Zeitschrift für Arbeits- u. Organisationspsychologie (2014) 58 (N. F. 32) 4, 186 – 205

Forschungsmethoden

Metaanalyse – praktische Schritte und Entscheidungen im Umsetzungsprozess

Anna F. Schewe, Ute R. Hülsheger und Günter W. Maier

Zusammenfassung. In diesem Beitrag wird der Durchführungsprozess metaanalytischer Techniken nach Hunter und Schmidt (2004) Schritt für Schritt beschrieben. In Form eines Tutoriums geben wir evidenzbasierte Empfehlungen, verweisen auf relevante Quellen und Hilfsmittel und bewerten alternative Vorgehensweisen in Bezug auf ihre Güte und Akzeptanz. Wir wenden uns an Forschende, die eine Metaanalyse mit der Besonderheit der Artefaktkorrekturen nach Hunter und Schmidt durchführen wollen und geben Anregungen für die Verbreitung metaanalytischer Ergebnisse sowohl im Wissenschaftskontext als auch im Austausch mit Praktikern.
Schlüsselwörter: Psychometrische Metaanalyse, Anleitung, Methodologie

Meta-analyses: Practical steps and judgment calls along the way

Abstract. This article describes the process of conducting meta-analyses following the approach of Hunter and Schmidt (2004). For each step we present focal citations and examples of good practice, and compare distinct procedures for typical judgment call situations. We address researchers that seek practical guidance when conducting a meta-analysis with artefact corrections and suggest strategies for communicating meta-analytical results in science as well as practical contexts.
Key words: psychometric meta-analysis, tutorial, methodology

Die Menge an wissenschaftlichen Veröffentlichungen zu arbeits- und organisationspsychologischen Fragestellungen hat im letzten Jahrzehnt stark zugenommen. Während im Jahr 2003 in peer-review-Zeitschriften der Arbeits- und Organisationspsychologie (*AO-Psychologie*) 521 Artikel erschienen, waren es in 2013 1.686 Artikel (Datenbank PsycInfo; Classification Code 3600). Mit fortschreitender Nutzung der englischen Sprache für die Wissenschaftskommunikation stehen Anwender und Wissenschaftler vor der Aufgabe, immer mehr relevante Forschungsarbeiten zu einem Thema gleichzeitig zu identifizieren, zu beachten und zu interpretieren (vgl. Krampen, Fell & Schui, 2012).

Metaanalytische Methoden, also die statistische Integration von Effektstärken aller verfügbaren Forschungsarbeiten zu einer inhaltlichen Fragestellung, helfen bei dieser Herausforderung. Neben der effizienten Zusammenfassung und Integration von Forschungsergebnissen erlauben metaanalytische Methoden auch das Überwinden typischer Probleme und Validitätsgefährdungen von Primärstudien: Die oft geringe Stichprobengröße in durchgeführten Studien erschwert das Auffinden kleiner Effekte mit gängigen Signifikanztests und das gewählte Setting und Design schränken die Generalisierung der Ergebnisse ein. In Metaanalysen werden viele Effektgrößen unabhängig von der erreichten Signifikanz statistisch zusammengefasst und kommen mit einem über viele Studien hinweg gemittelten Wert dem allgemein gültigen Populationseffekt näher. Der Einfluss der Stichprobenfehler von Untersuchungen wird in einer Metaanalyse verringert, da die Effektgrößen anhand ihrer Präzision oder Stichprobengröße gewichtet in den metaanalytischen Mittelwert eingehen.

Aus praktischer Sicht reduzieren Metaanalysen die Vielfalt von Befunden auf eine leicht überschau- und interpretierbare quantitative Zusammenfassung aller vorliegenden Primärstudien und bilden damit die Basis für evidenzbasiertes Handeln und Entscheiden im professionellen Alltag.

Metaanalytische Methoden sind heute multidisziplinär anerkannt und verbreitet (vgl. Borenstein, Hedges, Higgins & Rothstein, 2009) und in der AO-Psychologie sowohl in thematischen als auch in Überblickszeitschriften (z. B. *Psychological Bulletin*, *Academy of Management Review* oder *Organizational Psychology Review*) zahlreich vertreten. Es gibt eine ganze Reihe metaanalytischer Auswertungsstrategien (z. B. Glass, McGaw & Smith, 1981; Hedges & Olkin, 1985; Rosenthal, 1991), die einen

