse einbezogen werden kann. Daneben können auch allgemeine Auswahlkriterien wie die Art der einzubeziehenden Publikationstypen (z.B. nur Zeitschriftenartikel oder auch andere Publikationen wie Konferenzbeiträge, Working Papers), der zeitliche Rahmen relevanter Untersuchungen oder der kulturelle und insbesondere auch linguistische Kontext zur Eingrenzung der Grundgesamtheit herangezogen werden (Lipsey & Wilson 2001, S. 16ff).

Nur eine breit angelegte Recherche und Sammlung von Primärstudien innerhalb des festgelegten Rahmens gewährleistet, dass möglichst alle einschlägigen Arbeiten zu einer Fragestellung berücksichtigt werden und somit keine systematische Verzerrung der Untersuchungsergebnisse entsteht. Je nachdem, wo Untersuchungen veröffentlicht werden bzw. ob sie überhaupt veröffentlicht werden, hängt nämlich zum Teil auch von den Ergebnissen selbst ab, ein Problem, auf das später noch im Rahmen der

Kritik des *publication bias* eingegangen wird. Daher ist man in den meisten Metaanalysen um eine relativ gründliche und umfangreiche Recherche relevanter Primärstudien bemüht. Die möglichen Recherchestrategien dabei sind recht vielfältig (vgl. Hunter & Schmidt 1990, S. 490ff.; Rosenthal 1994a). Heutzutage sind neben systematischen Suchen in relevanten Periodika und Bibliographien vor allem begriffliche Suchen in elektronischen Datenbanken gängig. Um auch weitere Arbeiten, die durch diese systematischen Suchen nicht gefunden werden können (insbesondere also (noch) nicht veröffentlichte Untersuchungen bzw. graue Literatur), mit einzubeziehen, können die angegebenen Quellen in den bereits gefundenen Arbeiten durchsucht werden ("Schneeballprinzip") oder Forscherinnen und Forscher direkt kontaktiert werden, die im Rahmen der Recherche als Expertinnen und Experten für die jeweilige Fragestellung identifiziert werden konnten.

Beispiel:
In der Metaanalyse von Brown & Peterson (1993) wurde die Grundgesamtheit anhand der vorab definierten Variable Arbeitszufriedenheit festgelegt, wobei es um die Arbeitszufriedenheit innerhalb der Gruppe von Verkäuferinnen und Verkäufern (ebenfalls begrifflich vorab definiert und eingegrenzt) ging. Es sollten alle empirischen Studien einbezogen werden, die diese Arbeitszufriedenheit als abhängige oder unabhängige Variable untersuchten bzw. allgemein eine Beziehung (Korrelation) von Arbeitszufriedenheit mit anderen Variablen untersuchten. Als allgemeines Auswahlkriterium haben sich die Autoren implizit auch auf einen linguistischen Kontext festgelegt, indem sie nur englischsprachige Untersuchungen einbezogen haben.
Für die Recherche wurden die elektronischen Datenbanken *ABI-Inform* und *PsychLit* nach Begriffen durchsucht sowie acht Zeitschriften und zwei Konferenzbände aus dem Bereich Marketing systematisch durchsucht. Weitere Untersuchungen wurden durch eine anschließende Suche in den Quellenangaben der gefundenen Artikel identifiziert.

3.3 Codierung und Bewertung der Studien

In einem nächsten Schritt erfolgt eine Codierung und Bewertung der gefundenen Untersuchungen in der schließlich auch die Brauchbarkeit einzelner Untersuchungen für die Zwecke der Metaanalyse überprüft wird. Insbesondere werden dabei alle nötigen Informationen codiert, die zur Berechnung eines metaanalysierbaren statistischen Kennwerts erforderlich sind (z.B. bei Experimentaluntersuchungen die Stichprobengrößen, Mittelwerte und Standardabweichungen bei Experimental- und Kon-

trollgruppe). Fehlen nötige Informationen in einer Studie, kann es zum Ausschluss der Studie kommen oder aber man setzt geeignete Schätzverfahren ein, die die fehlenden Informationen ersetzen können (vgl. Hedges 1990; Pigott 1994 und Abschnitt 4.3). Weiterhin werden auch alle Merkmale der Studien codiert, von denen man annimmt, dass sie für die Erklärung der Varianz der Einzelergebnisse relevant sein könnten (Lipsey 1994). Diese Moderatoren können inhaltlicher Art sein, insbesondere also aus theoretischen Vorüberlegungen abgeleitet worden sein, als auch methodischer Art sein (z.B. die Operationalisierung und Messung der untersuchten Variablen). Ein Teil dieser Moderatoren kann anhand von weitgehend objektiven Eigenschaften der Untersuchungen ermittelt werden (so genannte *low-inference codings*, z.B. die Anzahl der Items bei der Messung der untersuchten Variablen), andere Moderatoren unterliegen zum Teil einem erheblichen subjektiven Bewertungsspielraum (*high-inference codings*, z.B. bestimmte Merkmale der Beurteilung der Studienqualität wie etwa die Sorgfalt bei der Datenerhebung) und erfordern eine entsprechende Genauigkeit und Überprüfung bei der Codierung dieser Merkmale (vgl. Cooper 1989, S. 32f.; Orwin 1994). Meist werden daher auch vor allem die *high-inference codings* von mehr als einer Person codiert, anschließend überprüft (meist mittels der so genannten Intercoderreliabilität) und gegebenenfalls abgeglichen. Für die Codierung und Aufbereitung sowie die anschließende Analyse metaanalytischer Daten stehen heutzutage eine Reihe von speziellen Softwarelösungen zur Verfügung, aber auch Standardsoftwarepakete der Statistik (SPSS, SAS) sind zu diesem Zwecke sinnvoll einsetzbar (vgl. die Übersichten bei Normand 1995; Sterne et al. 2001; Sutton et al. 2000).

Beispiel:

Brown & Peterson (1993) identifizieren in ihrer Recherche 89 relevante Studien, schließen aber nur 59 Studien in die endgültige Analyse ein. Unter anderem mussten Studien ausgeschlossen werden, weil sie keine relevanten Informationen zur Berechnung eines metaanalysierbaren Parameters lieferten. Daneben gab es noch eine Reihe weiterer Ausschlussgründe, wie z.B. das Verwenden des gleichen Datensatzes in unterschiedlichen Studien.

Diese 59 Studien wurden anschließend von beiden Autoren unabhängig voneinander hinsichtlich der nötigen Informationen zur Berechnung statistischer Kennwerte und der vorab festgelegten Moderatorvariablen codiert. Moderatorvariablen waren u. a. der Verkäuferty-

pus oder die Art der Messung der Arbeitszufriedenheit. Dabei gab es keine Abweichung zwischen beiden Codierern.

3.4 Datenanalyse

Die Datenanalyse umfasst im Wesentlichen zwei Schritte. Zum einen die Integration der Einzelergebnisse, zum anderen die Untersuchung der Varianz der Einzelergebnisse. Grundsätzlich gibt es neben der Verwendung von Effektstärkenmaßen zur Integration der Einzelergebnisse auch die Möglichkeit der Integration von Signifikanzniveaus (vgl. Becker 1994) oder aber die Möglichkeit, signifikante und nichtsignifikante Ergebnisse einfach auszuzählen (so genanntes *vote counting*, vgl. Bushman 1994). Letztere Methode gilt jedoch als wenig genau, da z.B. die Stichprobengröße einzelner Untersuchungen oder die Größe des Effekts völlig unberücksichtigt bleibt, weshalb sie kaum noch angewandt wird (Hedges & Olkin 1980). Auch die Integration von Signifikanzniveaus liefert keine brauchbare Information über die Größe von Effekten bzw. deren Variabilität und ist somit weniger aussagekräftig als die heute gängige Verwendung von Effektstärken. Gängige Effektstärkenmaße bzw. integrierbare Parameter sind:

- standardisierte Mittelwertsdifferenzen
- Korrelationskoeffizienten
- Elastizitäten (vgl. Farley & Lehmann 1986),
- Verhältnisse oder Differenzen von Wahrscheinlichkeiten für kategoriale Daten (vgl. Fleiss 1994)
- univariate Maßgrößen wie Anteilswerte (vgl. Lipsey & Wilson 2001, S. 38ff.)

Man kann davon ausgehen, dass nicht alle einbezogenen Studien die gleichen Effektstärkenmaße bzw. Parameter ausweisen. Integriert werden können aber nur einheitliche Maße. Lassen sich aufgrund der Informationen in den Untersuchungen die erforderlichen Maße nicht direkt berechnen, so können diese mit Signifikanzniveaus oder mit anderen Effektstärkenmaßen, die in den Studien ausgewiesen werden, an Hand bekannter Berechnungsprozeduren in einheitliche Maße umgerechnet werden (vgl. zu den Berechnungsprozeduren z.B. Glass 1977, S. 35; Glass et al. 1981, S. 149f.; Wolf 1994).

