

### 2.2.2.3 Metaanalyse

Wissenschaftlicher Erkenntnisfortschritt kann nicht alleine durch empirische Untersuchungen und theoretische Überlegungen vorangetrieben werden. Angesichts der Flut von Publikation zum gleichen Gegenstandsbereich mit zum Teil widersprüchlichen Ergebnissen, gilt es auch, sich einen Überblick zu verschaffen und somit den aktuellen Forschungsstand zu ermitteln. Das vertraute Vorgehen besteht in der Erstellung eines *reviews*, also eines Überblickartikels (Sammelreferats), in welchem die vorfindbare Literatur besprochen und integriert wird. Der Nachteil dieser Vorgehensweise besteht in der Subjektivität. Reviews verschiedener Wissenschaftler zum gleichen Thema können durchaus einen unterschiedlichen Forschungsstand vermitteln.

Besonders groß ist diese Gefahr beim *narrativen Review*, da dort die Gefahr besteht, dass die Autoren solcher Artikel die Literatur oft so auswählen, dass ihre vorgefassten Schlussfolgerungen bestätigt werden. Unliebsame Studien, die das Gegenteil beweisen, werden dagegen einfach nicht beachtet. Beim *systematischen Review* ist diese Gefahr geringer, da auf der Basis einer systematischen Literatursuche der aktuelle Wissensstand zusammengetragen und interpretiert wird. Diese Reviews sind weniger anfällig für Verzerrungen und Subjektivität.

Dem Ziel der Erstellung eines aktuellen Forschungsstands dient auch die *Metaanalyse*, die in den letzten Jahrzehnten fortwährend weiterentwickelt wird und von der man sich mehr Objektivität erhofft, weil sie verstärkt statistische Überlegungen einbezieht. Man spricht daher gelegentlich auch von quantitativer Ergebniszusammenfassung.

#### Definition: Metaanalyse

„Die Metaanalyse ist eine an den Kriterien empirischer Forschung orientierte Methode zur quantitativen Integration der Ergebnisse empirischer Untersuchungen sowie zur Analyse der Variabilität dieser Ergebnisse“ (Drinkmann, 1990, S. 11).

Nach Glass (1976, S. 3) versteht man unter Metaanalyse eine Art Tertiäranalyse. „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. I use it to refer to the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings.

Ein Beispiel aus dem Bereich der Therapieforschung soll diesen Ansatz verdeutlichen. Die Fragestellung beschäftigt sich mit dem Phänomen, dass es Kindergartenkinder gibt, die zu Gleichaltrigen keinen Kontakt aufnehmen können, wohl aber zu Erwachsenen. Die Methode der Wahl scheint die Verhaltenstherapie xyz zu sein, wonach jedes Verhalten, das auf die Kontaktaufnahme mit Gleichaltrigen hinweist, durch Lob direkt verstärkt wird und die Versuche, Kontakt mit Erwachsenen (Erziehern) aufzunehmen, geschwächt (ignoriert) werden.

Die Effizienz der therapeutischen Maßnahme wird an der AV „Dauer des Kontakts mit Gleichaltrigen pro Stunde“ gemessen. Der Untersuchung wird ein einfaktorieller Plan mit Experimental- und Kontrollgruppe (Stufen der UV B: B<sub>1</sub> mit bzw. B<sub>2</sub> ohne Therapie) zu Grunde gelegt (vgl. Kapitel 2.2.1.4), der einen Vergleich der AV vor und nach der Intervention ermöglicht. Tabelle 2.x9 zeigt diesen Versuchsplan.

**Tabelle 2.x9:** Der einfaktorielles Kontrollgruppenversuchsplan zum Beispiel der Prüfung der Therapie xyz bei kontaktgestörten Kindergartenkindern

<b>UV B: Therapie xyz</b>	
B <sub>1</sub> : mit (Experimentalgruppe)	B <sub>2</sub> : ohne (Kontrollgruppe)
AV der Vp 1	AV der Vp 31
AV der Vp 2	AV der Vp 32
.....	.....
AV der Vp 30	AV der Vp 60

Ein Blick in die Fachliteratur zeigt, dass zu dieser Fragestellung und zu diesem Vorgehen bereits eine Reihe von Untersuchungen vorliegen. Allerdings sind die Ergebnisse zum Teil widersprüchlich, da der postulierte Therapieeffekt sich mit unterschiedlicher Relevanz (Effektgröße; nähere Ausführungen dazu im weiteren Verlauf dieses Abschnitts) bzw. überhaupt nicht zeigt. Welcher dieser (angenommen) 85 Untersuchungen soll man nun vertrauen bzw. welche Aussage zum aktuellen Forschungsstand kann man treffen?