DOI: 10.1026/0932-4089/a000165

unterschiedlichen Verbreitungsgrad in den Teildisziplinen der Psychologie haben. Innerhalb der AO-Psychologie ist der Ansatz der Validitätsgeneralisierung und psychometrischen Metaanalyse nach Schmidt und Hunter (1977; siehe auch Hunter & Schmidt, 2004) der verbreiteteste. Nach einer Überblicksarbeit, in der 196 Metaanalysen zu organisationalen Themen aus den Jahren 1982 bis 2009 berücksichtigt wurden (publiziert in vier einflussreichen Zeitschriften der Managementforschung und AO-Psychologie), wurden 84 % der Effektstärken mit der Schmidt-Hunter-(*SH*)-Methode ausgewertet (Aguinis, Dalton, Bosco, Pierce & Dalton, 2011). Das Alleinstellungsmerkmal der SH-Methode liegt in der Korrektur von Mess- und Stichprobenartefakten. Diese Artefaktkorrekturen sind auch in andere Ansätze integrierbar, stehen bei Hunter und Schmidt (2004) jedoch konzeptionell wie auch praktisch im Vordergrund. Obwohl aus Sicht vieler Statistiker Einschränkungen bezüglich theoretischer Annahmen der Methode bestehen (z. B. Algera, Jansen, Roe & Vijn, 1984; Hedges & Vevea, 1998), schneidet sie in Vergleichen mit anderen Ansätzen nicht unpräziser oder generell schlechter ab (Brannick, Yang & Cafri, 2011; Field 2001, 2005; Hall & Brannick, 2002; Schulze, 2004, 2007).

Bekannt im Bereich der AO-Psychologie wurde der Einsatz der SH-Korrekturen und ihrer Methode zunächst im Bereich der Validitätsgeneralisierung bei Prognose- und Erfolgskriterien der beruflichen Leistung und dem beruflichen Trainingserfolg (Schmidt & Hunter, 2003). Trotz der weiten Verbreitung der SH-Methode im Bereich der AO-Psychologie (Aguinis et al., 2011; Geyskens, Krishnan, Steenkamp & Cunha, 2009), werden im Gesamtbereich des Faches Psychologie nur ca. 10 % der publizierten Metaanalysen nach SH durchgeführt (Cafri, Kromrey & Brannick, 2010; Schmidt, Oh & Hayes, 2009). In vielen anderen Disziplinen (z. B. der Medizin und Biologie) ist die SH-Methode gar nicht verbreitet. Aus dieser großen Diskrepanz in der Anwendungshäufigkeit ergibt sich auch eine Herausforderung für alle, die methodische Hilfe zu Metaanalysen nach der SH-Methode suchen: Allgemeine Standardwerke (z. B. Borenstein, et al., 2009; Cooper, Hedges & Valentine, 2009; Hunter & Schmidt, 2004) legen Schwerpunkte auf theoretische und statistische Details, während die durchführungsorientierten Anleitungen (Field & Gillet, 2010; Lipsey & Wilson, 2001; Quintana & Minami, 2006; Rosenthal, 1995; Viechtbauer, 2010) wenig oder keinen Bezug auf SH-spezifische Fragen wie z. B. Artefaktkorrekturen nehmen.

Mit diesem Beitrag wollen wir diese Lücke schließen, indem wir eine Anleitung für die Durchführung einer SH-Metaanalyse geben sowie Entscheidungshilfen, Best Practice Beispiele und Evaluationskriterien (basierend auf Cafri et al. 2010; Dieckmann, Malle & Bodner, 2009; Geyskens et al., 2009) präsentieren. Grundlegende Vorkenntnisse über die Methode der Metaanalyse werden dabei vorausgesetzt. Lesern, die mit der Methodik noch nicht vertraut sind, empfehlen wir, sich zunächst mit Einführungstexten zu befassen (z. B. anhand von Field, 2005 oder Bortz & Döring, 2006), bevor sie sich unseren Durchführungshinweisen widmen. Wir schließen mit einem Ausblick auf weitergehende metaanalytische Methoden und Empfehlungen zur Verbreitung metaanalytischer Ergebnisse – nicht nur im Forschungskontext.

Praktische Durchführungsschritte

Im Folgenden beschreiben wir die Arbeitsschritte in der Durchführung einer SH-Metaanalyse wie in Abbildung 1 gegliedert. Wenngleich die Abbildung eine lineare Abfolge der Arbeitsschritte nahelegt, ist das Vorgehen meist eher zyklisch, da ausgehend von einem Zwischenergebnis ein früherer Schritt überarbeitet wird (z. B. von der Datenextraktion zurück zur Ergänzung des Kodiermanuals). Die Ausführungen beziehen sich sowohl auf Metaanalysen der *r*-Familie (Zusammenhänge von Konstrukten) als auch der *d*-Familie (Unterschiede zwischen Gruppen oder Interventionseffekte), wenn nicht abweichend angegeben.

