Hansueli Stamm, Thomas M. Schwarb*

Metaanalyse. Eine Einführung**

Die Metaanalyse ist eine Sekundäranalysemethode mit deren Hilfe quantitative Ergebnisse aus empirischen Untersuchungen zusammengefaßt und deren Variabilität untersucht werden. In diesem Beitrag wird - mit einem Beispiel aus der Personalauswahl - die Methode der Metaanalyse erklärt und diskutiert. Ziel ist es, einen Einblick in das Verfahren sowie den Nutzen und die Probleme dieser Technik zu ermöglichen. Es wird aufgezeigt, daß weniger die eigentlichen Ergebnisse der Metaanalyse der Wissenschaft wichtige Impulse gegeben haben, sondern vielmehr Nebeneffekte wie die Moderatorenanalyse, methodische Standardisierungen und Qualitätskriterien empirischer Forschung sowie Hinweise auf Forschungslücken.

Meta-analysis is a secondary analysis method which is used to integrate a number of empirical studies and to test the variability of quantitative results. This article explains and discusses the basics of meta-analysis, in respect to ist applications in human resource management. The main objective of the article is to give an insight into the procedures, the advantages, and problems of this method. It is demonstrated that major impulses of meta-analysis came not so much from the actual results, but rather from side effects such as the evidence of moderators, quality control, standard setting for single-studies, and indication of research gaps.

---

* cand. rer. pol. Hansueli Stamm, Jahrgang 1966, Student der Wirtschaftswissenschaften an der Universität Basel.

lic. rer. pol. Thomas M. Schwarb, Jahrgang 1960, bis 1994 Lehrassistent für Personalmanagement bei Prof. Dr. Werner R. Müller, Institut für Betriebswirtschaft, Wirtschaftswissenschaftliches Zentrum (WWZ) der Universität Basel, Petersgraben 51, CH - 4051 Basel.

Forschungsschwerpunkt: „Personalauswahl". Publikation u.a.: „Arbeitsinhalt und Leistungswille: Verdeckte Botschaften aus der Personalpraxis" (zusammen mit W.R. Müller) in Lattmann, Charles; Probst, Gilbert & Tapernoux, Frédéric (Hrsg.): „Die Förderung der Leistungsbereitschaft des Mitarbeiters als Aufgabe der Unternehmensführung", Physica-Verlag, 1992

1. Einleitung

Die Metaanalyse ist eine Sekundäranalyse-Methode, ohne welche die Integration quantitativer Resultate aus der Forschung in den Sozialwissenschaften nicht mehr denkbar ist. Speziell auf dem Gebiet der Arbeits- und Organisationspsychologie ist sie ein beliebtes Mittel, die große Anzahl von bereits existierenden Primäruntersuchungen zusammenzufassen.

Woher kommt die Idee zu dieser „neuen" Technik? Wie entsteht eine Metaanalyse, und wie liest man eine solche? Wo liegen ihre Stärken und Schwächen? Was können wir einer Metaanalyse als neue Erkenntnisse entnehmen? Auf diese Fragen soll dieser Beitrag eine kurze Antwort geben.

Dabei wird der Leser in die Metaanalyse eingeführt, so daß er eine solche verstehen und qualifizieren kann und über den Stand der gegenwärtigen Diskussion dieser Technik informiert ist. Die Ausführungen sollen ihm deren Nutzen und Probleme aufzeigen sowie die Möglichkeit geben, entsprechende Anwendungsfelder für die Forschung im Bereich des Personalmanagements kennenzulernen. Zur Illustration des Verfahrens wird ein Beispiel aus dem Gebiet der Personalauswahl vorgestellt. Die methodischen Grundsätze gelten aber analog für andere Gebiete. Es wird versucht, die Darstellung dieser Einführung so zu gestalten, daß sie auch ohne besondere Statistikkenntnisse problemlos verstanden werden kann. Leser und Leserinnen, welche sich nicht mit der eigentlichen (statistisch-mathematischen) Technik der Metaanalyse auseinandersetzen wollen, können das Kapitel 4.4 überspringen.

2. Fragestellung und Begriff

Diese Suche nach neuen oder verbesserten wissenschaftlichen Erkenntnissen kann durch die Anwendung verschiedenster der Problemstellung möglichst angepaßter Methoden erfolgen. Vor allem in den Sozialwissenschaften hat man dabei grundsätzlich die Wahl, ob man eine „Primäruntersuchung" oder eine „Sekundäranalyse" durchführen will.

Erstere ist ein von anderen Untersuchungen unabhängiger Originalbeitrag eines Forschers (Fricke & Treinies (1985:16)). Dazu müssen Daten mittels Experimenten, Beobachtungen, Befragungen usw. erhoben, mit verschiedenen, oft statistischen Verfahren ausgewertet und anschließend vom Wissenschaftler interpretiert werden. Eine solche Untersuchung ist meist relativ aufwendig und daher teuer. Dazu kommt, daß vielfach mit zu kleinen Stichproben gearbeitet wird und nur vermeintlich signifikante Resultate entstehen (vgl. Hunter & Schmidt (1991)).

Letztere bezieht ihr Datenmaterial nicht aus der Labor- oder Feldforschung, sondern aus den Resultaten bereits vorliegender Primäruntersuchungen. Eine Sekundäranalyse hat also zum Ziel, aus den Forschungsergebnissen schon verfügbarer Untersuchungen (neue) Fragen zu beantworten, einen Überblick zu geben, was zu einem Thema bereits an Forschung existiert oder unabhängige Resultate zusammenzuführen, um allgemeingültigere Ergebnisse zu erhalten (Glass (1976)).

2.1 Verwandte Methoden der Sekundäranalyse

In Anlehnung an Hakim (1987) sollen hier drei Typen von Sekundäranalyse unterschieden werden. Diese weisen nicht nur unterschiedliche Methoden, sondern auch verschiedene Zielsetzungen auf. Ein Problem stellt die Namensgebung dar, die sowohl im Deutschen als auch im Englischen nicht immer eindeutig ist.

1. Beim „*Desk Research*“ werden zu einer bestimmten Problemstellung (z.B. aus dem Bereich der Marktforschung) alle verfügbaren Informationen aus amtlichen und halbamtlichen Statistiken, Verbandsstatistiken usw. zusammengetragen. Die Daten werden in eine vergleichbare Form gebracht (Unbrauchbares ausgeschieden, anderes skaliert, usw.). Die so verdichteten Daten können dann interpretiert werden. Im Unterschied zu den beiden folgenden Verfahren entstehen beim Desk Research wie bei Primäruntersuchungen Antworten auf *neue* Fragen.
2. Beim „*Research Review*“ geht es nicht darum, neue Fragen zu beantworten, sondern sich einen Überblick zu verschaffen über alle zu einer bestimmten Problemstellung bisher geleisteten, aus den verschiedensten wissenschaftlichen Disziplinen stammenden Arbeiten. Der Begriff „Research Review“ wird oft (z.B bei Bangert-Drowns (1986)) synonym mit dem eigentlich den Oberbegriff darstellenden Ausdruck „Sekundäranalyse“ verwendet. Ein großes Problem bei solchen narrativen Reviews ist die Subjektivität, mit der brauchbare von unbrauchbaren Studien getrennt werden und mit der die verschiedenen Arbeiten interpertiert werden (vgl. dazu z.B. Hakim (1987:18), Glass (1976:4) oder Fricke & Treinies (1985:14)).
3. Die dritte Gruppe von Methoden der Sekundäranalyse sind die „*Integrative Reviews*“, welche versuchen, mittels statistischer Methoden Generalisierungen (Allgemeingültigkeiten) zu speziellen Themen aus möglichst vielen empirisch-numerischen Untersuchungen abzuleiten. Das Ziel ist eine integrierte und quantifizierte Zusammenfassung der Forschungsergebnisse zu einem speziellen Forschungsgegenstand mit Berücksichtigung von statistischen Signifikanzen und Effektstärken. Ein Problem bei den „Integrative Reviews“ ist, daß sie auf Studien mit qualitativen Resultaten oder Studien, bei denen quantitative Daten nur einen Teil der gesamten Daten ausmachen, nicht anwendbar sind.

   Die Methode der „Metaanalyse“ gehört in die dritte Klasse von Sekundäranalysen.