Bei der Integration der nun einheitlichen Effektstärken der Einzelergebnisse T_i (Effektgrößen) kann zur Berücksichtigung der unterschiedlichen Größen der einzelnen Studien ein Gewichtungsfaktor w_i – üblicherweise die Stichprobengröße (vgl. Hunter & Schmidt 1990, S. 100) oder die Inverse der Varianz der einzelnen Effektgröße (vgl. Hedges 1994a) – mit einbezogen werden. Außerdem kann ein weiterer Gewichtungsfaktor q_i mit einbezogen werden, der mögliche Verzerrungen und Artefakte der Messung berücksichtigt, indem er z.B. die Reliabilitäten der gemessenen Konstrukte erfasst (Hunter & Schmidt 1994)[2]. Eine Berechnung des integrierten Wertes $\overline{T}$ erfolgt dann mittels der allgemeinen Formel (Hedges 1994a; Shadish & Haddock 1994):

$$(1) \qquad \overline{T} = \frac{\sum_{i=1}^{k} q_i w_i T_i}{\sum_{i=1}^{k} q_i w_i}$$

Die dazu gehörige gewichtete Varianz berechnet sich an Hand der folgenden Formel:

$$(2) \qquad s_{\overline{T}}^2 = \frac{\sum_{i=1}^{k} q_i^2 w_i}{(\sum_{i=1}^{k} q_i w_i)^2}$$

Um zu testen, ob der integrierte Wert tatsächlich von Null verschieden ist (Test auf Populationsnulleffekt), kann unter der Annahme der Normalverteilung die dem α-Niveau entsprechende Variationsbreite des Populationseffekts bestimmt werden (Shadish & Haddock 1994). Ein Konfidenzintervall wird dann mittels folgender Formel berechnet:

$$(3) \qquad \overline{T} - z_{\alpha/2} s_{\overline{T}} \leq \tau \leq \overline{T} + z_{\alpha/2} s_{\overline{T}}$$

Umschließt das Konfidenzintervall den Wert Null, ist davon auszugehen, dass der Mittelwert der Effektgrößen nicht signifikant von Null verschieden ist (Wolf 1994, S. 27). Wird dabei eine Varianz, die Messfehler und Artefakte berücksichtigt, verwendet, spricht man auch vom *credibility interval*, das bei entsprechender Größe auch als

[2] Eine ausführliche Beschreibung von weiteren Artefakten in Untersuchungen, die bei der Integration berücksichtigt werden können, findet sich bei Hunter & Schmidt (1990, S. 44ff.).

ein Indikator für die Existenz von Moderatorvariablen interpretiert werden kann (vgl. Arthur et al. 2001, S. 87ff.). Als äquivalenter Test auf einen Populationsnulleffekt kann auch der *z*-Wert der Standardnormalverteilung berechnet werden, der bei einer Irrtumswahrscheinlichkeit von weniger als 5% größer als der Wert 1,96 sein muss um auf einen vorhandenen Populationseffekt schließen zu können (Shadish & Haddock 1994):

$$z = \frac{\overline{T}}{s_{\overline{T}}} \tag{4}$$

Für signifikante Ergebnisse der Metaanalyse kann außerdem der *fail-safe N* berechnet werden, ein Wert, der angibt, wie viele Effektgrößen mit einem mittleren Effekt der Größe Null noch vorhanden sein müssten, damit der Gesamttest nicht signifikant wird. Dazu geht Rosenthal (1979) von der Aufaddierung der dem Signifikanzniveau der vorhandenen *k* Effektgrößen entsprechenden *z*-Werte der Standardnormalverteilung aus, so dass sich ein integrierter Wert *Z* ergibt aus:

$$Z = \frac{k\bar{z}}{\sqrt{k}} = \sqrt{k}\bar{z} \quad \text{mit} \quad \bar{z} = \frac{\sum_{i=1}^{k} z_i}{k} \tag{5}$$

Für *Z* setzt man nun 1,645 ein, wenn man für ein 5%-Signifikanzniveau einseitig testen möchte und für *k* setzt man *k* + *X*, wobei *X* die gesuchte Anzahl zusätzlicher Effektgrößen mit Nulleffekt ist. Nach entsprechender Umformung ergibt sich für *X*:

$$X = \frac{k}{2{,}706}(k\bar{z}^2 - 2{,}706) \tag{6}$$

Als Richtwert für ein Toleranzlevel schlägt Rosenthal (1984, S. 110) einen Wert $X \geq 5k + 10$ vor. Dieser Wert stellt eine Möglichkeit zur Untersuchung der Wahrscheinlichkeit eines *publication bias* dar, eine Problematik, auf die im nächsten Kapitel noch genauer eingegangen wird.

Nur bei Homogenität der einbezogenen Effektgrößen stellt der integrierte Wert einen akzeptablen Schätzer des wahren Populationseffekts dar. Um zu prüfen, ob von einer tatsächlichen Variation (Heterogenität) zwischen den einzelnen Effektgrößen auszugehen ist, wird zunächst die Varianz berechnet, die sich aufgrund des Stichproben-

fehlers der einzelnen Untersuchungen ergibt. Ist diese stichprobenbedingte Varianz zum Großteil für die Gesamtvarianz der Effektgrößen verantwortlich, so kann von einer Homogenität der integrierten Effektstärken ausgegangen werden. Hier kann die 75%-Regel angewandt werden, die davon ausgeht, dass mindestens 75% der Gesamtvarianz durch die stichprobenbedingte Varianz erklärt sein sollten (vgl. Hunter & Schmidt 1990, S. 414). Daneben kommen heute meist Testverfahren zur Überprüfung der Homogenität der Effektgrößen zum Einsatz, die im Prinzip aber ebenfalls auf dem beschriebenen Varianzvergleich beruhen. Ein sehr gängiger Homogenitätstest, der einer χ^2-Verteilung mit $k-1$ Freiheitsgraden folgt, berechnet sich folgendermaßen (Hedges 1994a; Shadish & Haddock 1994):

$$(7) \qquad Q = \sum_{i=1}^{k} \frac{(T_i - \overline{T})^2}{s_T^2} \quad \text{bzw.} \quad Q = \sum_{i=1}^{k} w_i T_i^2 - \frac{(\sum_{i=1}^{k} w_i T_i)^2}{\sum_{i=1}^{k} w_i}$$

Das Gewicht w entspricht dem Gewichtungsfaktor, der bereits bei der Integration der Effektstärken berücksichtigt wird (vgl. Formel (1)).

Bei Vorliegen von Heterogenität sollte die Varianz zwischen den Effektgrößen mit inhaltlichen und methodischen Moderatorvariablen untersucht werden, anhand derer alle Effektgrößen in Subgruppen unterteilt werden können, die – falls sie bedeutsam sind – zu einer geringeren Varianz in den Gruppen im Vergleich zur Gesamtvarianz aller Effektgrößen führen sollten. Die Erklärungskraft einer kategorialen Moderatorvariable (z.B. Geschlecht der Untersuchungspersonen) kann dann beispielsweise durch eine Varianzanalyse berechnet werden mit dem Ziel, die Varianz der Effektstärken durch die auf Grund der Moderatorvariable gebildeten Gruppen (im Beispiel eben die Gruppe weiblicher und die Gruppe männlicher Untersuchungspersonen) entsprechend zu reduzieren (vgl. Hedges 1982b). Für mehrere Moderatorvariablen kann auch eine Regressionsanalyse berechnet werden, bei der die Effektstärken T_i die abhängige Variable, die Moderatoren methodischen Ursprungs M und inhaltlichen Ursprungs C die unabhängige Variablen darstellen (vgl. Hedges 1994a; Lipsey & Wilson 2001, S. 122f.).

$$(8) \qquad T_i = \beta + \sum_{k=1}^{m} \beta_k M_k + \sum_{l=1}^{m} \beta_l C_l + u_i$$

Zur Berücksichtigung der unterschiedlichen Datengrundlage der einzelnen Effekte wird dazu üblicherweise eine *weighted least squares regression* (WLS) durchgeführt, bei der die Effektgrößen mit einem Gewicht w, nämlich der Inverse ihrer Varianz gewichtet werden (Hedges 1994a). Der Einfluss und die Einflussrichtung einzelner Moderatorvariablen kann dann an Hand der unstandardisierten Regressionskoeffizienten überprüft werden (Hedges & Olkin 1985, S. 238ff.). Die gesamte Erklärungskraft des Models kann anhand von Q_R, der erklärten Streuung des Modells, überprüft werden, ein Wert, der einer χ^2-Verteilung mit p Freiheitsgraden folgt, wobei p der Anzahl der erklärenden Variablen (Moderatoren) entspricht. Die unerklärte Varianz Q_E diente als Homogenitätstest. Der Wert folgt ebenfalls einer χ^2-Verteilung mit $k - p$ - 1 Freiheitsgraden. Ist Q_E signifikant, dann ist von einer Varianz auszugehen, die über die reine stichprobenfehlerbedingte Varianz hinausgeht, d.h. es besteht weiterhin Heterogenität.

Ziel der Moderatoranalysen ist es, die Heterogenität der Effektgrößen möglichst vollständig aufzuklären. Gelingt dies nicht, so kann von einer Unterspezifizierung des Moderatormodells ausgegangen werden und es sollte nach weiteren potenziellen Moderatoren gesucht werden. Eine weitere Möglichkeit um zu homogenen Effektstärkengruppen zu gelangen, ist der systematische Ausschluss von Ausreißern in den Effektgrößen (vgl. Arthur et al. 2001, S. 117ff.; Hedges & Olkin 1985, S. 256). Es kann aber auch die Annahme unterstellt werden, dass die Varianz der Effektgrößen zum einen bedingt ist durch die Stichprobenvarianz der einzelnen Studien und zum anderen durch eine Varianz der einzelnen Effektgrößen selbst, die womöglich auch nur eine Stichprobe aus allen potenziell möglichen Untersuchungen darstellen, wobei die Quelle dieser Varianz aber nicht etwa durch Moderatoren identifiziert werden kann. Die Anwendung eines entsprechenden *random effect model* sollte aber auf konzeptionellen Überlegungen über die unterstellte Grundgesamtheit, über die man die Studienergebnisse generalisieren möchte, beruhen und entsprechend begründbar sein (vgl. Hedges 1994b; Hedges & Vevea 1998).

Beispiel:
Brown & Peterson (1993) rechnen alle 254 Effektstärkenmaße in den von ihnen als relevant identifizierten Studien in den Produkt-Moment-Korrelationskoeffizienten nach Pearson um. Sie identifizieren weiterhin 28 Beziehungen zwischen der Arbeitszufriedenheit des Verkaufspersonals und jeweils einer anderen Variable und integrieren für jede dieser

Beziehung die Korrelationskoeffizienten, wobei sie einmal eine Varianzgewichtung und einmal eine Gewichtung an Hand der Varianz und der Reliabilitätskoeffizienten vornehmen. Beispielsweise ermitteln sie für den Zusammenhang zwischen der Arbeitszufriedenheit und der Neigung zu kündigen 19 Effektgrößen, die auf einer Gesamtstichprobe von 3992 Personen beruhen. Der integrierte Korrelationskoeffizient wird mit -0,36 (varianzgewichtet) bzw. -0,46 (varianz- und reliabilitätsgewichtet) ausgewiesen. Das 95%-Konfidenzintervall umfasst den Wertebereich von -0,54 bis -0,18 (varianzgewichtet) bzw. von -0,66 bis -0,26 (varianz- und reliabilitätsgewichtet). Da das Intervall die Null nicht umschließt, kann von einem von Null verschiedenen negativen Populationseffekt für die Beziehung zwischen Arbeitszufriedenheit und der Neigung zu kündigen ausgegangen werden.