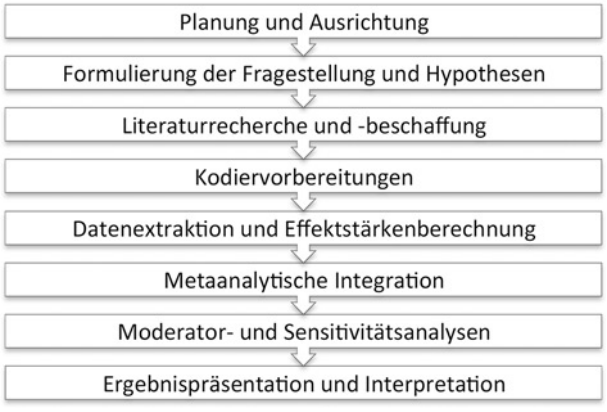


Abbildung 1. Praktische Schritte einer Metaanalyse.

Planung und Ausrichtung

Mit Hilfe metaanalytischer Techniken kann geklärt werden, (1) ob es einen signifikanten Effekt in der Population gibt, (2) wie hoch ausgeprägt, (3) wie variabel dieser Effekt ist und (4) ob die Unterschiede zwischen den Primärstudieneffekten größtenteils auf Stichprobenfehler und Artefakte zurückgehen oder ob bestimmte Merkmale der Stichprobe oder Studie die Größe des Effekts moderieren (vgl. Field & Gillet, 2010). In einer MA zum Zusammenhang zwischen Arbeitszufriedenheit und Leistung (Judge, Thoresen, Bono & Patton, 2001) beantworten die Autoren diese Fragen zum Beispiel so: (1) Ja, es gibt einen signifikanten Zusammenhang zwischen Arbeitszufriedenheit und Leistung, (2) dieser Effekt ist mit $\rho = .3$ mittelstark ausgeprägt, (3) die Effektstärke ist in den un-

tersuchten Studien unterschiedlich hoch ausgeprägt und damit variabel, (4) diese Unterschiede gehen z. T. darauf zurück, wie komplex die Arbeitstätigkeit ist.

Manchmal existieren schon Metaanalysen zu einer Fragestellung, aber es besteht Bedarf an einer Aktualisierung, der Überprüfung potentieller Moderatoren oder einer Verfeinerung der metaanalytischen Methoden (wie bei Van Iddekinge, Roth, Raymark & Odle-Dusseau, 2012). Neben dem Finden der Fragestellung ist in dieser Phase die Wahl eines Auswertungsansatzes relevant, damit dessen methodische Besonderheiten im Durchführungsprozess, z. B. bei der Kodierung oder Berechnung, beachtet werden.

Wahl des Auswertungsansatzes: Wie unterscheidet sich die SH-Methode von anderen Ansätzen?

Im Folgenden gehen wir auf drei zentrale Besonderheiten ein, durch die sich die SH-Methode deutlich von anderen unterscheidet und welche Konsequenz die Wahl der Methode für die Verbreitung innerhalb und außerhalb der AO-Psychologie hat.

Erstens gehen Hunter und Schmidt (2004) davon aus, dass in Metaanalysen Zusammenhänge auf der Ebene von Konstrukten untersucht werden sollen anstatt sie ausschließlich auf der Ebene der beobachteten Werte zu interpretieren. Deshalb werden sogenannte Studienartefakte definiert und korrigiert, die über den üblichen Stichprobenfehler hinausgehen (Schmidt & Hunter, 1996). Im Sinne der klassischen Testtheorie argumentieren die Autoren, dass sich der beobachtete Effekt einer Maßnahme oder eines Zusammenhangs aus dem „wahren“ Wert und Fehlereffekten zusammensetzt. Bei der Durchführung ihrer metaanalytischen Integration, auch psychometrische Metaanalyse genannt, werden typische Fehlerquellen isoliert und ihr Einfluss auf den Mittelwert und die Varianz des gemittelten Effekts bestimmt. So ist es möglich, für die mangelnde Messgenauigkeit zu korrigieren, mit denen Prädiktoren und Kriterien in Primärstudien erhoben wurden. Das Vorgehen erlaubt ebenfalls, den Einfluss von Varianzeinschränkung und Varianzerweiterung der zu Grunde liegenden Stichproben und die Erfassung kontinuierlicher Variablen in dichotomer Form zu korrigieren – falls ein Vorliegen dieser systematischen Verzerrungen zutrifft. Beim Vergleich zwischen um Artefakte korrigierten und unkorrigierten Effektstärken ist zu beachten, dass sich die Effektstärke systematisch vergrößert und die Varianz normalerweise abnimmt – je nach Stärke der vorgenommenen Korrekturen kann dies zu erheblichen Unterschieden in den berechneten Werten und damit in der Interpretation führen. Artefaktkorrekturen dürfen deshalb nicht leichtfertig vorgenommen werden, wenn z. B. Validitätsprobleme zu einer niedrigen Skalenreliabilität führen. Durch die Artefaktkorrektur würde dann die Effektstärke künstlich und fälschlicherweise erhöht.