2.2 Ursprung und Begriff

Der Begriff „Metaanalyse“ wird erstmals im Jahre 1976 von Glass verwendet. „Meta-analysis refers to the analysis of analyses. I use it to refer to the statistical analysis results of a large collection of analysis results from individual studies for the purpose of integrating the findings“ (Glass 1976:3). Die Vorsilbe „Meta“ soll ausdrücken, daß es sich dabei um einen Vorgang handelt, der auf sich selbst nochmals angewendet wird.

Die Idee, numerische Resultate aus empirischer Forschung mit statistischen Methoden (wie etwa der Kombination von Wahrscheinlichkeiten aus Signifikanztests) zusammenzufassen, datiert Bangert-Drowns (1986) in die Dreißigerjahre zurück. Damals wurde versucht, Ergebnisse aus landwirtschaftlichen Experimenten zu kombinieren. In

den Fünfzigerjahren wurden diese Methoden erstmals in den Sozialwissenschaften angewandt. Diese Versuche lösten aber die methodologisch vielfältigen Probleme der Integration und Akkumulation von Forschungsergebnissen über viele Studien hinweg nicht zufriedenstellend. Praktisch gleichzeitig mit Glass arbeiteten Schmidt & Hunter (1977) an einer metaanalytischen Methode zur Integration von quantitativen Ergebnissen aus Primärstudien. In der Folge entstanden weitere Varianten von Metaanalysen, z.T. von Kritikern der bereits publizierten Methoden (vgl. Bangert-Drowns (1986)).

Obwohl die Diskussionen um die verschiedenen Ausprägungen, Methoden, Anwendungsgebiete und Kritikpunkte noch nicht abgeschlossen sind, ist bezüglich der Ideen, die hinter der Metaanalyse stehen, und der groben Ziele, die damit verfolgt werden, Übereinstimmung festzustellen. Eine sehr weite und umfassende Definition, welche die gesamte Methodenfamilie der Metaanalyse einschließt, leitet z.B. Drinkmann (1990:11) her:

„*Metaanalyse soll sein: eine an den Kriterien empirischer Forschung orientierte Methode zur quantitativen Integration der Ergebnisse empirischer Untersuchungen sowie zur Analyse der Variabilität dieser Ergebnisse.*“

3. Einsatzgebiete und Nutzen der Metaanalyse im Bereich Personalwirtschaft

3.1 Anwendungsbereiche der Metaanalyse für personalwirtschaftliche Fragen

Die Metaanalyse ist ein Instrument, welches sich in der Arbeits- und Organisationspsychologie mittlerweile etabliert hat. Verglichen damit sind Metaanalysen im Personalmanagement noch wenig verbreitet. Im folgenden sollen deshalb einige personalwirtschaftliche Anwendungen skizziert werden, welche möglicherweise einen Impuls für neue Metaanalysen geben.

In diesem Beitrag wird (in Kapitel 4) die Metaanalyse am Beispiel der Personalauswahl illustriert. Die Methode eignet sich aber überall dort, wo bereits viele Primärstudien vorhanden sind und Zusammenhänge vermutet werden, welche nicht situationsspezifisch sind. Sie kann auch eingesetzt werden, wenn Faktoren untersucht werden sollen, welche vermutete Zusammenhänge beeinflussen. Diese Bedingungen sind bei Fragen des Personalmanagements oft erfüllt.

Die Messung der *Arbeitszufriedenheit* und deren Zusammenhänge mit dem *Absentismus*, mit der *Arbeitsleistung* oder der *Fluktuation* (und vielen weiteren Faktoren) sind Gegenstand zahlreicher empirischer Studien. Entsprechend ist es möglich, diese Zusammenhänge metaanalytisch zu erforschen. Eine der größeren Metaanalysen, welche den Zusammenhang von Arbeitsleistung und -zufriedenheit untersuchte, wurde von Iaffaldano & Muchinsky (1985) durchgeführt. Mitra et al. (1992) haben den Zusammenhang von Absentismus und Fluktuation metaanalytisch untersucht und so Erkenntnisse über das Rückzugsverhalten von Mitarbeitern gewonnen.

Von den betrieblichen *Lohnsystemen* wird neben der Abgeltung der Arbeitsleistung auch eine motivierende Wirkung erhofft. Die Wirkung von Lohnsystemen wurde empirisch untersucht. Wiersma (1992) hat beispielsweise in einer Metaanalyse nachgewiesen, daß zusätzliche extrinsische Belohnungen die Stärke der intrinsische Motivation (z.B. durch eine interessante Arbeit) nicht negativ beeinflussen.

Metaanalysen sind ebenfalls im Rahmen der Diskussion von *Führungstheorien* angezeigt, da mittels empirischer Studien versucht wird, einen Nachweis für deren Gültigkeit zu erbringen. Die Metaanalyse von Lord et al. (1986) belegt z.B., daß die in jüngerer Zeit eher gering geschätzte Eigenschaftstheorie den Führungserfolg erstaunlich gut erklären kann. Ebenso wurden Studien über Fiedlers kontingenztheoretischen Führungsansatz (Strube & Garcia (1981), Peters et al. (1985)) und über die Weg-Ziel-Theorie (Wofford & Liska (1993)) metaanalytisch ausgewertet - allerdings mit weniger eindeutigen Ergebnissen. Bemerkenswert sind auch die Metaanalysen zu Fragestellungen aus dem Gebiet Führung und Geschlecht (z.B. Eagly & Johnson (1990), Eagly et al. (1992)).

Die *Arbeitsgestaltung* und die *Personal- und Organisationsentwicklung* haben oft Produktivitätssteigerungen oder Kostensenkungen zum Ziel. Die Wirksamkeit von Maßnahmen in diesen Bereichen, sowie deren Situationsspezifizität kann ebenfalls metaanalytisch untersucht werden. Ziel ist es dabei, gesicherte Aussagen über die Übertragbarkeit des zu erwartenden Erfolgs der entsprechenden Maßnahmen zu machen. Fried (1991) untersuchte metaanalytisch verschiedene Methoden der Arbeitsanalyse im Zusammenhang mit Arbeitszufriedenheit und Leistung. Neumann et al. (1989) studierten mittels einer Metaanalyse die Auswirkungen von Programmen der Organisationsentwicklung und der Arbeitsgestaltung.

Diese kurze Zusammenstellung zeigt die breite Anwendbarkeit der Metaanalyse. Die Möglichkeiten sind aber längst nicht ausgeschöpft. Vielmehr liegen noch viele Daten aus Studien brach. Die Metaanalyse kann insbesondere dort einen entscheidenden Erkenntnisgewinn bringen, wo verschiedene Primärstudien zu widersprüchlichen Resultaten gelangt sind. Mit einer metaanalytischen Auswertung der verfügbaren Daten kann oft gezeigt werden, daß die widersprüchlichen Ergebnisse auf Artefakte, wie unzulängliche Stichprobengrößen oder systematische Einflüsse (Moderatoren, vgl. Anhang 1) usw. zurückzuführen sind.

3.2 Die Bedeutung der Metaanalyse für die Personalpraxis

Die Metaanalyse ist vorab ein Instrument der Forschung, direkte Anwendungen in der Praxis dürften selten sein. Hingegen können die Unternehmen von den Ergebnissen profitieren. Dank des Einsatzes von Metaanalysen ist es im Bereich der Personalauswahl zu einem eigentlichen Paradigmenwechsel gekommen: Es kann dank der Meta- und der sie begleitenden Moderatoranalyse gezeigt werden, daß die empirisch festgestellten unterschiedlichen Validitäten von Auswahlverfahren durch Moderatoren (vgl. Anhang 1) und statistische Effekte erklärt werden können. Die alte, über fünfzig Jahre gültige Auffassung, daß Ergebnisse von Validitätsstudien aufgrund ihrer Situationsspe-

zifität nicht übertragbar sind (vgl. z.B. Ghiselli (1966)), mußte aufgegeben werden.[1] Somit kann bei Instrumenten der Personalauswahl verläßlich festgestellt werden, für welche Stellenfamilien sich diese eignen. Das bedeutet, daß die einzelnen Betriebe nicht mehr eine eigene Validierungsstudie, sondern „nur" noch Arbeitsanalysen durchführen müssen.

Große Bedeutung haben die Ergebnisse von Metaanalysen auch für die monetäre Bewertung von Personalprogrammen (z.B. Arbeitsgestaltung, Lohnsysteme, Schulung, Personalentwicklung). Mit den Kenntnissen über die Situationsspezifizität, die Moderatoren und deren Wirkungen lassen sich Erfolgswirkungen von Programmen im Personalbereich besser schätzen (vgl. dazu aber Kapitel 5.5).