Für alle 28 untersuchten Variablenbeziehungen führen die Autoren anschließend Homogenitätstests durch. Bei einem Ergebnis, das auf Heterogenität verwies, eliminierten die Autoren zunächst jeweils einen Ausreißer. Es verblieben noch elf Beziehungen, die auf Heterogenität verweisen. Sie wählten drei dieser Beziehungen aus, nämlich diejenigen, die auf mindestens zehn Effektgrößen beruhten, und versuchten die Heterogenität durch Moderatoren zu erklären. So konnten sie u.a. zeigen, dass der Zusammenhang zwischen Rollenkonflikt und Arbeitszufriedenheit bei Verkaufspersonal im Business-to-Business-Sektor signifikant größer war als im Konsumgütersektor. Insgesamt wurden vier Moderatorvariablen herangezogen, die im Rahmen einer Regressionsanalyse die Heterogenität der drei Effektgrößengruppen aufklären sollten. Bis auf eine Gruppe gelang dies auch. Für diese heterogene Gruppe unterstellten die Autoren ein unterspezifiziertes Modell, also den Bedarf nach Einbeziehung von weiteren Moderatorvariablen.

In den letzten Jahren werden metaanalytisch ermittelte Daten auch häufiger mit weiterführenden statistischen Verfahren ausgewertet, insbesondere kausalanalytische Verfahren kommen dabei zum Einsatz (vgl. Becker & Schram 1994; Cheung 2002; Shadish 1996). Die Datengrundlage für diese Analysen stellt eine metaanalytisch synthetisierte Korrelations- bzw. Kovarianzmatrix dar. Auf der Basis dieser Daten können dann Pfadanalysen oder Kausalanalysen berechnet werden. Bisher sind diese Verfahren jedoch noch mit einigen Problemen behaftet, etwa mit der Frage nach der Fallzahl, die der Analyse zugrunde gelegt werden soll wenn die Korrelationen aus unterschiedlichen Studien mit unterschiedlicher Fallzahl stammen. Auch die Vollständigkeit der Korrelationsmatrix kann Schwierigkeiten bereiten, wenn beispielsweise zu einer bestimmten Variablenbeziehung keine statistische Information in den in die Metaanalyse einbezogenen Studien zu finden ist. Aber auch die Anforderun-

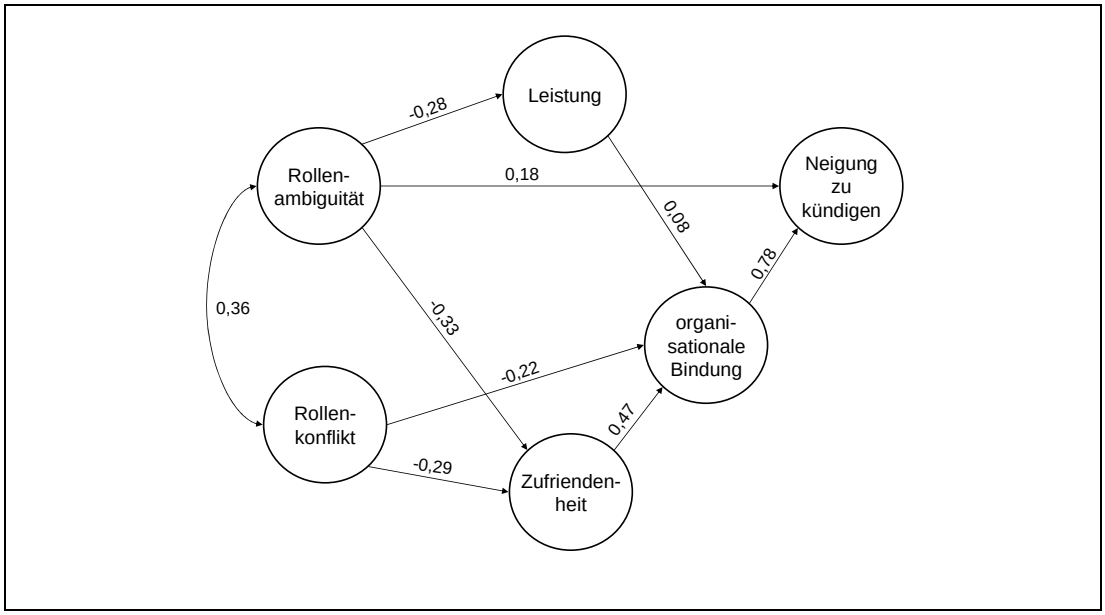
gen an die Kausalität der Beziehungen stellt möglicherweise ein Problem dar, auf das später noch eingegangen wird (vgl. Abschnitt 5).

Beispiel:

Brown & Peterson (1993) untersuchten in ihrer Metaanalyse ein vorab auf der Basis vorhandener Literatur aufgestelltes Modell über die Wirkungszusammenhänge zwischen Rollenambiguität, Rollenkonflikt, Leistung, Zufriedenheit, organisationaler Bindung und der Neigung zu kündigen. Dazu wurden die Korrelationskoeffizienten der relevanten Beziehungen in einer Korrelationsmatrix zusammengestellt. Die folgende Dreiecksmatrix weist für jede Beziehung die Korrelationskoeffizienten, darunter die Anzahl der Studien, auf denen die Korrelationen beruhen und darunter die kumulierte Stichprobengröße dieser Studien aus.

	Rollenamb.	Rollenk.	Leistung	Zufriedenh.	org. Bindung	Neig. zu K.
Rollenambiguität						
Rollenkonflikt	0,28 9 1245					
Leistung	-0,24 7 1204	-0,07 8 1251				
Zufriedenheit	-0,36 15 2431	-0,33 17 2641	0,13 29 7621			
organisationale. Bindung	-0,28 6 654	-0,34 7 915	0,15 7 863	0,50 11 1587		
Neigung zu kündigen	0,36 4 414	0,28 3 357	-0,12 9 1571	-0,36 19 3992	-0,70 4 423	

Für die Kausalanalyse legten die Autoren den Median der kumulierten Stichprobengrößen zugrunde (N=1251). Auf dieser Basis wurde dann das ursprünglich unterstellte Modell kausalanalytisch mittels des Softwarepakets LISREL geschätzt. Aufgrund der schlechten Modellanpassung wurde das Modell in mehreren Schritten modifiziert, z.B. wurden nicht signifikante Beziehungen ausgelassen. Das endgültige Modell erreichte mit einem GFI von 0,991 und einem RMSR von 0,021 eine sehr gute Modellanpassung. Die folgende Abbildung gibt die standardisierten Koeffizienten, die alle auf signifikante Beziehungen verweisen, wieder. Auch die in der Literatur oftmals strittige Richtung der Beziehung zwischen Zufriedenheit und organisationaler Bindung konnten die Autoren über ein entsprechend erweitertes Modell überprüfen, in dem sie eine reziproke Beziehung zwischen diese beiden Konstrukten unterstellten und die sich ergebenden Koeffizienten verglichen. Dabei ergab sich eine eindeutige Bevorzugung einer Beziehung, die davon ausgeht, dass die Arbeitszufriedenheit die organisationale Bindung beeinflusst und nicht umgekehrt.



3.5 Präsentation und Interpretation der Ergebnisse

Bei der Präsentation und Interpretation der Ergebnisse schließlich sollten die methodischen Schritte beschrieben werden, die Ergebnisse dargestellt und zusammengefasst werden, Implikationen für Theorie und Praxis aufgezeigt werden, insbesondere auch auf mögliche Forschungslücken sowie Ansätze für weitere Forschungsmöglichkeiten verwiesen werden und problematische Aspekte der Untersuchungsmethode diskutiert werden (vgl. Halvorsen 1994). Die Darstellung der einzelnen Studien mit relevanten Studienmerkmalen sowie der Ergebnisse selbst erfolgt meist in tabellarischer Form. Daneben gibt es auch eine Reihe von grafischen Darstellungsmöglichkeiten (vgl. Light et al. 1994). Für die Darstellung der Einzelergebnisse und deren Varianz sind neben *funnel plots* (vgl. Abschnitt 4.3, insbesondere Abb. 3) vor allem *stem-and-leaf-plots* gängig, sollen neben der Verteilung auch die integrierten Werte dargestellt werden, eignen sich *box-plots*.

Beispiel:

Brown & Peterson (1993) weisen für die Ergebnisse der Integration der Effektstärken sowie der Moderatorvariablen als auch für die Kausalanalyse jeweils Tabellen aus, die im Text entsprechen interpretiert werden.

Die nachfolgende Abbildung stellt ein *stem-and-leaf-plot* über die 83 Effektgrößen in Form von Korrelationskoeffizienten einer (fiktiven) Metaanalyse dar. Die Korrelationen sind alle positiv, die kleinste Korrelation beträgt 0,02, die größte 0,82, so dass von einer relativ großen Streuung der Effektgrößen auszugehen ist.

Stem	*Leaf*	*n*
0,8	2	1
0,7	017	3
0,6	114688	6
0,5	0224557	7
0,4	0233455667789	13
0,3	001233455566778889	18
0,2	00011234445566677889	20
0,1	0124556689	10
0,0	2589	4

4 Die "Metaanalyse-Diskussion": Argumente für und wider die Metaanalyse

Sinn und Wert der Metaanalyse sind insbesondere seit den siebziger Jahren ständig und intensiv diskutiert worden, wobei einer ersten Phase der Propagierung der Methode eine Phase zum Teil sehr scharfer Kritik folgte, die wiederum von einer Phase der Konsolidierung, Etablierung und Grenzziehung der Methode abgelöst wurde (Drinkmann 1990, S. 17f.). Was genau spricht nun eigentlich vor dem Hintergrund des aktuellen Stands der methodischen Entwicklung der Metaanalyse für ihre Anwendung, was spricht dagegen?