Zweitens legt man sich mit der Methode von SH auf das Modell zufälliger Effekte fest. Metaanalytische Integrationsmodelle lassen sich danach unterscheiden, wie Varianzanteile innerhalb der integrierten Studien berechnet werden, mit welcher Gewichtung Studien in die Integration eingehen und inwiefern Ergebnisse auf leicht unterschiedliche Anwendungssituationen generalisiert werden dürfen. Das Modell des festen Effekts nimmt an, dass allen Studien derselbe Populationseffekt zu Grunde liegt, während das Modell zufälliger Effekte eine Verteilung von sehr ähnlichen, aber nicht gleichen Populationseffekten postuliert (Borenstein, Hedges, Higgins & Rothstein, 2010). Nach langjähriger methodischer Debatte sind sich Forscher zwischenzeitlich einig, dass für die angewandten Wissenschaften mit wenigen Ausnahmen das Modell zufälliger Effekte am angemessensten ist (National Research Council, 1992; Schulze, 2007). In Verbindung mit Moderatoren wird es auch Modell gemischter Effekte genannt. Da der SH-Methode das Modell zufälliger Effekte zugrunde liegt, können ermittelte Effektstärken auch auf zukünftige Situationen generalisiert werden, die mit denen der aufgenommenen Primärstudien vergleichbar sind.

Drittens kritisieren Hunter und Schmidt (2004) den Einsatz zweier gängiger Vorgehensweisen im Rahmen metaanalytischer Techniken. Dies sind die Transformation von Korrelationskoeffizienten vor der Integration in Z-Werte, die von den Autoren aufgrund einer positiven Verzerrung des Populationsmittelwertes kritisiert wird, und der Einsatz von Signifikanztests bei der Durchführung von Metaanalysen. Sie propagieren stattdessen das Bilden von Kredibilitätsintervallen, um Ergebnisse zu interpretieren. Beide genannten Punkte sind eher „Glaubensfragen“ der SH-Methode und führen im Regelfall nicht zu erheblichen Unterschieden in den Ergebnissen selbst oder der Bewertung dieser.

Bei der Wahl der Methode sollten aber nicht nur die inhaltlichen Unterschiede zwischen SH und anderen Methoden bedacht werden, sondern auch, ob die Arbeit vorwiegend im Bereich der AO-Psychologie oder ebenso darüber hinaus Verbreitung finden soll. Forscher außerhalb der AO-Psychologie und besonders außerhalb der psychologischen Disziplin (z. B. Medizin) greifen zumeist auf die metaanalytische Methodik von Hedges und Olkin (1985) zurück und sind daher mit SH-spezifischen Kennwerten wie z. B. dem Kredibilitätsintervall oder der sog. 75 %-Regel, wenig vertraut. Die Interpretation der Ergebnisse sowie die Evaluation des Vorgehens werden dann nicht nur für Leser, sondern auch für Gutachter erheblich erschwert. Auch herrschen innerhalb der Disziplinen gewisse Vorbehalte gegenüber den metaanalytischen Methoden, die von der jeweils anderen Disziplin vornehmlich angewendet werden: So wird außerhalb der AO-Psychologie häufig Kritik an der SH-Methode geübt, wohingegen innerhalb der AO-Psychologie von Lesern, Gutachtern und Herausgebern überwiegend die SH-Methode favorisiert wird. Es ist daher empfehlenswert, bei

der Wahl der Methodik auch die Zielgruppe einer Publikation mit zu berücksichtigen.