Ebenfalls können die Unternehmen ihre Daten von Personalinformationssystemen und Personalbefragungen besser auswerten. Die metaanalytischen Studien rund um die Thematik Arbeitszufriedenheit können zusammen mit den betrieblichen Daten dazu dienen, verläßliche Diagnosen zu stellen und die Wirkung möglicher Therapien zu beurteilen.

4. Vorgehensweise beim Erstellen einer Metaanalyse

Die Schritte bei der Durchführung einer Metaanalyse sind grob die folgenden: Definition der Fragestellung, Literatursuche- und Bewertung, statistische Auswertung, Korrektur störender Einflüsse und Interpretation der Ergebnisse. Ausführliche Anleitungen finden sich z.B. in Glass et al. (1981), Hedges & Olkin (1985), Hunter & Schmidt (1990) oder auf deutsch in Fricke & Treinies (1985).

Die folgenden Ausführungen sollen am Beispiel des (vereinfacht dargestellten) Verfahrens von Hunter & Schmidt den Ablauf einer Metaanalyse aufzeigen. Dieses hat zum Ziel, die in vielen Primäranalysen zu einem Thema gefundenen Korrelationen zwischen unabhängigen und abhängigen Variablen zusammenzufassen (Hunter & Schmidt (1990:43)), von störenden Einflüssen zu befreien und zu allgemeingültigeren Aussagen zu kommen. Die theoretischen Erläuterungen sollen mittels eines Beispiels aus dem Bereich der Personalauswahl von Funke et al. (1987) illustriert werden. Die zur Illustration verwendeten Teile des Beispiels sind jeweils kursiv hervorgehoben. In der Übersichtsgrafik in Abbildung 1 sind links die Kapitel angegeben, in welchen die entsprechenden Schritte beschrieben sind.

Abb. 1: Übersicht über den Ablauf einer Metaanalyse (nach Funke et al. (1987)) (: „Judgement Call", genaueres siehe Kapitel 5.5)*

---

[1] Schmidt & Hunter (1977:529) formulieren diesen grundlegenden Sachverhalt folgendermaßen: „Personnel psychologists have traditionally believed, that employment test validities are situation specific. This study presents a Bayesian statistical model which allows one to explore the alternate hypothesis that variation in validity outcomes from study to study for similar jobs and tests is artifactual in nature."

4.1 Formulierung der Fragestellung

Bei einer Sekundäranalyse muß die Fragestellung bereits zu Beginn präzis definiert sein. Erst so ist es möglich, mit der systematischen Suche nach Primärstudien zu beginnen. Da eine Metaanalyse ein statistisches Verfahren ist, kommen nur Untersuchungsgegenstände in Frage, die numerisch erfaßbar sind, d.h. in den Primärstudien durch Messungen analysiert wurden. Die Metaanalyse versucht dann einerseits, auf dieselben Fragestellungen wie in den Primärstudien zuverlässigere Antworten zu erhalten und andererseits, falls entsprechende Daten vorhanden sind, Einflußfaktoren (sog. Moderatoren, vgl. Anhang 1) zu finden, welche die Korrelation der Einzeluntersuchungen beeinflußt haben.

Im Beispiel von Funke et al. (1987) werden vier metaanalytisch untersuchbare Fragestellungen aufgelistet:

1) Können diagnostische Verfahren [bei der Personalauswahl] einen Beitrag zur Prognose wissenschaftlich-technischer Leistung im Bereich Forschung und Entwicklung leisten? Wie hoch kann die durchschnittliche Validität angesetzt werden?
2) Welche Untergruppen von Prädiktoren[2] erweisen sich als valide und wie hoch liegen diese speziellen Validitätskoeffizienten?
3) Welche Leistungskriterien erweisen sich als durch die Prädiktoren prognostizierbar?
4) Sind Moderatoren auffindbar, welche die Validitätskoeffizienten beeinflussen?

4.2 Primärstudiensuche und -bewertung

Das Ziel einer Metaanalyse ist es, möglichst alle Studien zum entsprechenden Thema zu finden. Während des Suchvorgangs sollte noch keinerlei Wertung oder Ausschluß von Arbeiten vorgenommen werden; dies erfolgt erst später. Wichtig ist, daß auch unveröffentlichte sog. „graue Literatur“ wie z.B. Institutsreihen, Dissertationen, Studien mit nichtsignifikanten Resultaten usw. berücksichtigt werden (vgl. dazu auch Kapitel 5.4). Es gibt unterschiedliche Strategien, wie die benötigten Forschungsresultate beschafft werden können. Die verschiedenen Wege führen über Datenbank- und CD-ROM-Recherchen, Verzeichnisse von Forschungsprojekten, Übersichtsartikel oder das Schneeballprinzip, bei dem man von Literaturverzeichnis zu Literaturverzeichnis in die Vergangenheit „zurückwandert“. Um unveröffentlichte Studien zu finden, muß vor allem auf Beziehungen zu anderen Forschern zurückgegriffen werden.

Funke et al. (1987) haben bei ihrer Analyse auf den verschiedensten Wegen rund 70 Studien eruiert, von denen aber etwa 20 (vor allem unveröffentlichte) nicht beschafft werden konnten.

Im nächsten Schritt geht es darum, die in die Analyse einzubeziehenden Studien anhand eines zuvor erstellten Kriterienkataloges auszuwählen. *In unserem Beispiel mußten folgende Bedingungen erfüllt sein:*

---

[2] Als Prädiktoren werden hier die unterschiedlichen Verfahren der Personalauswahl (Persönlichkeitstests, biographische Fragebogen, usw.) bezeichnet. Die Resultate aus diesen Verfahren werden mit den Kriterien verglichen, d.h. mit den späteren Leistungen am Arbeitsplatz, welche z.B. mittels Vorgesetztenurteil erhoben werden.

- Die Stichprobe durfte nur Naturwissenschaftler oder Ingenieure umfassen (Studien an Studenten und nichtakademischem Personal wurden ausgeschlossen).
- Als Prädiktoren waren psychologische Tests, Fragebogen oder andere eignungsdiagnostische Verfahren verlangt.
- Es mußten konkrete Kriterien wissenschaftlich-technischer Leistung oder Kreativität in Form von Leistungsmessungen oder Leistungsbeurteilungen vorliegen.
- Ebenso war erforderlich, daß jede Studie mindestens ein Effektmaß aufwies, das quantitative Angaben über die Relation zwischen Prädiktor und Kriterium (z.B. in Form eines Validitätskoeffizienten) enthielt. Dieses Effektmaß mußte genau spezifizierte methodische Mindestanforderungen erfüllen.
- Schließlich wurden eindeutige Doppelveröffentlichungen ausgeschlossen.

Angaben zum Vorgehen bei der Auswahl kann man der Metaanalyse von Bliesener (1992) entnehmen. Von den 423 bei ihm identifizierten Studien konnten 91 nicht beschafft werden. Von den verbleibenden 332 Arbeiten mußten 133 Studien ausgeschieden werden, weil sie den Kriterien nicht standhielten.

4.3 Kodieren der Studien und Berechnen der Effektstärken und der Varianz

Sind die verwendbaren Studien ausgemacht, müssen diese in eine vergleichbare Form gebracht werden. Der Zusammenhang zwischen einer abhängigen und einer unabhängigen Variable (z.B. der Zusammenhang der Leistung in einem Assessment Center und derjenigen an der späteren Arbeitsstelle oder der Zusammenhang der Arbeitszufriedenheit mit der entsprechenden Arbeitsleistung) wird Effektstärke genannt und z.B. als Korrellationskoeffizient ausgedrückt. Da in jeder der unterschiedlichen Primärstudien meist mehrere Effektstärken an der selben Stichprobe berechnet werden, müssen die Ergebnisse der Studien eine „Vorbehandlung“ über sich ergehen lassen. Am Schluß sollte nur eine Gesamteffektstärke je Studie vorhanden sein. In dieser Vorbehandlung werden die Effektstärken der von abhängigen oder identischen Stichproben erhobenen Untersuchungen gemittelt.

Werden auf eine Stichprobe mehrere Tests angewendet, so ist es notwendig, die Resultate von erfolglosen Untersuchungen, d.h. solchen mit nichtsignifikanten Ergebnissen zu erhalten, damit die Effektmittelung nicht verzerrt wird. Funke et al. (1987) haben in ihrer Analyse Stichproben gefunden, auf die bis zu 18 Tests angewendet wurden. Meist sind solche Studien aber nicht erhältlich,und es muß mit Schätzungen weitergearbeitet werden.