Anbetrachts des Ziels der Metaanalyse, einen integrativen Überblick über vorhandene Forschung zu geben, ergibt sich der wesentliche Vorzug der Metaanalyse gegenüber herkömmlichen Reviews, die ja das gleiche Ziel verfolgen, vor allem aus der quantitativen Orientierung (Beaman 1991). Diese Quantifizierung bei der Integration von Untersuchungen und deren Ergebnissen ermöglicht auch bei nicht übereinstimmenden oder widersprüchlichen Partialbefunden zu einem eindeutigen Gesamtergebnis zu gelangen und die Unterschiedlichkeit der Partialbefunde anhand von Moderatorvariablen zu erklären. Da metaanalytische Befunde zudem auf einer umfangreicheren Fallzahl als die Befunde aus Einzelstudien beruhen, ist ihr Bewährungsgrad auch Entscheidungsträgern außerhalb der Wissenschaft unmittelbar einleuchtend, weshalb Metaanalysen für eine Reihe von Entscheidungsträgern z.B. im medizinischen Bereich bereits heute eine große Rolle spielen (vgl. Mann 1990, 1994). Als empirisch orientierte Methode zeichnet sich die Metaanalyse auch durch Replizierbarkeit und Objektivität (approximiert durch Intersubjektivität) der Analyse in all ihren Einzelheiten und Analyseschritten aus, womit auch die Kriterien der Explizierbarkeit, Systematik, Standardisiertheit und Quantifizierbarkeit abgedeckt werden (Drinkmann 1990, S. 22). Durchführung und Ergebnisse lassen sich daher auch auf Seiten des Lesers bei entsprechender Methodenkenntnis sehr gut nachvollziehen (Cooper & Rosenthal 1980). Aus dieser empirischen Orientierung folgt auch die Handhabbarkeit und Ökonomie der Methode, denn einzelne Arbeitsschritte lassen sich leicht partitionieren und delegieren.

Fast durchgängig werden in der Literatur zur Metaanalyse die folgenden vier Punkte als die wichtigsten Kritikpunkte oder Gegenargumente zur Metaanalyse aufgeführt (vgl. z.B. Beelmann & Bliesener 1994; Drinkmann 1990, S. 20ff.; Glass et al. 1981, S. 217ff.; Hunter & Schmidt 1990, S. 506ff.; Plath 1992, S. 16ff.; Wolf 1994, S. 14ff.):

- Durch Metaanalysen werden nicht vergleichbare Untersuchungen integriert (*apples and oranges*- bzw. "Äpfel und Birnen"-Argument, Uniformitätsproblem).
- Methodisch gute und schlechte Arbeiten werden bei der Integration nicht unterschieden (*garbage in - garbage out*).
- Die Selektivität von Wissenschaftlern und Herausgebern repräsentiert nicht den wahren Forschungsstand, vielmehr werden überwiegend nur signifikante Ergebnisse in den Integrationsprozess einbezogen (*publication bias*, *file drawer problem*).
- Abhängige Daten werden in die Metaanalyse mit einbezogen und wie unabhängige Daten behandelt (Problem der *nonindependent effects* oder *multiple effect sizes*).

Auf dem heutigen Stand der metaanalytischen Diskussion können diese Probleme als weitgehend ausdiskutiert betrachtet werden und es finden sich durchwegs Vorschläge, wie diese Probleme sinnvoll zu lösen sind. Auf die einzelnen Probleme und die jeweiligen Problembewältigungsstrategien wird nachfolgend eingegangen.

4.1 "Apples and Oranges" – Das Uniformitätsproblem

Wenngleich Metaanalysen Untersuchungsergebnisse zu einem Forschungsproblem integrieren, so hat man es dabei doch nicht immer mit identischen Replikationsstudien zu tun, vielmehr werden Studien vermengt, die sich z.B. durch Operationalisierungen, Eigenschaften von Untersuchungspersonen in den Stichproben oder durch Auswertungsmethoden unterscheiden, wodurch womöglich ein Problem der Vergleichbarkeit entsteht. Zur Handhabung dieses Uniformitätsproblems lassen sich in der wissenschaftlichen Diskussion zwei verschiedene Positionen identifizieren (vgl. Hunter 2001). Auf der einen Seite gibt es die vor allem in naturwissenschaftlichen Disziplinen zu findenden Befürworter einer recht strikten Herangehensweise, die nur

so genannte perfekte Replikationen für metaanalysierbar halten, also ausschließlich Untersuchungen, die die gleichen Variablenbeziehungen mit jeweils gleichen Messmethoden untersuchen (vgl. Lipsey & Wilson 2001, S. 9f.). Dem halten Glass et al. (1981, S. 218ff.) aber entgegen, dass die Forderung nach einer Integration von in allen Aspekten gleichen Arbeiten ja eigentlich sinnlos sei, da gleiche Arbeiten bis auf den Stichprobenfehler auch gleiche Ergebnisse bringen würden und von daher eine Ergebniszusammenfassung wenig neue Informationen liefern würde. Eine Generalisierbarkeit von Ergebnissen über Studien mit sich unterscheidenden Untersuchungsmerkmalen ist damit natürlich auch nicht möglich. Man erkauft sich die methodische Strenge aber auch durch eine Reduzierung der Anzahl der integrierbaren Ergebnisse, da perfekte Replikationen in der Wissenschaft aufgrund mangelnder Beachtung und Anerkennung eher selten durchgeführt werden. Dies trifft gerade auf die verhaltenswissenschaftlichen oder sozialwissenschaftlichen Disziplinen zu, in einigen Naturwissenschaften, insbesondere in der Medizin, sind dagegen perfekte Replikationen noch eher gängig. Daher ist gerade in verhaltenswissenschaftlich und sozialwissenschaftlich orientierten Disziplinen eine andere Position bezüglich des Uniformitätsproblems vorzufinden, die von der Einbeziehung imperfekter Replikationen mit inhaltlichen und methodischen Unterschieden zwischen den einzelnen Studien ausgeht und die versucht, die dadurch möglicherweise auftretende Heterogenität der Ergebnisse an Hand von Moderatorvariablen zu erklären (Farley et al. 1981; Greenland 1994). Tatsächlich steht heute im Gegensatz zu früheren Metaanalysen auch weniger die Aggregation und Integration von Effektstärken im Vordergrund als vielmehr die Analyse der Varianz dieser Effektstärken (Lipsey & Wilson 2001, S. 8f.). Die Anwendung der Metaanalyse ist bei dieser Position nicht nur für Replikationsstudien mit einigen methodischen Unterschieden möglich, sondern alle Studien, die im Sinne der Fragestellung des Forschers als inhaltlich und methodisch homogen angesehen werden können, können auch analysiert werden. Smith et al. (1980, S. 47) formulieren dies sehr pointiert in ihrer häufig zitierten Aussage: "Indeed the approach does mix apples and oranges, as one necessarily would do in studying fruits."

4.2 "Garbage in – Garbage out" – Die Integration von Studien unterschiedlicher Qualität

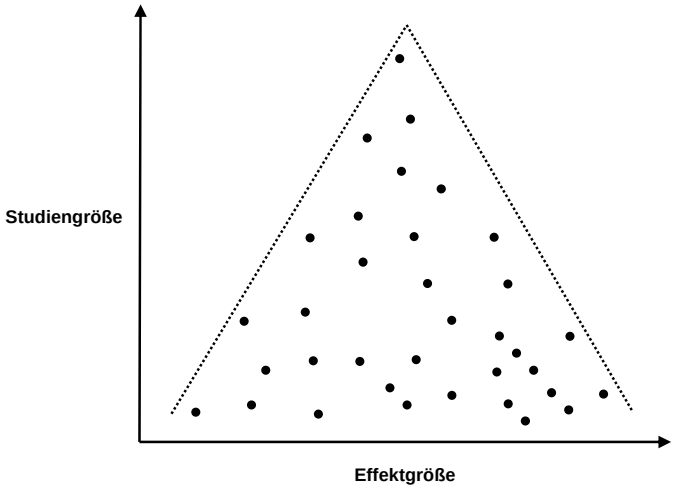
Der zweite Kritikpunkt bezieht sich auf die Unterschiede bezüglich der methodischen Qualität der zu integrierenden Studien. Die Problematik wird vor allem dadurch begründet, da von einem Zusammenhang zwischen methodischer Qualität und dem Ergebnis einer Studie auszugehen ist: je höher die methodische Qualität der Studie, desto stärker fallen die Effekte aus (Fricke & Treinies 1985, S. 171). Grundsätzlich gibt es zur Lösung des Problems natürlich die Möglichkeit, Studien von minderer Qualität auszuschließen. Daneben kann die Studienqualität aber auch bei der Integration der Effektstärken als ein Gewichtungsfaktor berücksichtigt werden oder aber man bezieht die Studienqualität als eine Moderatorvariable zur Erklärung der Heterogenität der integrierten Ergebnisse mit ein. Ein a priori Ausschluss von Studien minderer Qualität, wie er bei einer sehr stringenten Herangehensweisen der Metaanalyse gefordert wird (z.B. Slavin 1986), geht dabei allerdings auch immer mit einem Informationsverlust einher, da ja die Gesamtheit der möglicherweise einzubeziehenden Studien entsprechend verringert wird (Glass et al. 1981, S. 220ff.). Die Berücksichtigung der Studienqualität durch Qualitätskorrekturen erfolgt beispielsweise durch die Einbeziehung von Messfehlerkorrekturen etwa in der Form von Reliabilitätskoeffizienten bei der Ergebnisintegration (vgl. dazu insbesondere Formel (1) und (2) in Abschnitt 0). Werden Messartefakte bei der Integration der Effektstärken berücksichtigt, spricht man auch von einer *full metaanalysis* im Gegensatz zur *bare bone metaanalysis*, die eben keine entsprechende Fehlerkorrekturen berücksichtigt (Hunter 2001). Schließlich kann die Studienqualität auch als Moderatorvariable einbezogen werden um damit die Varianz der Studienergebnisse zu erklären. Zur Beurteilung der Studienqualität sind mittlerweile recht umfangreiche Codierungsschemata entwickelt worden, die die problematische Subjektivität bei der Beurteilung bestimmter Merkmale der Studienqualität (z.B. die Sorgfalt bei der Datenerhebung oder die Qualität der Datenauswertung) einzuschränken versuchen (Wortman 1994). Allerdings verbleiben ein gewisser Ermessensspielraum bei der Qualitätsbeurteilung und damit auch die Gefahr von nicht kontrollierbaren Einflüssen auf die Ergebnisvarianz (Jackson 1980). Dennoch bietet die Einbeziehung der Studienqualität gerade in Form einer Moderatorvariable die Möglichkeit, auch die unterschiedliche Studienqualität zur metaanalytisch verwertbaren Information aufzuarbeiten, die mög-

licherweise eine Erklärung für die Variabilität der vorgefundenen Effektstärken liefern kann, so dass der Prozess *garbage in – garbage out* umgeformt wird in einen Prozess des *garbage in – information out*.