Literaturaufbereitung

Bei der anfänglichen Sichtung der Literatur zum Thema stehen vier Ziele im Vordergrund: Erstens sollten relevante Theorien, Modellannahmen, potentielle Prädiktoren, Korrelate, Konsequenzen und Moderatoren des zu bearbeitenden Konstrukts gesichtet werden. Das Ableiten theoretisch fundierter Annahmen für das metaanalytische Vorgehen erfüllt die Anforderungen deduktiven empirischen Arbeitens, erleichtert spätere Handlungsentscheidungen (z. B. bezogen auf Selektionskriterien) und bietet inhaltliche Vorlagen für das spätere Kodieren. Zweites Ziel ist, einen Überblick über die Befundlage, also Beschaffenheit und Anzahl von potentiell relevanten Primärstudien, zu erlangen um zu entscheiden, ob genügend Studien quantitative Daten berichten, ob vergleichbare Studiendesigns vorliegen, in welcher Effektstärke sich eine Integration anbietet und welches die üblichen Gefährdungen der Studienqualität im spezifischen Forschungsbereich sind. Drittens sollte überprüft werden, ob es schon Überblicksarbeiten, Reviews oder Metaanalysen zur Fragestellung gibt, um den Mehrwert der eigenen Arbeit zu bestimmen und ggf. die Vorgehensweise zur Orientierung und angeführte Literatur als Vorauswahl für sich zu nutzen. Implizit liefert dieser Schritt viertens meist die zentralen Quellen, die in relevanten Primärstudien i. d. R. zitiert werden. Diese Quellen können später für die systematische Literatursuche genutzt werden.

Wann ist eine Metaanalyse sinnvoll?

Erst nach dieser ersten Literatursichtung ist es möglich zu entscheiden, ob und in welcher Art eine Metaanalyse zum angedachten Thema sinnvoll erscheint. Zur kritischen Zahl von Primärstudien, die in eine Metaanalyse eingehen sollten, gibt es keine allgemeingültige Angabe (vgl. Borenstein et al., 2009, S. 357 ff.). Letztlich hängt es daher von der individuellen Fragestellung, Zielsetzung und dem Themenbereich ab, wie viele Studien eine Metaanalyse enthält. Das Beispiel einer Metaanalyse zu kausalen Beziehungen zwischen Arbeitseinstellungen und Leistung (Riketta, 2008) zeigt, dass auch eine Integration mit nur 16 Studien sinnvoll sein kann. Field (2001) demonstrierte jedoch, dass bei einer Studienanzahl unter 16 die Power der Metaanalyse signifikant abnimmt. Zu beachten ist, dass sich die Zahl der nach Prüfung des Titels und Abstracts relevanten Primärstudien bei Anwendung der Selektionskriterien noch einmal stark dezimieren kann (z. B. auf 30 %, vgl. Van Iddekinge et al., 2012).

Zu den Vorüberlegungen gehört auch die Frage, inwiefern die geplante Metaanalyse einen Beitrag für die Forschung oder Praxis liefern kann. Statt nur eine gemittelte Effektstärke zu liefern, sollte die Arbeit die Anforderungen eines Reviews erfüllen: Die vorhandene Literatur unter anderem Blickwinkel neu verknüpfen oder ordnen, zu sinnvollen Aussagen integrieren und zukünftig lohnende Forschungspfade ableiten (Humphrey, 2011). So unterscheiden Aguinis und Kollegen (2011) Metaanalysen etwa danach, ob sie gezielt eine existierende Theorie (oder Teile dieser) überprüfen (z. B. Ng & Feldmann, 2012) oder anhand der untersuchten Effektstärken und Merkmale der Primärstudien ein neues Modell anbieten (z. B. Kluger & DeNisi, 1996).

Formulierung der Fragestellung und Hypothesen

Welche Fragestellungen können mittels einer Metaanalyse untersucht werden? Chan und Arvey (2012) geben Beispiele für Zielsetzungen, etwa zur Klärung der Fragen (a) wie hoch die Korrelation zwischen zwei Variablen ist (z. B. Ilies, Nahrgang & Morgeson, 2007), (b) wie wirksam eine Intervention ist (z. B. Arthur, Bennett, Edens & Bell, 2003), (c) ob postulierte Theorien empirisch bestätigt werden (z. B. Fried & Ferris, 1987), (d) wie valide Messinstrumente oder Verfahren sind (z. B. Kinicki, McKee-Ryan, Schriesheim & Carson, 2002) oder (e) ob Drittvariablen einen Effekt moderieren (z. B. Wallace, Paulson, Lord & Bond, 2005). Nachdem geklärt ist, welche Erkenntnisse mit der geplanten Arbeit gewonnen werden sollen, können die Fragestellung formuliert und explizite Hypothesen auch zu potentiellen Moderatoren (z. B. O'Brien, Biga, Kessler & Allen, 2010) abgeleitet werden. Wie bei anderen empirischen Arbeiten müssen auch Hypothesen in Metaanalysen inhaltlich hergeleitet und begründet werden, z. B. mit zu Grunde liegenden Wirkmechanismen, Regulationsprozessen oder anderen Theorien. Zu diesem Zeitpunkt ist auch abzusehen, in welcher Art die Studienergebnisse größtenteils vorliegen, ob also eine Unterschiedshypothese (*d*-Effektstärke) oder Zusammenhangshypothese (*r*-Effektstärke) untersucht werden soll.