Relativ große Übereinstimmung besteht darin, dass eine Metaanalyse folgende Verfahrensschritte umfassen sollte:

- Bestimmung der empirisch-inhaltlichen Hypothese,
- umfassende Literatursuche,
- Bewertung und Codierung der Studien,
- statistische Analyse,
- Interpretation.

Gemäß des gewählten Beispiels bestünde die Konkretisierung der Forschungsfrage etwa in der *empirisch-inhaltlichen Hypothese* (vgl. Kapitel 1.5.3): Wenn Kindergartenkinder mit Schwierigkeiten in der Kontaktaufnahme zu altersgleichen Kindern für erkennbare Versuche zur Kontaktaufnahme mit Altersgleichen verstärkt (belohnt) und für entsprechende Versuche mit Erwachsenen nicht verstärkt (ignoriert) werden, dann erhöht sich zumeist die Dauer der Kontakte mit Altersgleichen. Die Hypothese legt die UV und die AV einschließlich ihrer jeweiligen Operationalisierungen fest und beschreibt auch den Kausalcharakter der Relation zwischen UV und AV. Dieser Schritt ist wichtig, da die zu integrierenden Primärstudien unterschiedliche Operationalisierungen der UVn und AVn und Hypothesenformulierungen enthalten können, die Vergleiche bzw. Rekonstruktionen erfordern.

Eine möglichst umfassende und nachvollziehbar dokumentierte *Literatursuche* ist die unerlässliche Basis jeder Metaanalyse. Neben den entsprechenden *Fachzeitschriften* (z. B. Psychological Review) sind psychologische *Datenbanken* (z. B. PsychLit, Psyndex usw.), so genannte *graue Literatur* wie Institutszeitschriften (Kölner Psychologische Studien, Trierer Psychologische Berichte), *Internetrecherchen* usw. einzubeziehen. Denn wie in den Reviews besteht auch hier die Gefahr, dass die absichtliche oder unabsichtliche Nichtbeachtung von Studien zu Verzerrungen führen kann. „Es konnte gezeigt werden, dass Autoren aus nicht-englischsprachigen Ländern signifikante Ergebnisse bevorzugt in angloamerikanischen Zeitschriften publizieren, so dass die nichtsignifikanten Ergebnisse dann in deutschen, französischen oder auch spanischen Zeitschriften „verschwinden“. Da Medline seinen Schwerpunkt bei angloamerikanischen Journals hat, führt also eine unvollständige (auf Medline beschränkte; Anmerkung des Verfassers) Literatursuche tendenziell zu einer Überschätzung des Behandlungseffekts. Diese ernst zu nehmende Form von Verzerrung wird als „Language Bias“

bezeichnet“ (Sauerland, 2004). Ein vergleichbarer *Publikationsbias* liegt vor, wenn nicht-signifikante Untersuchungsergebnisse tendenziell eher in grauer Literatur (Institutszeitschriften) veröffentlicht werden. Aus methodologischer Sicht tragen dagegen nichtsignifikante Ergebnisse ebenso zum Erkenntnisfortschritt einer Wissenschaft bei, wie signifikante Ergebnisse (vgl. Hussy & Jain, 2002, S. 275ff).

In die *Bewertung* der gefundenen Studien gehen vor allem methodische Gütekriterien mit ein. Die Primärstudien werden nach dem Ausmaß der Kontrolle von Störvariablen (z. B. Randomisierung, Kontrollgruppe) ebenso beurteilt wie nach der Güte der Operationalisierung der UV und AV. Dabei ist insbesondere die Operationalisierung der AV von großer Bedeutung: Im gewählten Beispiel war es „die Dauer der Kontakte mit gleichaltrigen Kindern pro Stunde. In einer anderen ausgewählten Studie könnte „die Anzahl der Kontakte mit gleichaltrigen Kindern am Vormittag“ und in einer weiteren Studie das Urteil der Kindergärtnerin „zur Kontaktfähigkeit mit Gleichaltrigen“ als Operationalisierungsform herangezogen worden sein. Ersichtlich sind es vor allem die Kriterien der internen Validität und der Variablenvalidität, die die Bewertungsgrundlage bilden.

Welche Studien bleiben auf Grund solcher Überlegungen in der weiteren Analyse und welche sind auszuschließen? Diese *Auswahlproblematik* kreist um drei Aspekte, nämlich das

- „Müll rein Müll raus“ - Problem und das
- „Äpfel und Birnen“ – Problem und die
- Abhängigkeitsproblematik.