Zum Schluß dieser Vorbereitungen müssen die unterschiedlichen Effektstärke maße in ein einheitliches Maß umgerechnet werden.

In verschiedenen Metaanalysen, z.B. in jener von Baron-Boldt (1988) oder in jener von Bliesener (1992), findet sich auch noch eine Bewertung der methodischen Qualität der einzelnen Primärstudien nach genau vorgegebenen, validitätsmindernden Kriterien (z.B. „Bewußte oder unbewußte Ergebniserwartungshaltung“ oder „mangelhafte Prüfung der statistischen Voraussetzungen“ usw. (vgl. Fricke & Treinies (1985:58)). Diese Bewertung wird zusätzlich als Gewicht der jeweiligen Studie für die weitere Berechnung verwendet.

4.4 Artefaktkorrektur, Gesamteffekt und Erklärung der Residualvarianz (Moderatoranalyse)

Mit den nun vorbereiteten Ergebnissen aus den Einzelstudien kann mit den folgenden, hier vereinfacht dargestellten Berechnungsschritten der statistische Teil der Metaanalyse durchgeführt werden. Dazu werden zunächst Gesamteffekt und -varianz über alle Studien berechnet (Schritte 1 und 2). Dann wird versucht, statistische Artefakte, d.h. der Statistik innewohnende Störgrößen wie z.B. Stichprobenfehler oder Meßfehler, welche einen Einfluß auf die Varianz haben, zu eliminieren (Schritte 3 und 4). Ist die verbleibende Varianz noch groß, so weist dies auf systematisch wirkende Einflußfaktoren (sog. Moderatoren, vgl. Anhang 1) hin. In der dann notwendigen Moderatoranalyse werden für alle möglichen solchen Faktoren Subgruppen gebildet und diese einzeln neu metaanalysiert (Schritte 5 und 6).

Im folgenden werden diese Schritte der Einfachheit halber mit einem Lehrbuchbeispiel, das demjenigen von Cascio (1991:169) nachempfunden wurde und von der Metaanalyse von Funke et al. (1987) unabhängig ist, illustriert (vgl. auch Weinert (1987:391ff)). Die Korrelation r_i kann man sich als Effektstärke zwischen Prädiktor und Kriterium vorstellen.

Studie	1	2	3	4	5
Stichprobenumfang (N_i)	823	95	72	197	206
Korrelation (r_i)	0.147	0.155	0.278	0.329	0.20

K = Anzahl der Korrelationen = 5 i = 1,..., K

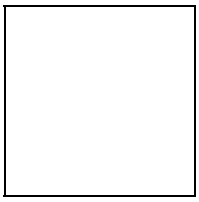
1) Gewichte die Effektstärke (Korrelation r_i) jeder Einzelstudie mit dem jeweiligen Stichprobenumfang N_i und berechne das Mittel $\bar{r}$ dieser gewichteten Effektstärken. Dies ergibt den mittleren Validitätskoeffizienten:

$$\bar{r} = \frac{\sum_{i=1}^{K} (N_i r_i)}{\sum_{i=1}^{K} N_i} = 0.188$$

2) Berechne die Gesamtvarianz σ_r^2 in den Effektstärken über alle Studien:

$$\sigma_r^2 = \frac{\sum_{i=1}^{K} \left[N_i (r_i - \bar{r})^2 \right]}{\sum_{i=1}^{K} N_i} = 0.0043$$

3)Bestimme die Größe der Varianz aufgrund des Stichprobenfehlers:



$$\sigma_e^2 = \frac{K(1-\bar{r}^2)^2}{\sum_{i=1}^{K} N_i} = 0.0033$$

Daraus läßt sich die Populationsvarianz, d.h. eine um den Stichprobenfehler korrigierte Gesamtvarianz bestimmen:

$$\sigma_p^2 = \sigma_r^2 - \sigma_e^2 = 0.0043 - 0.0033 = 0.0010$$ [3]

4) Bestimme die weiteren statistischen Artefakte und korrigiere entsprechend Mittel und Varianz.

   Beim Lehrbuchbeispiel ist dies aufgrund fehlender weiterer Angaben nicht möglich.

5) Vergleiche die korrigierte mittlere Varianz mit dem ursprünglichen Mittel der Varianz, um den Aufklärungsgrad der Artefaktkorrekturen zu bestimmen.

   Im Beispiel ist dieser $\frac{\sigma_e^2}{\sigma_r^2} = \frac{0.0033}{0.0043} = 0.77$, d.h. 77% der Gesamtvarianz wird erklärt.

6) Bleiben trotz der Artefaktkorrekturen große Abweichungen (d.h. weniger als 75% der ursprünglichen Varianz (= Faustregel, vgl. Hunter & Schmidt (1990:414)) kann durch Artefakte aufgeklärt werden), so identifiziere mögliche Moderatorvariablen, unterteile die Einzelstudien in Untergruppen (eine je Moderatorvariable) und führe für jede solche Subgruppe eine eigene Metaanalyse durch.
   Im Fall des Lehrbuchbeispiels ist dieser Schritt nicht nötig.

Bei der Metaanalyse von Funke et al. über die Eignungsdiagnostik in Forschung und Entwicklung wurden in den 50 auswertbaren Studien 89 verschiedene eignungsdiagnostische Verfahren insgesamt 142 mal eingesetzt. Die Stichprobengröße der Studien lag zwischen N=16 und N=769. Der Streubereich der Validitätskoeffizienten reichte von -0.05 bis +0.85.

Tabelle 1 zeigt die Ergebnisse der Metaanalyse. Daraus wird ersichtlich, daß die artefaktkorrigierte, mittlere Validität der Gesamtstichprobe 0.38 beträgt. Dies ist also der Schätzwert der generellen Prognostizierbarkeit wissenschaftlicher Leistung in Forschung und Entwicklung aufgrund „aller" bisherigen Untersuchungen. Allerdings fällt auf, daß die Varianz nach der Korrektur der Artefakte größer ist als zuvor. Dies läßt die Existenz von Moderatorvariablen vermuten.

---

[3] Hier kann es geschehen (wie etwa im Originalbeispiel von Cascio), daß die Populationsvarianz negativ wird, die ursprüngliche Varianz also „übererklärt" wird. In einer solchen Situationen darf gemäß Cascio als Näherung davon ausgegangen werden, daß die Varianz gleich Null sei. Für neuere Ansätze in der Varianzaufklärung vgl. Hunter & Schmidt (1994).

Tab. 1: Ergebnisse der Metaanalyse für die Gesamtgruppe und die Prädiktor-/Kriterienuntergruppen (aus Funke et al. (1987))

			ungewichtet		gewichtet		artefaktkorrigiert	
	Anz. unabh. Stpr	Stichpr.-größe	mittlere Validität	Varianz	mittlere Validität	Varianz	mittlere Validität	Varianz
Gesamt	48	6724	0.33	0.0393	0.33	0.0295	0.38	0.0305
Prädiktoren:								
Mehrdim. Persönl.-Tests	8	757	0.22	0.0045	0.21	0.0054	0.24	0.0000
Spezielle Persönl.-Tests	12	1065	0.24	0.0191	0.22	0.0172	0.25	0.0088
Motivationstests	14	1436	0.24	0.0579	0.26	0.0245	0.30	0.0206
Persönl.-tests gesamt	22	2296	0.26	0.0195	0.26	0.0174	0.30	0.0114
Intelligenztests	11	949	0.15	0.153	0.14	0.0148	0.16	0.0047
Kreativitätstest	15	848	0.28	0.565	0.26	0.0412	0.30	0.0335
Fachbez. Fähigk./Kreat.	10	764	0.27	0.0183	0.28	0.0130	0.32	0.0018
Biogr. Fragebogen	13	3297	0.46	0.0315	0.41	0.0250	0.47	0.0279
Kriterien:								
Vorgesetztenurteil	26	4020	0.37	0.0326	0.37	0.0236	0.44	0.0251
Gleichgestelltenurteil	13	987	0.30	0.0193	0.29	0.0203	0.34	0.0123
Ergebniskriterien	17	2953	0.35	0.0384	0.35	0.0266	0.37	0.0246

*Wie schon in der Fragestellung angekündigt, wurden einerseits sieben Kategorien von Prädiktortypen und drei Kriterien als Moderatoren behandelt und je getrennt metaanalysiert. Nach dieser Aufspaltung zeigt sich, daß die Artefaktkorrektur mit zwei Ausnahmen (*biographischer Fragebogen *und* Vorgesetztenurteil*) bei allen Prädiktoren und Kriterien zu einer verminderten Varianz geführt hat.*

Die beiden linken Spalten von Tabelle 2 geben Hinweise auf die Verläßlichkeit der korrigierten mittleren Validitätskoeffizienten. Die erste Spalte gibt denjenigen Wert an, oberhalb dessen 90% aller korrigierten Validitätswerte unter Zugrundelegung der korrigierten Standardabweichungen liegen. Er ist für alle Prädiktoren und Kriterien größer Null. Jede der Variablen hat damit mit 90% Wahrscheinlichkeit wenigstens minimale Validität, unabhängig von der speziellen Situation einer Primäruntersuchung. Diese Bedingung muß mindestens erfüllt sein, damit die mittlere korrigierte Validität (zweitletzte Spalte in Tabelle 1) generalisiert, d.h als von den Umständen einer Primäruntersuchung unabhängig betrachtet werden darf. Für alle Variablen in Tabelle 2 ist also Validitätsgeneralisierung gegeben.