4.3 "Publication Bias" – Die Verzerrung zugunsten signifikanter Ergebnisse

Das in der Kritik des *publication bias* formulierte Dunkelziffer-Argument bezieht sich auf die Selektionsmechanismen im Forschungs- und Publikationsprozess, wodurch die Publikation signifikanter Ergebnisse gefördert wird, während nicht signifikante Ergebnisse meist unveröffentlicht in der Schublade der Forscherinnen und Forscher verbleiben (daher auch die Bezeichnung als *file drawer problem*). Neben den publizierten Untersuchungen existiert also vermutlich eine gewisse Dunkelziffer an nicht zugänglichen Untersuchungen mit vermutlich eher nicht signifikanten Ergebnissen (Wolf 1994, S. 37f.). Diesem Problem wird heutzutage mit einer recht umfangreichen Recherchestrategie begegnet, die möglichst auch unveröffentlichte Untersuchungen erfassen sollte, so wie bereits in Abschnitt 3.2 beschrieben. Eine einfache Möglichkeit zur Überprüfung des Vorhandenseins eines *publication bias* stellen Trichter-Grafiken (*funnel-graphs*) dar. Dabei handelt es sich um einfache Streudiagramme, wobei die gefundenen Ergebnisse gegen die dazugehörige Studiengröße angetragen werden (Light & Pillemer 1984, S. 63ff). Idealiter sollte sich im Diagramm die Form eines umgekehrten Trichters ergeben, da die Ergebnisse bei kleineren Studien aufgrund der Zufallsschwankungen breiter streuen als bei größeren Studien (vgl. Abb. 3). Gibt es einen *publication bias* und es fehlen tatsächlich nicht signifikante Ergebnisse, dann ist der Trichter nicht vollständig und im Streudiagramm besteht eine Lücke bei kleinen Effektgrößen. Natürlich kann die Überprüfung auch rechnerisch bzw. analytisch erfolgen, indem eine entsprechende Verteilung unterstellt wird und die ermittelten Werte auf ihre Anpassung an diese Verteilung überprüft werden (zur *weighted distribution theory* vgl. Begg 1994; vgl. auch Rust et al. 1990).

Abb. 3: Trichter-Grafik (*funnel graph*) zur Überprüfung des *publication bias*



Der *publication bias* bzw. das *file drawer problem* führt schließlich auch zur Vermutung, dass die vorhandenen signifikanten Ergebnisse nur zufällig zustande gekommen seien, in der Wirklichkeit aber die Gültigkeit der Nullhypothese zu erwarten sei, da ja eine ganze Reihe an Untersuchungen mit nicht signifikanten Ergebnissen ja gar nicht aufzufinden waren. Dieses Problem verliert an Bedeutung, je mehr Untersuchungen nötig sind, um die aufgrund der vorhandenen Untersuchungen gezogenen Rückschlüsse zu widerlegen. Dazu hat Rosenthal (1979) den Kennwert *fail-safe N* entwickelt, ein Wert, der angibt, wie groß die Zahl der noch nicht entdeckten, nicht signifikanten Ergebnisse sein müsste, um die Zahl der entdeckten signifikanten Ergebnisse als Zufallsfehler deklarieren zu können (vgl. zur Berechnung Abschnitt 0). Green & Hall (1984) berichten beispielsweise von einer Metaanalyse, bei der 345 Untersuchungen integriert wurden und dazu ein *fail-safe N* von 65000 berechnet wurde, so dass erst ab einer zusätzlichen Zahl von 65000 nicht signifikanten Ergebnissen die Signifikanz des integrierten Ergebnisses in Frage zu stellen wäre. Dabei stellt der *fail-safe N* aber keine Teststatistik dar, sondern liefert lediglich einen interpretationsbedürftigen Richtwert.

In Zusammenhang mit dem Problem des *publication bias* steht auch das Problem der *missing data*, dem Fehlen nötiger Informationen in einzelnen Untersuchungen also, insbesondere von statistischen Kennwerten, die für eine metaanalytische Weiterverarbeitung notwendig sind. Drinkmann (1990, S. 113) verweist in diesem Zusammen-

hang auf den problematischen Aspekt, dass gerade bei nicht signifikanten Ergebnissen die mitgeteilten Informationen in der Untersuchung meist geringer sind als bei signifikanten Ergebnissen: häufig wird hier nämlich nur auf die fehlende Signifikanz im Text verwiesen ("n.s."), während bei signifikanten Ergebnissen meist auch statistische Kennwerte angegeben werden. Grundsätzlich ist daher bei *missing da*ta zu unterscheiden, ob die Werte eher zufällig fehlen oder aufgrund einer bestimmten Beziehung zu den Daten, insbesondere weil sie nicht signifikant sind (vgl. Hedges 1990; Pigott 1994). Im ersten Fall ist ein Ausschluss der Ergebnisse unproblematisch, da ja keine Verzerrung des integrativen Ergebnisses zu erwarten ist. Im zweiten Fall ist der Einsatz statistischer Schätzverfahren möglich, die fehlende Werte auf der Basis der vorhanden Werte z.B. regressionsanalytisch schätzen (Pigott 1994). Alternativ kann bei fehlender Signifikanz als konservativer Schätzer auch eine Null vergeben werden, was allerdings tendenziell zu einer Unterschätzung des integrierten Ergebnisses führt (Lipsey & Wilson 2001, S. 70).

4.4 "Nonindependent Effects" – Die Integration abhängiger Daten

Schließlich ist bei der Integration der Untersuchungsergebnisse die Problematik der Abhängigkeit von Ergebnissen (*nonindependent effects* oder *multiple effect sizes*) zu berücksichtigen. Der letzte der zentralen Kritikpunkte an der Metaanalyse bezieht sich auf die Tatsache, dass mehrere relevante Ergebnisse in einer Studie, die bei den gleichen Untersuchungsobjekten erhoben wurden, statistisch nicht voneinander unabhängig sind und es bei einer Integration dieser abhängigen Ergebnisse zu einem verzerrten Integrationsergebnis kommen kann. Diese Abhängigkeit von Ergebnissen kann auf unterschiedliche Art berücksichtigt werden. Zum einen können abhängige Ergebnisse zusammengefasst werden, z.B. durch Mittelwertbildung oder die Berechnung des Medians, oder aber es wird eine aus mehreren abhängigen Ergebnissen zufällig oder systematisch (z.B. immer die kleinste Effektstärke) ausgewählt und geht als ein unabhängiges Ergebnis in die metaanalytische Auswertung ein (Rosenthal 1994b; Rosenthal & DiMatteo 2001). Auf diese Weise verringert sich aber die Datenbasis bei vielen Metaanalysen jedoch häufig beachtlich, weshalb dann eben auch oft auf eine derartige Zusammenfassung abhängiger Ergebnisse verzichtet wird. Bijmolt & Pieters (2001) konnten in einer Monte-Carlo-Studie zeigen, dass metaana-

lytische Verfahren, die auf einem einzigen Messwert je Studie beruhen, nicht nur einen hohen Informationsverlust aufweisen, sondern auch zu vergleichsweise schlechten Schätzergebnissen kommen. Sie unterschätzen die Effekte von Moderatorvariablen erheblich und die zusammengefassten Effektstärken weichen sogar deutlich von den tatsächlichen Werten ab. Dagegen schneiden Metaanalysen, die alle vorliegenden Effektstärken einbeziehen, besser ab und liefern weitgehend unverzerrte integrierte Effektstärken. Die abhängigen Ergebnisse können dazu auch gewichtet werden, so dass z.B. alle abhängigen Ergebnisse einer unabhängigen Stichprobe mit dem gemeinsamen Gewicht Eins in die Metaanalyse eingehen. Eine letzte Möglichkeit ist die Berücksichtigung der Kovarianz abhängiger Ergebnisse. Im Unterschied zu unabhängigen Ergebnissen ist die Kovarianz der abhängigen Ergebnisse nämlich ungleich Null, weshalb abhängige Ergebnisse auch durch die Einbeziehung der Varianz-Kovarianzmatrix gewichtet werden können und so die Abhängigkeit bei der Integration sowie bei der Moderatoranalyse beispielsweise durch die Einbeziehung der Varianz-Kovarianzmatrix im Rahmen eines GLS-Regressionsmodells berücksichtigt werden kann (Becker & Schram 1994; Gleser & Olkin 1994; Raudenbush et al. 1988). Allerdings werden diese Kovarianzen in den meisten Studien nicht mitgeteilt, wodurch die praktische Anwendbarkeit des Ansatzes eher fraglich ist (Lipsey & Wilson 2001, S. 126).

4.5 Weitere Kritikpunkte

Weitere Punkte, die allerdings in der Literatur als weniger kritisch betrachtet werden, sind die Beschränktheit der Anwendung der Metaanalyse auf quantitative empirische Arbeiten oder auch die Konzentration auf Haupteffekte, während Interaktionseffekte, die oftmals die eigentlich interessierenden Effekte in vielen experimentellen Primäruntersuchungen darstellen, aufgrund des Fehlens metaanalytischer Techniken zur Analyse von Interaktionseffekten unberücksichtigt bleiben (Drinkmann 1990, S. 28ff.). Das mittlerweile recht umfangreiche Methodenarsenal der Metaanalyse bietet nicht nur recht unterschiedliche Analysestrategien, vielmehr stellt es den Forscher auch vor das Problem, welche Strategie denn für seine Fragestellung die geeignete ist, zumal durchaus auch von einer methodenbedingten Ergebnisvarianz auszugehen ist (Cooper 1984; Strube & Hartman 1982). Der Spielraum des Forschers bei der

Durchführung einer Metaanalyse kann also durchaus auch zu einem "personal bias" führen (Stamm & Schwarb 1995). Als problematischer Punkt hinsichtlich der praktischen Durchführung gilt es außerdem zu bedenken, dass Metaanalysen doch vergleichsweise aufwändig sind und auch ein entsprechendes Methodenwissen erfordern (Lipsey & Wilson 2001, S. 7).