Das „*Müll rein Müll raus*“ - Problem thematisiert die unterschiedliche Qualität von Studien. Kann man reliable und valide Ergebnisse erwarten, wenn die Primärstudien aus methodischer Sicht (gravierende) Mängel enthalten? Zwei Lösungswege werden besprochen: Das Benutzen von *Ausschlusskriterien* oder die Einführung einer *Moderatorvariablen*. Ausschlusskriterien stehen für Mindeststandards, die erfüllt sein müssen, damit die Studie weiter in der Analyse verbleiben kann. So wird in verschiedenen Metaanalysen bspw. gefordert, dass neben der Experimentalgruppe mindestens eine Vergleichsgruppe an der Untersuchung beteiligt ist, gleichgültig ob es sich dabei um eine Kontrollgruppe oder zweite Experimentalgruppe handelt. Im gewählten Beispiel existiert eine Kontrollgruppe, so dass die Studie bezüglich dieses Auswahlkriteriums in der Analyse bleiben könnte. Weitere Kriterien könnten z. B. die standardisierte Datenerhebung und/oder die Kontrolle der VL-Merkmale durch das Randomisieren etc. sein. Allerdings muss man bedenken, dass mit zunehmend strengen Kriterien die Anzahl der verbleibenden Studien schrumpft und die Gesamtanalyse damit an Aussagekraft verliert. Bei der Verwendung der Moderatorvariable „Studienqualität“ bleiben auch methodisch schwächere Arbeiten in der Analyse, erhalten aber gemäß ihrer Codierung entweder ein schwächeres Gewicht oder es erfolgt ein Vergleich der Analyseergebnisse mit und ohne die als schwächer codierten Arbeiten.

Das „*Äpfel und Birnen*“ – Problem besteht darin dass man bekannter Weise Äpfel und Birnen nicht zusammenzählen darf: 2 Äpfel + 2 Birnen = ?. Weshalb aber soll man dann Studien zusammenfassen dürfen, die sich inhaltlich teilweise deutlich unterscheiden, so etwa – wie schon angesprochen - hinsichtlich der Operationalisierung von UV und AV. Für die Operationalisierung der AV gilt, dass hier deutliche Abweichungen zum Ausschluss führen müssen. Im gegebenen Beispiel könnte „die Zeit, welche das Kind in seiner Gruppe verbringt (und nicht außerhalb)“ als inadäquate (variableninvalide) Operationalisierung der AV „Häufigkeit und Dauer des Kontakts mit Gleichaltrigen“ durchaus dazu führen, dass ein Ausschluss in Betracht gezogen wird. Bei den UVn und weiteren Merkmalen der Studien (Stichprobe, Therapeuten usw.) dagegen kann man argumentieren, dass dadurch die interne

Validität und der Geltungsbereich (im Sinne der besprochenen direkten und systematischen Replikation) gestärkt wird (vgl. z. B. Hall et al., 1994, S. 11).

Das „*Abhängigkeits-Problem*“ entsteht, wenn mehrere, nicht aus unabhängigen Stichproben gewonnene Ergebnisse (Effektgrößen) pro Studie in die Analyse eingehen. Vor allem, wenn eine einzelne Primärstudie viele Teilergebnisse (Effektgrößen) beisteuert, kann die durchschnittliche Effektgröße, das Hauptergebnis der Metaanalyse, stark verzerrt sein. Die Beschränkung auf eine Effektgröße pro Studie kann manchmal das Problem lösen. Im gegebenen Beispielfall lägen mehrere Ergebnisse abhängiger Stichproben dann vor, wenn der Vergleich von Experimental- und Kontrollgruppe an mehreren AVn vollzogen würde. Das bedeutet, dass der vorliegende Datensatz einer Stichprobe zu mehreren (abhängigen) Ergebnissen führt, die in die Gesamtanalyse eingehen und dieser Primärstudie dadurch besonderes Gewicht verleihen würden. Würde dieser wiederholte Vergleich mit verschiedenen AVn mit einer jeweils neuen (unabhängigen) Stichprobe erfolgen, so läge kein Abhängigkeitsproblem vor.

Die *Codierung* der Studien erfolgt aber nicht nur bezogen auf die methodische Qualität, sondern im Hinblick auf eine Vielzahl weiterer Merkmale (Stichprobenmerkmale, Untersuchungszeitraum, situative Merkmale usw.), insbesondere aber mit Blick auf die Ergebnisse. Dazu zählen alle Informationen, die zur Berechnung eines Gesamtkennwerts der Metaanalyse – der Effektgröße – erforderlich sind, also etwa Mittelwerte, Standardabweichungen und Stichprobengröße bei Experimental- und Kontrollgruppe. Fehlen solche zentralen Kennwerte einer Untersuchung, so kann man entweder versuchen, diese im Nachhinein zu berechnen, oder man schließt die Untersuchung aus der weiteren Analyse aus.