Festlegung der Selektionskriterien

Die Entscheidung, ob die Ergebnisse einer Primärstudie in die Metaanalyse eingehen oder nicht, wird mithilfe der Selektionskriterien so objektiv und replizierbar wie möglich gestaltet. Eine Grundvoraussetzung ist, dass in Studien eine Effektstärke oder ausreichend Informationen zur Berechnung berichtet werden. Weitere Kriterien können die Beschaffenheit der Stichprobe betreffen, das Studiendesign, die Operationalisierung der Untersuchungsvariablen oder andere Studienmerkmale. Dem häufigen Vorwurf an Forschende, in Metaanalysen Äpfel und Orangen zu mischen (vgl. Sharpe, 1997), also inhaltlich nicht vergleichbare Effekte zu integrieren, wird

durch eine vollständige Dokumentation der Selektionskriterien im späteren Artikel begegnet.

Literaturrecherche und -beschaffung

Um möglichst alle relevanten Forschungsarbeiten zu einer Fragestellung aufzufinden, werden zu Beginn dieser Phase anhand der Selektionskriterien Suchbegriffe (sowie deren Übersetzungen und Synonyme) definiert. Für eine gründliche und doch ökonomische Suche empfiehlt es sich, mehrstufig vorzugehen und zunächst Metadaten wie Titel, Abstract und Schlagwörter durchzusehen, bevor die Beschaffung der Quelle im Volltext angestoßen wird. Als Ergebnis stehen bei Abschluss dieser Phase die identifizierten potentiellen Quellen im Volltext zur Verfügung.

Suche in elektronischen Datenbanken

Die Suchmöglichkeiten in elektronischen Datenbanken erleichtern diese Aufgabe mit der Unterstützung von Booleschen Operatoren und Eingrenzungsfeldern. Ein Suchterm mit verknüpften Synonymen könnte etwa so gebildet werden (vgl. ADEPT, 2012):

„Im Titel/ Abstrakt: [(Persönlichkeit* OR personality OR „big five“) AND (Leistung* OR perform*)], Jahre: 1990 – 2013“ Durch die Verknüpfung ist eine zeitsparende und doch systematische Suche möglich, die auch von offenen Ressourcen wie google.scholar, scirus und worldwidescience unterstützt wird. Noch komfortabler wird die Suche allerdings mit der Verwendung von Fachdatenbanken wie PsycInfo, Psyndex, Social Web of Science und ggf. thematisch relevanten angrenzenden Datenbanken (z. B. ABI Inform, EconLit, WISO, SOC-Index, medline, ERIC, Education Research Complete). Bei diesen Abstraktdatenbanken sorgen Limitierungsfunktionen für einen besseren Ausschluss irrelevanter Quellen. Anhand englischer Suchterme werden in vielen Fachdatenbanken auch fremdsprachliche Artikel identifiziert. Diese sollten nicht kategorisch ausgeschlossen werden, sondern im Bestreben einer umfassenden Literaturrecherche nach Möglichkeit ebenso gesichtet werden.

Sowohl per Hand als auch in Datenbanken ist das rückwärtsgewandte Referenzverfolgen eine wertvolle Recherchestrategie. Dazu werden die Referenzen in einer thematisch besonders einschlägigen Arbeit nach relevanten Primärstudien durchsucht. Die vorwärts (im Zeitverlauf) gewandte Suche nach potentiellen Primärstudien ist dagegen nur elektronisch möglich, z. B. bei google.scholar, PsycInfo und SSCI. Basierend auf der Annahme, dass ein grundlegender Theorieartikel von allen folgenden empirischen Arbeiten zum Thema zitiert wird, kann man sich alle Arbeiten anzeigen lassen, die diese Quelle im Literaturverzeichnis führen.