*Tab. 2: Konfidenzintervalle und Varianzaufklärung nach Artefaktkorrektur (aus Funke et al. (1987)); * = p < 0.05*

	90%-Wahrsch'keit	90% Konfidenzintervall	Varianzaufklärung in % Stichprobenfehler	Unreliables Kriterium	Gesamt	χ^2-Test
Gesamt	0.16	0.09 - 0.67	19.2	2.5	21.7	249.80 *
Prädiktoren:						
Mehrdim. Persönl.-Tests	0.24	0.24 - 0.24	100	-	100	4.47
Spezielle Persönl.-Tests	0.13	0.10 - 0.40	49.3	1.8	51.1	20.23 *
Motivationstests	0.12	0.06 - 0.54	34.6	1.9	36.5	40.47 *
Persönl.-Tests gesamt	0.16	0.12 - 0.48	47.9	2.6	50.5	45.95 *
Intelligenztests	0.07	0.05 - 0.27	75.3	0.9	76.2	14.61
Kreativitätstests	0.07	0.00 - 0.60	37.3	1.1	38.4	40.19 *
Fachbez. Fähigk./Kreat.	0.27	0.25 - 0.39	85.3	4.0	89.5	11.69
Biogr. Fragebogen	0.26	0.20 - 0.74	10.9	4.5	15.4	119.10 *
Kriterien						
Vorgesetztenurteil	0.24	0.18 - 0.60	20.4	3.4	23.8	127.35 *
Gleichgestelltenurteil	0.20	0.16 - 0.52	54.4	2.4	56.8	23.88 *
Ergebniskriterien	0.17	0.11 - 0.63	16.7	0	16.7	102.01 *

Die 90%-Konfidenzintervalle der korrigierten mittleren Validitätskoeffizienten in der zweiten Spalte in Tabelle 2 stellen ein weiteres mögliches (restriktiveres) Kriterium für die Generalisierbarkeit dar: mit Ausnahme der Kreativitätstests *sind sie alle von Null verschieden und können somit generalisiert werden.*

*Da außer bei drei Prädiktoren (*mehrdimensionale Persönlichkeitstests, Intelligenztests *und* fachbezogene Fähigkeits- und Kreativitätstests*) weniger als 75% der Varianz mittels Artefaktkorrektur aufgeklärt werden konnte, wurden für den Gesamteffekt, die restlichen Prädiktoren und die Kriterien Moderatoranalysen durchgeführt. Die Signifikanzen des Chi^2-Homogenitätstests die in der letzten Spalte von Tabelle 2 dargestellt werden, bestätigen die Vermutung von Moderatorvariablen.*

Nach der bereits in der Fragestellung angekündigten und in Tabelle 1 und 2 dargestellten Analyse der Prädiktoren- und Kriterien-Moderatoren, wurden die in Tabelle 3 aufgelisteten Moderatorvariablen weiter untersucht.

Tab. 3: Moderatorvariablen mit mehreren, einzeln metaanalysierten Ausprägungen (aus Funke et al. (1987))

	Anz. unabh. Stpr.	Stichpr.-größe	gewichtet mittlere Validität	Varianz	artefaktkorrigiert mittlere Validität	Varianz	90%-W' keit	90%-Konf.int.
Korrelation	4	6074	0.31	0.0251	0.36	0.0252	0.16	.10- .62
Effektmaß	0	650	0.37	0.0701	0.42	0.0690	0.08	-.01- .85
Gruppendiff.	8							
>5	1	2144	0.17	0.0065	0.20	0.0000	0.17	.17- .17
Anzahl abh.	7	4580	0.40	0.0240	0.46	0.0209	0.27	.22- .70
Prädiktoren	3							
< 5	1							
Spezialverf.	1	2706	0.48	0.0158	0.55	0.0129	0.40	.36- .74
Spezifizität	4	3561	0.23	0.0146	0.26	0.0068	0.17	.12- .41
Standardverf.	3							
	3							
Staat	1	2411	0.31	0.0263	0.36	0.0256	0.16	.10- .62
Arbeits-	0	2385	0.34	0.0371	0.39	0.0316	0.16	.10- .68
bereich	2							
Industrie	7							
versch. Wiss.	1	3184	0.38	0.0273	0.44	0.0303	0.22	.15- .73
Berufsgruppe	5	1126	0.32	0.0774	0.37	0.0512	0.08	.00- .74
Ingenieure	1							
nur Naturwiss.	1	1492	0.31	0.0130	0.36	0.0088	0.24	24- .48
	1							
	5							

Für die in Tabelle 3 aufgeführten Moderatorvariablen wurde jede Ausprägung separat metaanalysiert. Das Kriterium für den Nachweis eines Moderators ist dabei das Absinken der beiden korrigierten Varianzen im Vergleich zur korrigierten Gesamtvarianz sowie die Differenz der korrigierten mittleren Validitäten. Die Anzahl abhängiger Prädiktoren *und die* Spezifität des Verfahrens *lassen sich als Moderatoren bestätigen. Daraus läßt sich folgern, daß höhere Validitäten dann resultierten, wenn pro Studie weniger Prädiktoren untersucht wurden und somit die Metaanalyse dank gezielterer Hypothesenprüfung weniger Effektmittelung erforderte (vgl. Funke et al. (1987:417)).*

Es wurden weitere mögliche Moderatorvariablen wie Qualität und Alter der Studie, Stichprobengröße *oder* Betriebszugehörigkeit *und* Berufserfahrung der Probanden *untersucht. Als einziger signifikanter Wert hat sich derjenige des* Alters der Probanden *herausgestellt. Da dieses aber nur in 16 Primärstudien angegeben wurde, ist die statistische Macht und somit die Zuverlässigkeit dieser Aussage gering.*

4.5 Interpretation der Ergebnisse

Nachdem der Ablauf und die Ergebnisse der Metaanalyse vorgestellt wurden, geht es jetzt darum, die wichtigsten Resultate zusammenzufassen, mit anderen Untersuchungen auf ähnlichen Gebieten zu vergleichen und auf spezielle, vielleicht unerwartete Erkenntnisse aufmerksam zu machen. Ein wichtiger Punkt dabei ist auch, auf mögliche

Schwachstellen der Studie, die schon den Daten, der Vorgehensweise oder dem Verfahren selbst anzulasten sind, hinzuweisen (vgl. dazu Kapitel 5).

Bei der vorgestellten Metaanalyse wird festgestellt, daß der korrigierte mittlere Validitätskoeffizient von 0.38 über die gesamte Studie die Vermutung stützt, daß Personvariablen prinzipiell einen bedeutsamen Beitrag zur Prognose individueller wissenschaftlicher Leistungen im Bereich Forschung und Entwicklung leisten können und daß das Ergebnis in der Größenordnung desjenigen eines Assessment-Centers liegt, welches allgemein als prognostisch valide gilt.

Im weiteren wird auf die Resultate der einzelnen Prädiktoren und Kriterien eingegangen und es werden die Auswirkungen der gefundenen Moderatoren zusammengefaßt. So wird z.B. festgestellt, daß Verfahren, die speziell für einen Einsatzzweck konstruiert wurden, wesentlich höhere Validitätskoeffizienten hatten als allgemein einsetzbare Standardtests. Daraus leiten die Autoren die Folgerung ab, daß die zukünftige Entwicklung solcher spezieller Verfahren vielversprechend in bezug auf die zu erwartenden Validitäten sein dürfte.