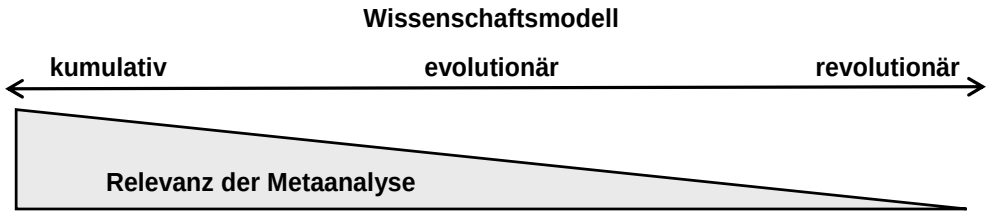
5 Wissenschaftstheoretische und methodologische Einordnung

Metaanalysen sind vor dem Hintergrund der stetig wachsenden und immer unübersichtlicher werdenden Menge an Publikationen in vielen Disziplinen als ein von pragmatischen Interessen geleitetes Werkzeug entwickelt worden. Auch die Diskussion und Weiterentwicklung ist stark von technischen und pragmatischen Themen und Fragestellungen geprägt. Dagegen fand eine Auseinandersetzung um die wissenschaftstheoretische Fundierung der Metaanalyse zunächst kaum statt. Mit der Weiterentwicklung der Metaanalyse wurden aber immer öfter auch Fragen zum Theorie- und Erkenntnisbeitrag der Metaanalyse thematisiert (z.B. Cook et al. 1992; Cooper & Hedges 1994b; Wachter & Straf 1990). Diese Überlegungen werden im Folgenden aufgegriffen.

Wissenschaftstheoretisch fußt die Grundidee der Metaanalyse auf dem Modell kumulativen wissenschaftlichen Fortschritts, in dem es um die Akkumulation und Ausdifferenzierung von Wissensbeständen geht. Hier ist ein Rückgriff auf bestehende Wissensbestände im Sinne einer bilanzierenden und qualifizierenden Analyse möglich und sinnvoll. Innerhalb eines revolutionären Fortschrittsmodells sensu Kuhn (1962) spielt die Metaanalyse dagegen praktisch keine Rolle, da hier revolutionäre Paradigmenwechsel zur Inkomensurabilität vorherigen Wissens führen und dabei die bearbeiteten Probleme, Konzepte und Variablen so grundlegend ändern, dass ein Rückgriff auf bisheriges Wissen nicht mehr erforderlich und auch nicht mehr möglich ist (Drinkmann & Groeben 1989, S. 190). Allerdings wird die Annahme der grundlegenden Inkomensurabilität von Wissensbeständen aus verschiedenen Paradigmen aus heutiger Sicht deutlich relativiert, können doch alte Theorien auch nach Paradigmenwechsel oftmals in neue Theorien eingebettet werden (Stegmüller 1980). Daher wird zunehmend auch ein Evolutionsmodell des Theorienfortschritts vertreten, das von einer Wissensakkumulation ausgeht, die jedoch nicht zielstrebig, systematisch, geordnet und aufeinander aufbauend verläuft, sondern ungeordneten und fragmentierten Prozessen der Selektion unterliegt, vergleichbar der natürlichen Selektion in der Evolution (Campbell 1974; Nersessian 1987). Theoriesprünge oder neue Orientierungen, die die bisher bearbeiteten Konzepte nicht grundsätzlich verwerfen, können dabei durchaus durch eine rückwärtsorientierte Integration erfasst werden (Drinkmann & Groeben 1989, S. 190). Bei durchgreifenden innovativen Theorieent-

wicklungen und Neuausrichtungen, die zur völligen Neu-Konzeptualisierung von potenziell metaanalysierbaren Konstrukten führen, stößt die Metaanalyse jedoch an ihre Grenzen. Offensichtlich gibt es dabei zwischen völlig revolutionären Neuorientierungen und kleineren Theoriesprüngen wohl eher einen fließenden Übergang als eine eindeutige Trennung, die auch begrifflich in der Wissenschaft nicht zuletzt durch das vieldeutige Konzept des Paradigmas nur schwer zu erfassen ist (vgl. Masterman 1974). Da Metaanalysen heute meist nur im Bezug auf eine bestimmte Fragestellung angewandt werden, die durch Forscher mit ähnlicher theoretischer und methodischer Grundorientierung innerhalb einer bestimmten zeitlichen Spanne bearbeitet wird, laufen Metaanalysen in ihrer Anwendung bisher kaum Gefahr, auf inkommensurable Wissensbestände zu treffen, obwohl diese Beschränkung natürlich potenziell besteht.

Abb. 4: Relevanz der Metaanalyse in verschiedenen Wissenschaftsmodellen



Weiterhin stellt sich aus wissenschaftstheoretischer Perspektive auch die Frage, welchen eigenen Beitrag Metaanalysen überhaupt zur Theorieentwicklung und –prüfung leisten können oder ob die Metaanalyse nicht per se einfach nur ein statistisches Instrument darstellt, das völlig a-theoretisch angelegt ist. Dazu unterscheiden Miller & Pollock (1994) zwischen drei Typen von Metaanalysen.

- Metaanalysen vom Typ A integrieren Studien zu einer Fragestellung und prüfen die Signifikanz der ermittelten integrierten Effektstärke. Ihr Theoriebeitrag ist also die Bestätigung einer vorhandenen Hypothese auf breiterer Stichprobenbasis und damit auch mit einer höheren statistischen Signifikanz der Ergebnisse und auch mit einem reduzierten Fehler zweiter Art (Cooper & Rosenthal 1980). Werden dabei in einer Metaanalyse Effektstärken zu mehreren bekannten theore-

tisch postulierten Beziehungen hinsichtlich einer Fragestellung zusammengetragen, kann auch der Erklärungsbeitrag der jeweiligen Theorie anhand der Stärke und der Homogenität der jeweils relevanten integrierten Effektstärke überprüft und verglichen werden (Hall et al. 1994).

- Metaanalysen vom Typ B untersuchen Moderatoren, die bereits in anderen Studien untersucht wurden und aus den Primäruntersuchungen auch direkt ermittelbar sind, wie z.B. das Geschlecht der Probanden. Die ermittelten Ergebnisse tragen dabei zu einer theoretischen Differenzierung oder Generalisierung bei, je nachdem, ob die untersuchten Moderatoren verantwortlich sind für die Variabilität der einzelnen Effektstärken oder nicht. Spielt beispielsweise das Geschlecht der Probanden keine Rolle für die Variabilität der ermittelten Effektstärken, so lässt sich die untersuchte Beziehung der Variablen über das Geschlecht verallgemeinern und die zugrunde liegende Theorie auch entsprechend generalisieren. Ist das Geschlecht dagegen ein signifikanter Moderator, der zur Erklärung der Varianz der Effektstärken beiträgt, so sollte die der Variablenbeziehung zugrunde liegende Theorie zukünftig das Geschlecht als differenzierenden Faktor mit berücksichtigen.

- Typ C-Metaanalysen schließlich testen neue, also in bisheriger Primärforschung noch nicht untersuchte Hypothesen. Dazu werden theoretisch relevante Moderatorvariablen auf der Basis von Informationen aus der Studie sowie des Kontexts neu generiert (vgl. Mullen et al. 1991). Ein bekanntes Beispiel hierzu stammt aus der Sozialpsychologie: Mullen et al. (1990) integrierten verschiedene Studien, die untersuchten, welchen Einfluss eine Person, die bei Rot über die Straße geht oder an der Ampel wartet auf die Bereitschaft der anderen Personen hat, bei Rot über die Straße zu gehen. Sie postulierten nun die bisher noch nicht untersuchte Hypothese, dass auch die Überfüllung des Gehwegs mit Menschen einen zusätzlichen Einfluss haben könnte, also eine Moderatorvariable darstellt, und ordneten dazu der jeweiligen Tageszeit, an der die Primäruntersuchung durchgeführt wurde, auf der Basis von Archivdaten über das Verkehrsaufkommen in Städten einen Grad der Überfüllung zu. Dabei konnten sie nachweisen, dass die Überfüllung des Gehwegs tatsächlich eine relevante Einflussgröße auf die untersuchte Bereitschaft der Personen, bei Rot über die Straße zu gehen, darstellt. Eine be-

triebswirtschaftliche Anwendung wäre beispielsweise in Untersuchungen zur Werbewirkung denkbar, wo aus der Kategorisierung der beworbenen Produkte als Such-, Erfahrungs- oder Vertrauensgüter abgeleitet werden könnte, wie hoch die Qualitätsunsicherheit auf Seiten des Konsumenten ist, eine Variable, die zur Erklärung des Informationsverhaltens von Konsumenten und damit auch der Werbewirkung herangezogen werden könnte.