Suche nach unveröffentlichter und grauer Literatur

Um dem Publikationsbias und der verzögerten Verbreitung von durchgeführten Forschungsarbeiten zu begegnen, ist es empfehlenswert, gezielt nach unveröffentlichter und grauer Literatur zu suchen. Eine mögliche Strategie ist das Durchsuchen von Abstraktbänden oder Datenbanken von Kongressen oder Tagungen ab einem bestimmten Zeitraum. Lohnenswert ist es, sich einen Überblick über die aktiv zum Thema forschenden Wissenschaftler zu verschaffen (falls die Anzahl einzugrenzen ist) und deren Websites nach relevanten Studien zu durchsuchen. Mit einer kurzen Beschreibung der geplanten Metaanalyse und dem Hinweis auf vertrauliche Behandlung der Daten kann direkt nach unveröffentlichten Studienergebnissen gefragt werden. Um noch mehr Personen zu erreichen, empfiehlt es sich, innerhalb des wissenschaftlichen Netzwerkes Kollegen gezielt aufzufordern, publizierte oder unpublizierte Studien für die Integration in die Metaanalyse zur Verfügung zu stellen. Dazu stehen Mailverteiler von Fachgruppen oder themenspezifische Mailverteiler, wie listserv, zur Verfügung. Auch wissenschaftliche Qualifikationsarbeiten sind für die Aufnahme in Metaanalysen geeignet, aufzufinden etwa über google.scholar, Dissertation Abstracts oder über die ZPID-Datenbank Diplomarbeiten. Mitunter ist es nötig, Autoren von Beiträgen, in denen relevante Angaben fehlen (z. B. bei unvollständigen Korrelationstabellen oder widersprüchlichen Aussagen zur Stichprobengröße), zu kontaktieren und um die fehlenden Angaben zu bitten.

Dokumentation der Suchergebnisse

Die Suche und das Verwalten der Literaturangaben mit dem aktuellen Bearbeitungsstand sollten systematisch erfolgen und dokumentiert werden, um den Vorgang später replizierbar berichten zu können. Gängige Literaturverwaltungsprogramme unterstützen den Forscher dabei mit Kategoriensystemen (z. B. zum Festhalten des Arbeitsstands, der Aufnahmeentscheidungen und Ausschlusskriterien) und Informationen zur Herkunft der Quellenangabe. Besonders übersichtlich gelingt die spätere summarische Dokumentation in Form von Flussdiagrammen, innerhalb derer verwendete Suchstrategien mit Datum und Trefferanzahl, aber auch die Häufigkeit der Exklusionsgründe aufgelistet werden (vgl. Lincoln, Suttner & Nestoriuc, 2008).

Eine häufige Unsicherheit besteht darin, wie weit zurückliegende Veröffentlichungen bei der Literatursuche beachten werden sollten. Generell umschließt eine umfassende Literaturrecherche Forschungsarbeiten jeden Alters. Das Publikationsjahr wird explorativ (z. B. Hülsheger, Maier, Stumpp & Muck, 2006) oder theoriegeleitet (z. B. Hülsheger, Anderson & Salgado, 2009) aber häufig als Moderator untersucht. Es gibt jedoch begründete

Ausnahmen, in denen ein Startjahr bei der Suche sinnvoll erscheint: etwa für Aktualisierungen schon publizierter Metaanalysen (z. B. Griffeth, Hom & Gaertner, 2000), bei der Neuausrichtung eines Konzepts nach Erscheinen eines theoretischen Grundlagenartikels (vgl. Nguyen & Ryan, 2008) oder ab der Veröffentlichung eines zentralen Messinstruments (Maier & Woschée, 2007). Weitere Ausführungen zu den Suchstrategien bei einer Metaanalyse finden sich z. B. im Herausgeberwerk von Cooper und Kollegen (2009, S. 49 ff.) oder mit spezifischen Ausführungen zu Suchtermen bei Arendt (2007).

Kodiervorbereitungen

In dieser Phase (die sich, wie auch die Kodierung selbst, meist zeitlich mit der Literaturrecherche und -beschaffung überschneidet) werden die Vorbereitungen getroffen, um alle relevanten Studienmerkmale objektiv und reliabel zu bestimmen und diese systematisch in einem für die Integration geeigneten Format aufzubereiten. Dazu gehört das Verfassen eines Kodiermanuals, das Erstellen einer Kodiertabelle oder Datenbank für die extrahierten Informationen und die Schulung weiterer Personen, die Studien kodieren werden. Die statistische Unabhängigkeit der Effektstärken, die in eine Metaanalyse eingehen, muss gewährleistet sein, d. h. jede Stichprobe darf nur eine Effektstärke zum Mittelwert beitragen (Details folgen im Abschnitt zur Kodierung). Daher muss jederzeit ersichtlich sein, aus welcher Stichprobe welche Werte in die Metaanalyse eingehen. Erreicht wird diese Anforderung mit der Vergabe von jeweils unabhängig voneinander vergebenen Identifikationsnummern (IDs) auf Studienebene, Stichprobenebene und Effektstärkenebene. Diese Ebenen sind hierarchisch aufgebaut, wobei es vorkommt, dass mehrere relevante Effektstärken in einer Stichprobe und mehrere Stichproben in einer Studie vorliegen. Letzteres ist kein Problem: Es dürfen mehrere Stichproben aus einer Untersuchung aufgenommen werden (da es sich um verschiedene Probanden handelt, sind die Effekte unabhängig). Zum Umgang mit multiplen Effektstärken aus einer Stichprobe sollte vorab geplant werden, ob im Rahmen der Metaanalyse ein metaanalytischer Mittelwert oder mehrere metaanalytische Mittelwerte bestimmt werden sollen. Hülsheger und Schewe (2011) haben in ihrer Arbeit diverse Variablen u. a. im Zusammenhang mit der Emotionsregulationsform Surface Acting betrachtet und dafür mehrere unabhängige metaanalytische Berechnungen auf Basis eines Primärstudienpools durchgeführt. In Bezug auf die Kodierung werden wir diese Arbeit im Folgenden als Beispiel zur Veranschaulichung nutzen.