Streng genommen liefert das statistische Verfahren der Metaanalyse als Ergebnis nur Zahlenwerte wie Prozentangaben, Wahrscheinlichkeitswerte oder Konfidenzintervalle; all dies meist auf mehrere Stellen nach dem Komma genau. Ein Beispiel dafür sind die Tabellen 1 bis 3 in Kapitel 4.4. Daß solche Resultate einer Interpretation und Relativierung bedürfen, liegt auf der Hand. Daher sollen in Kapitel 5 die grundsätzlichen Kritikpunkte, die der Metaanalyse angelastet werden, aufgezeigt und diskutiert werden.

5. Kritik am Verfahren der Metaanalyse

Obwohl die Technik der Metaanalyse eine „Marktlücke“ im Angebot der Sekundäranalyse-Methoden gefüllt hat und im allgemeinen auch akzeptiert wird, bleiben doch diverse Punkte, an denen zum Teil heftige Kritik einsetzt. Eine Zusammenstellung von „Forty Questions about Validity Generalization and Meta Analysis“ mit entsprechenden Antworten und Repliken haben Schmidt et al. (1985) gemeinsam mit Sackett et al. (1985) zusammengetragen. Auch in den beiden Büchern von Hunter & Schmidt (1990) und von Glass et al. (1981) findet sich je ein Kapitel, das sich mit der Kritik am Verfahren der Metaanalyse auseinandersetzt. In der deutschsprachigen Literatur finden sich entsprechende Zusammenstellungen etwa in Drinkmann (1990) oder in Fricke & Treinies (1985).

Die vier Kritikpunkte „Abhängigkeit der Primärergebnisse“, „Apples and Oranges“-Problem, „Garbage In - Garbage Out“-Problem sowie der „Publication Bias“ resp. das „Filedrawer“-Problem werden fast von allen Autoren erwähnt.

5.1 Abhängigkeit der Primärergebnisse

Ein häufig auftretendes Problem ist die Abhängigkeit der Daten in den Primäranalysen. Ein gutes Beispiel dafür ist die im letzten Kapitel beschriebene Studie von Funke et al. (1987). Von den 52 verwendbaren Studien hatten nur 29 eine von allen anderen völlig unabhängige Stichprobe. Eine der Ursachen für solche Abhängigkeiten ist, daß

auf die Stichprobe(n) einer Studie oft mehrere Untersuchungen angewendet werden. Vielfach bilden derartige Studien die Grundlage für mehrere Publikationen, meist auch von unterschiedlichen Autoren. Bei der Literatursuche zu einer Metaanalyse werden nun alle diese „verschiedenen" Studien gefunden und deren identische oder zumindest abhängige Stichproben jedesmal als „unabhängig" neu zur Gesamtstichprobe addiert. Werden solche abhängigen Stichproben erkannt, was auf Grund mangelnder Angaben in den Primäruntersuchungen nicht immer möglich ist, müssen Korrekturverfahren eingesetzt werden.

Eine Methode liegt in der Gewichtung paralleler Ergebnisse, so daß die Stichprobe nur einmal mit dem Gewicht Eins und die an ihr erforschten Ergebnisse in der Summe ebenfalls mit Eins in die Metaanalyse eingehen. Eine andere Möglichkeit ist, nur ein Ergebnis der Studie resp. der an der Stichprobe vorgenommenen Studien zu übernehmen. Eine dritte Variante ist das Bilden von homogenen Subgruppen, in denen dann jeweils eine Studie nur mit einem Ergebnis vertreten ist. Diese Subgruppen werden dann getrennt analysiert.

Im Fall der Studie von Funke et al. (1987) wurden Teil- und Doppelveröffentlichungen mit Hilfe von Stichprobencharakteristika identifiziert und durch Effektmittelung zusammengefaßt. Doppelveröffentlichungen wurden ausgeschlossen. Zudem wurden zusätzlich zum Gesamteffekt die Effekte einzelner Subgruppen analysiert, wie dies oben als „dritte Variante" beschrieben wurde. Die dabei unterschiedlich auftretenden Abhängigkeiten in den Stichproben sind denn auch der Grund, warum sich in Tabelle 1 die Stichprobengröße der einzelnen Prädiktoren und Kriterien nicht zum Umfang der Gesamtstichprobe aufsummieren lassen.

5.2 „Apples and Oranges" - Problem

Der Metaanalyse wird vorgeworfen, sie integriere nicht vergleichbare Untersuchungen oder zumindest Studien mit nicht-identischen, sondern höchstens ähnlichen Fragestellungen und Vorgehensweisen.

Glass et al. (1981:218ff) argumentieren, daß es im Grunde genommen nur von der jeweiligen Fragestellung des Forschers abhänge, ob bestimmte Arbeiten metaanalytisch integriert werden dürfen. Fricke & Treinies (1985:170) zitieren Glass aus einer anderen Quelle: „Indeed the approach does mix apples and oranges, as one necessarly would do in studying fruits" und fügen noch hinzu: „Man kann es jedoch nicht der Methode ‚Metaanalyse' als Mangel ankreiden, wenn einzelne Anwender nicht begründen können, warum sie es für sinnvoll halten, gerade diese „Äpfel" und „Birnen" metaanalytisch zu integrieren".

Drinkmann (1990:24) führt als weiteres Gegenargument an, daß „Apples and Oranges"-Mischungen eigentlich in allen Formen von integrativen Reviews auftreten. Nur habe die Metaanalyse gegenüber traditionellen Review-Formen den Vorteil, daß Unterschiede in den Primäruntersuchungen nicht nur explizit gemacht werden, sondern prinzipiell auch kontrollierbar und vor allem analysierbar seien (Artefaktkorrektur, Moderatoranalyse, usw.).

5.3 „Garbage In - Garbage Out“ - Problem

Ebenso wird der Metaanalyse vorgeworfen, sorgfältig und aufwendig durchgeführte Studien von hoher methodischer Qualität und qualitativ „schlechte“ Studien würden mit gleichem Gewicht in die Analyse eingehen; das metaanalytische Ergebnis würde darunter leiden.

Um dieser Tendenz entgegenzuwirken, schlagen die meisten Autoren vor, die Einzelstudien einer numerischen Bewertung zu unterziehen. Dieser Wert gibt dann das Gewicht an, mit welchem die Primäruntersuchung in die Metaanalyse eingeht. Glass et al. (1981) widmen diesem Problem ein ganzes Kapitel. Das Problem jeder Bewertung sind aber die Kriterien und die Objektivität, mit der diese angewendet werden können (vgl. dazu Kapitel 5.5).

Durch diese Kodierung, resp. Gewichtung, und das abschließende „Herausrechnen“ der Varianz aus der Metaanalyse, soll die auch „schlechten“ Studien innewohnende Information entlockt werden, ohne daß diese das Gesamtresultat negativ beeinflussen. Glass wird in Fricke & Treinies (1983:171) zitiert, wie er den Ausspruch „garbage in - garbage out“ im Zusammenhang mit den Funktionen der Metaanalyse in „garbage in - information out“ umformuliert.

5.4 Der „Publication Bias“ oder das „File Drawer“ - Problem

Da eine Metaanalyse für sich in Anspruch nimmt, für ihren Analysezeitraum ein allgemeingültiges Resultat zu einem Untersuchungsgegenstand zu liefern, sollte sie theoretisch *alle* entsprechenden Ergebnisse von Primärstudien integrieren oder zumindest eine *repräsentative* Stichprobe aus der Gesamtheit aller Ergebnisse zusammenfassen. Praktisch kann dieser Forderung fast nie entsprochen werden. Einerseits ist es selten möglich, an alle identifizierten Studien heranzukommen, wie es im in Kapitel 4 dargestellten Beispiel der Fall war. Andererseits ist jedoch noch viel gravierender, daß die publizierten Arbeiten eine Verzerrung (bias) zugunsten signifikanter Ergebnisse aufweisen, da Primärstudien mit signifikanten Ergebnissen von Autoren und Herausgebern eher publiziert werden als solche mit unsignifikanten Ergebnissen; letztere landen normalerweise in der Schublade (file drawer).