Zwei problematische Aspekte sind bei der metaanalytischen Theorieentwicklung und –prüfung jedoch zu bedenken. Erstens können Moderatoreffekte nicht völlig isoliert werden, wie es eigentlich für eine strenge Ursache-Wirkungs-Prüfung nötig wäre, da sich Untersuchungen ja typischerweise in mehr als einem Merkmal unterscheiden, so dass streng genommen eine Kausalität auch nicht nachweisbar ist (Shadish & Sweeney 1991). Diese Problematik wird auch relevant, wenn nachgeschaltete statistische Verfahren der Kausalanalyse zur Untersuchung von Struktur- und Prozessmodellen mit moderierenden und mediierenden Effekten eingesetzt werden, die auf metaanalytisch ermittelten Kovarianzen oder Korrelationen beruhen (vgl. zur Kausalanalyse auch Abschnitt 0). Streng genommen müsste jede der untersuchten Kovarianzen bzw. Korrelationen auf Ergebnissen aus streng kontrollierten Experimenten beruhen, um tatsächlich von Kausalitäten ausgehen zu können (Miller & Pollock 1994). Zweites ist davon auszugehen, dass die Bildung von Hypothesen für die Metaanalyse sehr wahrscheinlich auf der Basis der von der Forscherin bzw. dem Forscher gesichteten und einbezogenen Untersuchungen getroffen wurde oder zumindest durch die Sichtung dieser Untersuchungen beeinflusst wurde. Dieselben Daten dürfen aber nicht gleichzeitig zur Entwicklung und Testung von Theorien verwendet werden, sondern beide Schritte sollten auf unabhängigen Daten beruhen, weshalb die Metaanalyse bei der Überprüfung von Theorien auch keine weiteren Primärstudien ersetzen kann, sondern diese vielmehr erforderlich macht (Cooper & Hedges 1994a; Cooper & Lemke 1991). Daher sind Metaanalysen zwar als eine Synthetisierung vergangener Forschung zu verstehen, die aber gleichzeitig auch weitere Forschung anregen soll. Die Hinweise auf zukünftige Forschungsbemühungen müssen dabei nicht nur auf der Basis der in der Metaanalyse untersuchten oder vorgeschlagenen Moderatoren erfolgen, auch durch die reine Bilanzierung von Befunden vor einem theoretischen Kontext können Bereiche der empirischen Unterdeterminiertheit dieser Theorien aufgezeigt werden (Eagly & Wood 1994).

Aus methodischer Sicht stellt die Metaanalyse nicht etwa ein einzelnes, klar abgegrenztes Verfahren dar, vielmehr handelt es sich um eine Methodenfamilie, die verschiedene Phasen eines Analyseprozesses umfasst und auch mehrere alternative Methoden innerhalb einzelner Phasen ermöglicht (Drinkmann 1990, S. 6). Glass et al. (1981, S. 21) verstehen Metaanalysen daher auch in erster Linie als Aufforderung, integrative Zusammenfassungen durch exakte statistische Methoden zu ergänzen, weshalb Metaanalysen eher eine Forschungsperspektive als eine spezielle Technik darstellen, bei der die verschiedensten statistischen Methoden eingesetzt werden. Entsprechend haben sich auch mehrere metaanalytische Schulen mit unterschiedlicher methodischer Schwerpunktbildung herausgebildet. Unterscheiden lassen sich hierbei insbesondere die Glasssche Variante der Effektstärkenintegration (Glass et al. 1981), die Variante der Integration von Irrtumswahrscheinlichkeiten (Rosenthal 1978), die Variante der Effektstärkenintegration mit Einbeziehung von Homogenitätstests und Bildung von Subgruppen (Hedges 1981, 1982a, b; Hedges & Olkin 1985) und die Variante der Effektstärkenintegration mit der Korrektur von Stichproben- und Messfehlern (Hunter & Schmidt 1990; Hunter et al. 1982). Wenngleich die meisten Metaanalysen einer bestimmten methodischen Schule folgen, so kombinieren sie doch auch verschiedene Aspekte aus diesen Schulen, beispielsweise die Messfehlerkorrekturen und die Homogenitätstests. Die Vielfalt der durchgeführten Metaanalysen kann daher besser an Hand anderer Unterscheidungsmerkmale eingeordnet werden. Eine sehr umfassende Kategorisierung schlagen Beelmann & Bliesener (1994) vor, die folgende methodischen Unterscheidungskriterien anführen:

- Primär- oder Sekundärfragestellung: Behandelt die Metaanalyse die Fragestellung der Primärstudie und/oder Fragestellungen zu relevanten Drittvariablen?
- Ziel der Fragestellung: Besitzt die Fragestellung eher explorativen, eher deskriptiven oder eher konfirmatorischen Charakter?
- Spezifität der Fragestellung: Werden eher breite Konstrukte untersucht oder eher definitorisch eng abgegrenzte Variablen?
- Spezifität der Studienauswahl: Werden alle verfügbaren Quellen einbezogen oder findet eine Vorauswahl durch allgemeine Auswahlkriterien wie z.B. an Hand des Veröffentlichungszeitraums statt?

- Grad der Studienauswahl: Werden Studien aufgrund der Studienqualität vorselektiert oder findet keine entsprechende Selektion statt?
- Analyseeinheit: Stellt die einzelne Studie oder die einzelne Effektgröße die Analyseeinheit dar?
- Basisparameter: Handelt es sich bei den zu integrierenden Parametern um Irrtumswahrscheinlichkeiten oder um Effektstärkenparameter?
- Effektstärkengewichtung: Gehen die einzelnen Basisparameter ohne Gewichtung in die Integration ein oder findet eine Gewichtung statt, wenn ja, wie wird gewichtet: an Hand der Stichprobengröße, Varianz, Studienqualität, etc.?
- Art der Varianzaufklärung: Werden ausschließlich globale Effektstärkenparameter ermittelt und/oder findet auch eine Analyse der Ergebnisvarianz statt?

Dieser Katalog verdeutlicht nicht zuletzt die methodische Breite von Metaanalysen und damit auch die Vielfalt an Anwendungsmöglichkeiten, die eine Anpassung der Methode an verschiedenste Forschungsprobleme ermöglicht.

6 Zusammenfassende Bewertung

Metaanalysen gelten heute als weitgehend konsolidierte Methode der Ergebniszusammenfassung, die im zunehmenden Maße zum Einsatz kommen. Metaanalysen helfen Wissenschaftlern bei der Informationsintegration und –bewertung im Rahmen ihrer Arbeit und unterstützen Praktiker bei der Entscheidungsfindung. Die problematischen Aspekte der Methode gelten heute als weitgehend ausdiskutiert, ihre Möglichkeiten und Grenzen sind umfassend definiert. Innerhalb dieser Grenzen kann die Metaanalyse auch einen eigenständigen Beitrag zum Erkenntnisfortschritt einer Disziplin leisten. Mit den vorangegangen Ausführungen wurde versucht, eine kurze Einführung in die Methode der Metaanalyse zu geben und dabei die Breite der Anwendungsmöglichkeiten der Methode aufzuzeigen. Die Diskussion der kritischen Aspekte der Metaanalyse sowie die wissenschaftstheoretische Verortung sollten die Grenzen und Möglichkeiten der Metaanalyse noch weiter verdeutlichen.

.

Literatur

Arthur, Winfred, Jr., Winston Bennett, Jr. & Allen I. Huffcutt (2001), *Conducting Meta-Analysis Using SAS*, Mahwah, NJ: Lawrence Erlbaum.

Bangert-Drowns, Robert L. (1986), Review of Developments in Meta-Analytic Method, *Psychological Bulletin*, 99, S. 388-399.

Beaman, Arthur L. (1991), An Empirical Comparison of Metaanalytic and Traditional Reviews, *Personality and Social Psychology Bulletin*, 17, S. 252-257.

Becker, Betsy Jane (1994), Combining Significance Levels, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 215-230.

Becker, Betsy Jane & Christine M. Schram (1994), Examining Explanatory Models Through Research Synthesis, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 357-381.

Beelmann, Andreas & Thomas Bliesener (1994), Aktuelle Probleme und Strategien der Metaanalyse, *Psychologische Rundschau*, 45, S. 211-233.

Begg, Colin B. (1994), Publication Bias, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 399-409.

Bijmolt, Tammo H. A. & Rik G. M. Pieters (2001), Meta-Analysis in Marketing when Studies Contain Multiple Measurements, *Marketing Letters*, 12, S. 157-169.

Brown, Stephen P. & Robert A. Peterson (1993), Antecedents and Consequences of Salesperson Job Satisfaction: Meta-Analysis and Assessment of Causal Effects, *Journal of Marketing Research*, 30, S. 63-77.

Bushman, Brad J. (1994), Vote-Counting Procedures in Meta-Analysis, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Reserach Synthesis*, New York: Russell Sage Foundation, S. 193-213.

Campbell, Donald T. (1974), Evolutionary Epistemology, In Paul Arthur Schilpp (Hrsg.), *The Philosophy of Karl Popper. Book I*, La Salle, Ill: Open Court, S. 413-463.

Cheung, Mike W. L. (2002), *Meta-Analysis for Structural Equation Modeling: A Two-Stage Approach,* Unpublished Doctoral Dissertation Dissertation, The Chinese University of Hong Kong, Hong Kong.

Cook, Thomas D., Harris Cooper, David S. Cordray, Heidi Hartman, Larry V. Hedges, Richard J. Light, Thomas A. Louis & Frederick Mosteller (1992), *Meta-Analysis for Explanation. A Casebook*, New York: Russell Sage Foundation.

Cooper, Harris (1984), *The Integrative Research Review: A Systematic Approach*, Beverly Hills, CA: Sage.

Cooper, Harris (1989), *Integrating Research: A Guide for Literature Reviews*, 2. Aufl., Newbury Park, CA: Sage.

Cooper, Harris & Larry V. Hedges (1994a), Potentials and Limitations of Research Synthesis, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 521-529.

Cooper, Harris & Larry V. Hedges (1994b), Research Synthesis as a Scientific Enterprise, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 3-14.

Cooper, Harris & Larry V. Hedges (Hrsg.) (1994c), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation.

Cooper, Harris M. (1982), Scientific Guidelines for Conducting Integrative Research Reviews, *Review of Educational Research*, 52, S. 291-302.

Cooper, Harris M. & Kevin M. Lemke (1991), On the Role of Meta-analysis in Personality and Social Psychology, *Personality and Social Psychology Bulletin*, 17, S. 245-251.

Cooper, Harris & Robert Rosenthal (1980), Statistical Versus Traditional Procedures for Summarizing Research Findings, *Psychological Bulletin*, 87, S. 442-449.

Drinkmann, Arno (1990), *Methodenkritische Untersuchungen zur Metaanalyse*, Weinheim: Deutscher Studien-Verlag.

Drinkmann, Arno & Norbert Groeben (1989), *Metaanalysen für Textwirkungsforschung. Methodologische Varianten und inhaltliche Ergebnisse im Bereich der Persuasionswirkung von Texten*, Weinheim: Deutscher Studien Verlag.

Durlak, Joseph A. & Mark W. Lipsey (1991), A Practitionner's Guide to Meta-Analysis, *American Journal of Community Psychology*, 19, S. 291-332.

Eagly, Alice H. & Wendy Wood (1994), Using Research Synthesis to Plan Future Research, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 485-500.