Die *Analyse der Daten* erfolgt zweischrittig:

- Bestimmung der Gesamteffektgröße und
- Bestimmung der Homogenität der Varianzen.

Die *Gesamteffektgröße* wird aus den Effektgrößen der Primäruntersuchungen ermittelt. Gängige Indizes zur Darstellung der Effektgröße sind der quadrierte Korrelationskoeffizient und die standardisierte Mittelwertsdifferenz  $d'$ . Die Effektgröße gibt den Anteil der Varianz in der Messwertreihe an, der durch die UV aufgeklärt wird. Sie bestimmt also die psychologisch-inhaltliche Relevanz eines gefunden Mittelwertsunterschieds. Der gleiche Mittelwertsunterschied kann als signifikant ausgewiesen werden, aber – je nach Stichprobengröße – 5%, 25%, 50% und mehr Varianz aufklären, also unterschiedlich bedeutsam sein. Leider werden nicht in allen Untersuchungen solche Effektgrößeindizes mitgeteilt bzw. unterschiedliche Indizes ermittelt, so dass zunächst für jede Primärstudie der gleiche Index zu berechnen ist (weitere Einzelheiten dazu z. B. bei Lipsey & Wilson., 2001). Daran schließt sich im entscheidenden Schritt die Integration der einzelnen Effektgrößen in die Gesamteffektgröße  $\Delta$  an. Die Variabilität und die Stichprobengrößen der einzelnen Studien können dabei gewichtend bzw. korrigierend einbezogen werden. Im dargestellten Beispiel würde eine Gesamteffektgröße  $\Delta = 0,57$  bedeuten, dass die beschriebene therapeutische Vorgehensweise bei kontaktgestörten Kindergartenkindern auf dem Hintergrund von z. B. 65 analysierten Studien einen Mittelwertsunterschied zwischen Experimental- und Kontrollgruppe in der postulierten Richtung erbringt, der 57% der Varianz in der Gesamtmesswertreihe erklärt. Der Vorteil des Ergebnisses dieser Metaanalyse im Vergleich zu einem angenommen Ergebnis in der verwendeten Beispielsstudie von  $d' = 0,62$  liegt darin, dass die interne Validität (Möglichkeit zur Kausalinterpretation) wesentlich gestärkt ist, der Geltungsbereich sehr gut anhand der Primärstudien geprüft und die Zuverlässigkeit der Ergebnisse kaum noch angezweifelt werden kann. Voraussetzung dafür ist allerdings ein noch zu erbringender Nachweis der *Homogenität der Varianz der Effektgrößen* (weitere Einzelheiten dazu z. B. bei Lipsey & Wilson., 2001). Liegt Varianzheterogenität vor, so kann durch Bildung von Subgruppen

anhand von inhaltlichen und/oder methodischen Moderatorvariablen versucht werden, ein differenzierteres Ergebnismuster zu erreichen, z. B. dergestalt, dass sich für die Operationalisierungsform der AV „Dauer des Kontakts mit Gleichaltrigen pro Stunde“ ein  $\Delta = 0,67$  und für „Anzahl der Kontaktaufnahmen am Vormittag“ ein  $\Delta = 0,47$  ergibt.

Die *Veröffentlichung* einer Metaanalyse muss die beschriebenen Verfahrensschritte im Einzelnen darstellen und begründen, so dass der Leser sich ein genaues Bild über die Sammlung und Auswahl der Studien machen kann. Die Interpretation muss differenzierte Aussagen über interne Validität und Geltungsbereich der Ergebnisse enthalten und die ermittelte Gesamteffektgröße in den jeweiligen thematischen Rahmen einordnen.

Die Metaanalyse ist eine Forschungsmethode, die in der vorgegebenen Systematik eine Sonderposition einnimmt: Zwar ist sie keine wirklich experimentelle Methode, denn ihre Untersuchungseinheiten sind keine Vpn sondern Primärstudien, aber die Primärstudien selber können auch (quasi)experimenteller Natur sein und ihre Bewertungskriterien jenen des Experiments folgen. Metaanalysen prüfen in der Regel keine spezifische Hypothese, sondern durch Integration vieler Primärstudien eine Reihe eng verwandter Hypothesen. Sie ermitteln und beschreiben den Stand der Forschung zu einem Themenbereich, wobei das ermittelte Ergebnis kausalen Charakter besitzen kann. Wenngleich ein großer Teil der Metaanalysen – insbesondere aus der Therapieforschung – sich mit Kausalhypothesen beschäftigt, finden sich andererseits auch viele Analysen, die deskriptive Zusammenhangsstudien zum Gegenstand haben.