Für die Lösung dieses Problems schlägt Rosenthal (1979) ein statistisches Verfahren vor, das es erlaubt, die Anzahl von Null-Ergebnissen (d.h. nichtsignifikanten Resultaten mit Korrelationskoeffizient r=0) zu schätzen, die zu einem metaanalytisch ermittelten Effekt hinzukommen müßten, damit dieser selbst an die Schwelle der Insignifikanz käme. Green & Hall (1984:47) verweisen auf eine Studie, in der 345 Primäruntersuchungen integriert wurden. Um deren Signifikanz in Frage zu stellen, bräuchte es nach dem oben beschriebenen Vorgehen rund 65‘000 Nullergebnisse. Bei kleineren und nicht sehr konsistenten Datenmengen braucht es jedoch relativ wenige Nullergebnisse, um die Signifikanz der Gesamtanalyse zu Fall zu bringen.

Ein Beispiel für eine Metaanalyse mit einer solch inkonsistenten Datensammlung erwähnen Wanous et al. (1989:263), bei der von den insgesamt 707 Korrelationen deren 528 (ca. 75%) aus einer unpublizierten Studie stammen. Durch das Weglassen dieser

einen Untersuchung verändert sich der Gesamteffekt von -0.07 auf -0.11. Es dürfte also sehr unwahrscheinlich sein, daß ein Wissenschaftler, der eine Metaanalyse zum selben Thema erstellt und die unveröffentlichte Studie nicht findet oder nicht auftreiben kann, zum identischen Ergebnis kommt. Eine Metaanalyse wird also um die fehlenden (oder einbezogenen) Daten verzerrt.

5.5 Weitere Kritik

Zu den vier oben beschriebenen Kritikpunkten, deren „Lösungen" als weitgehend akzeptiert gelten, gesellen sich mit zunehmender Verbreitung der Metaanalyse-Technik noch weitere. So haben bspw. Wanous et al. (1989) Metaanalysen von verschiedenen Autoren zum selben Thema verglichen und festgestellt, daß diese zu unterschiedlichen Ergebnissen kommen. Dies widerspricht aber der metaanalytischen Grundidee, ein von allen Einflüssen unabhängiges, jederzeit reproduzierbares Resultat zu liefern. Bei genauer Re-Analyse der verglichenen Metaanalysen kommen die drei Autoren zum Schluß, daß ein Großteil der Entscheide, die bei einer Metaanalyse gefällt werden müssen, Werturteile sind. So ist z.B. der Beschluß, welche Studien in eine Analyse mitaufgenommen werden und welche nicht, vom Urteil der jeweiligen Forscher abhängig und somit eine Quelle unterschiedlicher Endresultate. Neben dem „publication bias" existiert also auch noch ein „personal bias". In Abbildung 1 (Übersicht über den Ablauf einer Metaanalyse) sind diejenigen Schritte, die nach Wanous et al. (1989) mit einem Werturteil behaftet sind, mit einem Stern gekennzeichnet.

Weitere noch offene Fragen wirft Drinkmann (1990) auf. Sie beziehen sich auf den Umstand, daß die Metaanalyse allein Main-stream-Forschung unterstützt, für komplexere, ausgefallenere oder innovative Forschung jedoch kaum angewandt werden kann. So stößt die Metaanalyse bei qualitativen Forschungsmethoden oder solchen mit komplexen statistischen Auswertungsmethoden an ihre Grenzen. Ebenso kritisiert Drinkmann die „Rückwärtsorientiertheit" der Metaanalyse: Diese bindet potentiell progressiv einsetzbare wissenschaftliche Kapazität und leistet somit keine eigentlich „neuen" Ergebnisse, sondern bestätigt schon Dagewesenes.

Ein wichtiger Aspekt, der von verschiedenen Autoren erwähnt wird, ist die Frage nach der Zukunft der Forschung in Bereichen, in denen die Metaanalyse Fuß gefaßt hat. Die Motivation, auf metaanalytisch bereits „abgegrasten" Gebieten weitere Primärforschung zu betreiben, ist eher klein, da dank den Metaanalysen das „korrekte" Resultat ja ohnehin schon bekannt ist. Dazu kommt, daß es zur Zeit einfacher scheint, eine Metaanalyse publizieren zu können als eine Primäruntersuchung.

Ein weiteres Problem, das zwar nicht direkt dem Verfahren der Metaanalyse angelastet werden kann, das jedoch als Folge ihrer „präzisen" numerischen Resultate auftritt, ist dasjenige der unkritischen weiteren Verwendung dieser numerischen Werte. Das statistisch sicherste Resultat einer Metaanalyse ist ihr Hauptergebnis (z.B. der Validiätskoeffizient von 0.38 als Prädiktor von Personvariablen für die zukünftige Leistung im Bereich Forschung und Entwicklung (vgl. Kapitel 4.4 und 4.5)). Dieses stellt jedoch einen idealisierten, von allen störenden Einflüssen bereinigten Wert dar, wie er in der täglichen Praxis kaum zu erreichen sein wird. Als solcher Wert kann und sollte er in erster

Linie mit anderen mit der selben Methode ermittelten Resultaten verglichen werden. Mit anderen Worten, die Ergebnisse einer Metaanalyse sollten vor allem ordinal und höchstens zum Vergleich der Größeno*rdnungen* kardinal verwendet werden.

Analoges gilt auch für die Moderatoranalyse. Aufgrund des dabei erforderlichen Aufteilens der ursprünglichen Stichprobe in Subgruppen und des häufigen Fehlens der dazu notwendigen Detailinformationen wird die Datenbasis für allgemeingültige Aussagen oft zu klein, womit das ganze Verfahren wieder in Frage gestellt wird. Bei Funke et al. (1987) wurde zwar das Alter der Probanden als Moderator identifiziert, die Aussage stützt sich jedoch nur auf 16 der total 52 Untersuchungen (vgl. Kapitel 4.4).

Zwar eignen sich die Ergebnisse der Metaanalyse z.B. für die monetäre Bewertung des Erfolgs von Personalprogrammen. Allerdings müssen dazu die identifizierten Moderatoren auf jeden Fall mitberücksichtigt werden. Das Beispiel von Gerpott (1990) (er verwendet die metaanalytisch ermittelten Validitäten von traditionellen Interviews (TI) und Assessment Centers (AC) zur Berechnung der Ersparnisse beim Einsatz von AC‘s im Vergleich zum TI auf 100 DM pro Bewerber genau) zeigt aber, daß die Versuchung groß ist, nur das Hauptergebnis zu berücksichtigen und somit das Risiko in Kauf zu nehmen, u.U. sehr teure Fehler zu begehen.

6. Schlußbemerkungen

Die Methode der Metaanalyse hat sich in den letzten knapp zwanzig Jahren zu *der* Standardmethode bei den integrativen Sekundäranalysen entwickelt. Sie faßt mit statistischen Methoden die quantitativen Ergebnisse aus Primärstudien zusammen. Durch die Integration entstehen große Stichproben. Dank diesen ist es möglich, Artefakte (statistische Fehler, Meßfehler usw.) zu korrigieren und situationsspezifische Einflüsse (Moderatoren, vgl. Anhang 1) zu identifizieren. Auf diese Weise werden zuverlässige und von äußeren Faktoren unabhängige Resultate ermittelt.

Metaanalysen haben der Wissenschaft weitere wichtige Impulse geliefert:

- Die Moderatoranalyse (in Metaanalysen der Schmidt-Hunterschen Ausprägung) als Lieferantin von oft wichtigeren und interessanteren Resultaten als die der eigentlichen Metaanalyse.
- Das Aufzeigen von Lücken, in denen unbedingt noch Primärforschung notwendig ist.
- Das Etablieren von methodischen Standards, an denen sich künftige Primäruntersuchungen ausrichten können und sollen.
- Die Herausbildung von Qualitätskriterien, nach denen bereits existierende empirische Arbeiten zumindest auf deren Eignung zur Integration in Metaanalysen untersucht werden können.

Auch im Bereich der Personalforschung hat die Metaanalyse seit längerem Einzug gehalten. Es liegen aber noch viele geeignete Gebiete brach. Auf demjenigen der Personalauswahl hat sie jedoch - wie im Kapitel 3.2 beschrieben - zu einem eigentlichen Paradigmenwechsel geführt.

Dennoch ist, wie in Kapitel 5 gezeigt wurde, an gewissen Stellen Vorsicht geboten. Nicht nur Wanous et al. (1989) auch Schmitt et al. (1984) warnen vor blindem Ver-

trauen in die Resultate von Metaanalysen. Darum gilt es den Hinweis von Green & Hall (1984:52) stets zu beachten: „Statistical methodes, to be useful, must be used thoughtfully. Data analysis is an aid to thought, not a substitute."