Farley, John U. & Donald R. Lehmann (1986), *Meta-Analysis in Marketing. Generalization of Response Models*, Lexington, MA: Lexington Books.

Farley, John U., Donald R. Lehmann & Michael J. Ryan (1981), Generalizing from 'Imperfect' Replication, *Journal of Business*, 54, S. 597-610.

Fleiss, Joseph L. (1994), Measures of Effect Size for Categorical Data, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 245-260.

Fricke, Reiner & Gerhard Treinies (1985), *Einführung in die Metaanalyse*, Bern: Huber.

Glass, Gene V. (1976), Primary, Secondary, and Meta-Analysis of Research, *Educational Researcher*, 5, S. 3-8.

Glass, Gene V. (1977), Integrating Findings: The Meta-Analysis of Research, *Review of Research in Education*, 5, S. 351-379.

Glass, Gene V., Berry McGaw & Mary Lee Smith (1981), *Meta-Analysis in Social Research*, Beverly Hills, CA: Sage.

Gleser, Leon J. & Ingram Olkin (1994), Stochastically Dependent Effect Sizes, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 339-356.

Green, Bert F. & Judith A. Hall (1984), Quantitative Methods for Literature Review, *Annual Review of Psychology*, 35, S. 37-53.

Greenland, Sander (1994), Invited Commentary: A Critical Look at Some Popular Meta-Analytic Methods, *American Journal of Epidemiology*, 140, S. 290-296.

Hall, Judith A., Linda Tickle-Degnen, Robert Rosenthal & Frederick Mosteller (1994), Hypotheses and Problems in Research Synthesis, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 17-28.

Halvorsen, Katherine Taylor (1994), The Reporting Format, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 425-437.

Hedges, Larry V. (1981), Distribution Theory for Glass's Estimator of Effect Size and Related Estimators, *Journal of Educational Statistics*, 6, S. 107-128.

Hedges, Larry V. (1982a), Estimation of Effect Size From a Series of Independent Experiments, *Psychological Bulletin*, 92, S. 490-499.

Hedges, Larry V. (1982b), Fitting Categorical Models to Effect Sizes From a Series of Experiments, *Journal of Educational Statistics*, 7, S. 119-137.

Hedges, Larry V. (1990), Directions for Future Methodology, In Kenneth W. Wachter & Miron L. Straf (Hrsg.), *The Future of Meta-Analysis*, New York: Russell Sage Foundation, S. 11-26.

Hedges, Larry V. (1994a), Fixed Effect Models, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 285-299.

Hedges, Larry V. (1994b), Statistical Considerations, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 29-38.

Hedges, Larry V. & Ingram Olkin (1980), Vote-Counting Methods in Research Synthesis, *Psychological Bulletin*, 88, S. 359-369.

Hedges, Larry V. & Ingram Olkin (1985), *Statistical Methods for Meta-Analysis*, Orlando, FL: Academic Press.

Hedges, Larry V. & Jack L. Vevea (1998), Fixed- and Random-Effects Models in Meta-Analysis, *Psychological Methods*, 3, S. 486-504.

Hunt, Morton (1997), *How Science Takes Stock: The Story of Meta-Analysis*, New York: Russell Sage Foundation.

Hunter, John E. (2001), The Desperate Need for Replications, *Journal of Consumer Research*, 28, S. 149-158.

Hunter, John E. & Frank L. Schmidt (1990), *Methods of Meta-Analysis. Correcting Error and Bias in Research Findings*, Newbury Park, CA: Sage.

Hunter, John E. & Frank L. Schmidt (1994), Correcting for Sources of Artifical Variation Across Studies, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 323-336.

Hunter, John E., Frank L. Schmidt & Gregg B. Jackson (1982), *Meta-Analysis. Cumulating Research Findings Across Studies*, Beverly Hills, CA: Sage.

Jackson, Gregg B. (1980), Methodes for Integrative Reviews, *Review of Educational Research*, 50, S. 438-460.

Kuhn, Thomas S. (1962), *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

Light, Richard J. & David B. Pillemer (1984), *Summing up: The Science of Literature Review*, Cambridge, MA: Harvard University Press.

Light, Richard J., Judith D. Singer & John B. Willett (1994), The Visual Presentation and Interpretation of Meta-Analysis, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 439-453.

Lipsey, Mark W. (1994), Identifying Potentially Interesting Variables and Analysis Opportunities, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 111-123.

Lipsey, Mark W. & David T. Wilson (2001), *Practical Meta-Analysis*, Thousand Oaks, CA: Sage.

Mann, Charles C. (1990), Meta-Analysis in the Breech, *Science*, 249, S. 476-480.

Mann, Charles C. (1994), Can Meta-Analysis Make Policy?, *Science*, 266, S. 960-962.

Masterman, M. (1974), The Nature of a Paradigm, In I. Lakatos & A. Musgrave (Hrsg.), *Criticism and the Growth of Knowledge*, Cambridge, MA, S. 59-90.

Miller, Norman & Vicki E. Pollock (1994), Meta-Analytic Synthesis for Theory Development, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 457-483.

Mullen, Brian, Carolyn Cooper & James E. Driskell (1990), Jaywalking as a Function of Model Behavior, *Personality and Social Psychology Bulletin*, 16, S. 320-330.

Mullen, Brian, Eduardo Salas & Norman Miller (1991), Using Meta-Analysis to Test Theoretical Hypotheses in Social Psychology, *Personality and Social Psychology Bulletin*, 17, S. 258-264.

Nersessian, Nancy (Hrsg.) (1987), *The Process of Science*, Dordrecht: Nijhoff.

Normand, Sharon-Lise T. (1995), Meta-Analysis Software: A Comparative Review, *American Statistician*, 49, S. 298-309.

Orwin, Robert G. (1994), Evaluating Coding Decisions, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 139-162.

Pearson, Karl (1904), Report on Certain Enteric Fever Inoculation Statistics, *British Medical Journal*, 3, S. 1243-1246.

Pigott, Therese D. (1994), Methods for Handling Missing Data in Research Synthesis, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 163-175.

Plath, Ingrid (1992), *Understanding Meta-Analyses. A Consumer's Guide to Aims, Problems, Evaluation and Developments*, Baden-Baden: Nomos.

Price, Derek de Solla (1976), *Science since Babylon. Second Edition*, New Haven, London: Yale University Press.

Price, Derek de Solla (1986), *Little Science, Big Science ... and beyond*, 2. Aufl., New York: Columbia University Press.

Raudenbush, Stephen W., Betsy Jane Becker & Hripsime Kalaian (1988), Modelling Multivariate Effect Sizes, *Psychological Bulletin*, 103, S. 111-120.

Rosenthal, Marylu C. (1994a), The Fugitive Literature, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 85-94.

Rosenthal, Robert (1978), Combining Results of Independent Studies, *Psychological Bulletin*, 85, S. 185-193.

Rosenthal, Robert (1979), The "File Drawer Problem" and Tolerance for Null Results, *Psychological Bulletin*, 86, S. 638-641.

Rosenthal, Robert (1984), *Meta-Analytic Procedures for Social Research*, Beverly Hills, CA: Sage.

Rosenthal, Robert (1994b), Parametric Measures of Effect Sizes, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 231-244.

Rosenthal, Robert & M. R. DiMatteo (2001), Meta-Analysis: Recent Developments in Quantitative Methods for Literature Reviews, *Annual Review of Psychology*, 59, S. 59-82.

Rust, Roland T., Donald R. Lehman & John U. Farley (1990), Estimating Publication Bias in Meta-Analysis, *Journal of Marketing Research*, 27, S. 220-226.

Shadish, William R. (1996), Meta-Analysis and the Exploration of Causal Mediating Processes: A Primer of Examples, Methods, and Issues, *Psychological Methods*, 1, S. 47-65.

Shadish, William R. & C. Keith Haddock (1994), Combining Estimates of Effect Sizes, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 261-281.

Shadish, William R., Jr. & Rebecca B. Sweeney (1991), Mediators and Moderators in Meta-Analysis: There's a Reason We Don't Let Dodo Birds Tell Us Which Psychotherapies Should Have Prizes, *Journal of Consulting and Clinical Psychology*, 59, S. 883-893.

Slavin, Robert E. (1986), Best-Evidence Synthesis: An Alternative to Meta-Analytic and Traditional Reviews, *Educational Researcher*, 15, S. 5-11.

Smith, Mary L., Gene V. Glass & Thomas I. Miller (1980), *The Benefits of Psychotherapy*, Baltimore, MD: John Hopkins University Press.

Stamm, Hansueli & Thomas M. Schwarb (1995), Metaanalyse. Eine Einführung, *Zeitschrift für Personalforschung*, 9, S. 5-27.

Stegmüller, Wolfgang (1980), *Neue Wege der Wissenschaftsphilosophie*, Berlin: Springer.

Sterne, J. A. C., M. Egger & A. J. Sutton (2001), Meta-Analysis Software, In M. Egger, G. Davey Smith & D. G. Altman (Hrsg.), *Systematic Reviews in Health Care: Meta-Analysis in Context*, London: BMJ Books, S. 336-346.

Strube, Michael J. & Donald P. Hartman (1982), A Critical Appraisal of Meta-Analysis, *British Journal of Clinical Psychology*, 21, S. 129-139.

Sutton, Alexander J., Paul C. Lambert, Keith R. Abrams, David R. Jones & Martin Hellmich (2000), Meta-Analysis in Practice: A Critical Review of Available Software, In Darlene K. Stangl & Donald A. Berry (Hrsg.), *Meta-Analysis in Medicine and Health Policy*, New York: Marcel Dekker, S. 359-390.

Wachter, Kenneth W. & Miron L. Straf (Hrsg.) (1990), *The Future of Meta-Analysis*, New York: Russell Sage Foundation.

White, Howard D. (1994), Scientific Communication and Literature Retrieval, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 41-55.

Wolf, Fredric M. (1994), *Meta-Analysis. Quantitative Methods for Research Synthesis*, Newbury Park, CA: Sage.

Wortman, Paul M. (1994), Judging Research Quality, In Harris Cooper & Larry V. Hedges (Hrsg.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, S. 97-109.

ISBN 3-935058-77-2
Freie Universität Berlin
Garystr. 21
D-14195 Berlin