Anhang 1 Definition und Wirkung von Moderatorvariablen

Eine Moderatorvariable[4] ist eine (Stör-)Größe, die den Zusammenhang (Korrelation r) zwischen Prädiktor und Kriterium beeinflußt. Wenn z.B. bei der Voraussage über den Berufserfolg im gleichen Test für Männer und Frauen unterschiedliche Werte resultieren, so ist das Geschlecht eine Moderatorvariable.

Abb. 2: Geschlecht als Moderatorvariable (nach Cascio (1991:283))

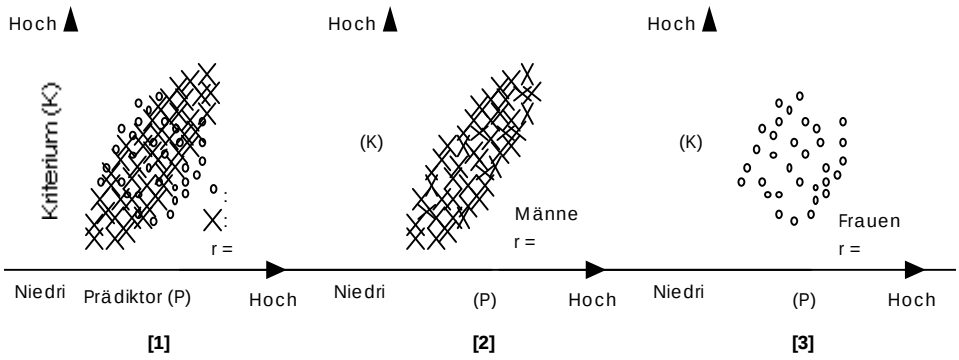


Diagramm [1] zeigt einen Gesamtzusammenhang zwischen Prädiktor und Kriterium von ca. r=0.50. Unterteilt man die ganze Gruppe jedoch in Männer und Frauen und betrachtet den Zusammenhang erneut, so wird klar, daß der Prädiktor für Männer relativ gut ist (Diagramm [2]), für Frauen jedoch nicht gebraucht werden sollte (Diagramm [3]).

Mit Hilfe solcher Moderatorvariablen (je nach Situation z.B. Geschlecht, Alter, Ausbildung, Hautfarbe usw.) können Untergruppen in einer Stichprobe identifiziert und im Falle einer Metaanalyse separat weiter untersucht werden.

Ein Problem der Moderatoranalyse ist, daß sie auf große Stichproben angewiesen ist, um nicht Gefahr zu laufen, selbst statistische Fehler zu produzieren.

Literatur

Bangert-Drowns, Robert L. (1986): Review of Developments in Meta-Analytic Method. Psychological Bulletin, 99, 388-499.

---

[4] Einen kurzen Einstieg vermittelt Cascio (1991:282); einen Überblick über die aktuelle Diskussion im Zusammenhang mit der Metaanalyse geben Sagie & Koslowski (1993).

Baron-Boldt, Jutta (1989): Die Validität von Schulabschlußnoten für die Prognose von Ausbildungs- und Studienerfolg: eine Metaanalyse nach dem Prinzip der Validitätsgeneralisierung. Frankfurt am Main, u.a.: Peter Lang.

Bliesener, Thomas (1992): Ist die Validität biographischer Daten ein methodisches Artefakt? Ergebnisse einer meta-analytischen Studie. Zeitschrift für Arbeits- und Organisationspsychologie, 36, 12-21.

Cascio, Wayne F. (1991): Applied Psychology in Personnel Management; 4.ed. London: Prentice Hall.

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

Eagly, Alice H. & Johnson, Blair T. (1990): Gender and leadership style: A meta analysis. Psychological Bulletin, 108, 233-256.

Eagly, Alice H.; Makhijani, Mona G. & Klonsky, Bruce G. (1992): Gender and the evaluation of leaders: A meta analysis. Psychological Bulletin, 111, 3-22.

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

Fried, Yitzhak (1991): Meta-analytic comparison of the Job Diagnostic Survey and Job Characteristics Inventory as correlates of work satisfaction and performance. Journal of Applied Psychology, 76, 690-697.

Funke, Uwe; Kraus, Julius; Schuler, Heinz & Stapf, Kurt H. (1987): Zur Prognostizierbarkeit wissenschaftlich-technischer Leistungen mittles Personvariablen: Eine Metaanalyse der Validität diagnostischer Verfahren im Bereich Forschung und Entwicklung. Gruppendynamik, 18, 407-428.

Gerpott, Torsten J. (1990): Erfolgswirkungen von Personalauswahlverfahren. Zur Bestimmung des ökonomischen Nutzens von Auswahlverfahren als Instrument des Personalcontrolling. Zeitschrift für Führung und Organisation, 59, 37-44.

Ghiselli, Edwin E. (1966): The Validity of Occupational Aptitude Test. New York: Wiley.

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

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

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

Hakim, Catherine (1987): Research Design: Strategies and Choices in the Design of Social Research. London: Allen & Unwin.

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

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

Hunter, John Edward & Schmidt, Frank L. (1991): Meta-Analysis. In Hambleton, Ronald K. & Zaal Jac N. (1991): Advances in Educational and Psychological Testing: Theory and Applications. Boston: Kluwer.

Hunter, John Edward & Schmidt, Frank L. (1994): Estimation of Sampling Error Variance in the Meta-Analysis of Correlations: Use of Average Correlation in the Homogenous Case. Journal of Applied Psychology, 79, 171-177.

Iffaldano, Michelle T. & Muchinsky, Paul M. (1985): Job satisfaction and job performance: Ameta-analysis. Psychological Bulletin, 97, 251-273.

Lord, Robert G.; DeVader, Christy L. & Alliger, George M. (1986): A meta-analysis of the relation between personality traits and leadership perceptions: An application of validity generalization procedures. Journal of Applied Psychology, 71, 402-410.

Mitra, Atul; Jenkins, G. Douglas & Gupta, Nina (1992): A meta-analytic review of the relationship between absence and turnover. Journal of Applied Psychology, 77, 879-889.

Neuman, George A.; Edwards, Jack E. & Raju, Nambury S. (1989): Organizational development interventions: A meta-analysis of their effects on satisfaction and other attitudes. Personnel Psychology, 42, 461-489.

Peters, Lawrence H.; Hartke, Darrell D. & Pohlmann, John T. (1986): Fiedlers Contingency Theory of Leadership: An application of the meta-analysis procedures of Schmidt and Hunter. Psychological Bulletin, 97, 274-285.

Rosenthal, Robert (1979): The ‚file drawer' problem and tolerance for null results. Psychological Bulletin, 86, 683-641.

Sackett, Paul R.; Schmitt, Neal; Tenopyr, Mary L.; Kehoe, Jerard & Zedeck Sheldon (1985): Commentary on Forty Questions About Validity Generalization and Meta-Analysis. Personnel Psychology, 38, 697-798.

Sagie, Abraham & Koslowski, Meni (1993): Detecting Moderators with Meta Analysis: An Evaluation and Comparison of Techniques. Personnel Psychology, 46, 629-640.

Schmidt, Frank L. & Hunter, John E. (1977): Development for a General Solution to the Problem of Validity Generalization. Journal of Applied Psychology, 62, 529-540.

Schmidt, Frank L.; Hunter, John E.; Pearlman, Kenneth & Rothstein Hirsh, Hannah (1985): Forty Questions About Validity Generalization and Meta-Analysis. Personnel Psychology, 46, 629-640.

Schmitt, Neal; Gooding, Richard Z.; Noe, Raymond A. & Kirsch, Michael (1984): Meta-Analyses of Validity Studies Published Between 1964 and 1982. Personnel Psychology, 37, 407-422.

Strube, Michael j. & Garcia, Joseph E. (1986): A meta-analytic investigation of Fiedlers contingency model of leadership effectiveness. Psychological Bulletin, 90, 307-321.

Wanous, John P.; Sullivan, Sherry E. & Malinak, Joyce (1989): The Role of Judgment Calls in Meta-Analysis. Journal of Applied Psychology, 74, 259-264.

Weinert, Ansfried B. (1987): Lehrbuch der Organisationspsychologie. 2. erw. Aufl., Weinheim: Psychologie Verlags Union.

Wiersma, Uco J. (1992): The effects of extrinsic rewards in intrinsic motivation: A meta-analysis. Journal of Occupational and Organizational Psychology, 65, 101-114.

Wofford, J. C. & Liska, Laurie Z. (1993): Path-goal theories of leadership: A meta-analysis. Journal of Management, 19, 857-876.