Aufbau des Kodiermanuals

Mithilfe des Manuals wird der Kodierprozess objektiv und replizierbar; Es führt in Protokollform durch den Kodierprozess, expliziert Entscheidungen und listet zur Auswahl stehende Bewertungskategorien auf (s. Abb. 2 für den Auszug aus dem Manual der beispielhaften Metaanalyse). Aus dem Manual sollte ersichtlich sein, wie ein Studienmerkmal aufgrund seiner Ausprägung und seines späteren Verwendungszwecks in der Tabelle kodiert wird: als kontinuierliche oder kategoriale Variable oder als Freitext. Alle zur Berechnung des mittleren korrigierten Effekts benötigten Informationen (Effektstärke, N und ggf. Artefaktinformationen wie Reliabilitäten und Varianzratios) müssen als kontinuierlicher Wert vorliegen; für die Überprüfung von moderierenden Effekten können sowohl kontinuierliche als auch kategoriale Angaben genutzt werden. Freitextangaben dienen zum besseren Überblick, lassen aber auch a posteriori noch die gemeinsame Auswertung ähnlicher Konstrukte in einer Kategorie zu. Dies kann sinnvoll sein, wenn nicht genügend Effektstärken zu einzelnen Konstrukten gefunden wurden, und wird im Methodenteil der Arbeit berichtet. In der Beispiel-Metaanalyse (Hülsheger & Schewe, 2011) wurde z. B. festgelegt, dass in der Kategorie „organizational attachment" die Konstrukte organisationale Bindung und Kündigungsabsichten (revers kodiert) zusammengefasst wurden. Im Idealfall können gebildete Kategorien theoretisch oder praktisch belegt werden, oder die Nutzung vorhandener Taxonomien ist möglich. Häufig genutzte Kodierkategorien betreffen etwa die Clusterung von Nationen bezüglich Kontinenten oder Kollektivismus/Individualismus-Scores (Hofstede 2001), Berufskategorienbildung nach O*Net (Peterson et al., 2001), ISCO-08 oder 2010-SOC (U.S. Bureau of Labor Statistics, 2010), oder Industriesektoren (NAICS; Federal Government of the U.S., 2012).

Beim Kodieren im Team ist es sinnvoll, zu Beginn des Dokuments eine kurze Beschreibung der Forschungsfrage zu geben und die Selektionskriterien aufzuführen, wenn diese nicht schon im Rahmen der Literaturrezeption systematisch überprüft worden sind. Anschließend werden im Itemformat und unter durchlaufender Nummerierung Studieninformationen abgefragt, bei kategorialen Variablen mit Angabe der Antwortalternativen. Konventionell (s. Abb. 2; vgl. auch Lipsey & Wilson, 2001, Appendix E) erfolgt zunächst die Abfrage von Quellenmerkmalen zur Zuordnung der kodierten Informationen zu einem Schriftstück, die Vergabe einer Studien-ID und relevanten Studienmerkmalen sowie die Vergabe einer Sample-ID und Abfrage relevanter Stichprobenmerkmale. Auf Ebene der Effektstärke sind üblicherweise kodierte Merkmale die Höhe, Bestimmungsart und ggf. Ausgangswerte der Effektstärke, Operationalisierung und Messmethoden der abhängigen Variablen (aV) und unabhängigen Variablen (uV). Neben den hier genannten Merkmalen müssen jeweils themenspezifische Variablen, auch solche, die später als Einflussgröße auf die Höhe des Effekts untersucht werden sollen (Moderatoren), im Kodiermanual abgefragt werden.

Kodiervariablen: Studienmerkmale

a) Stud_ID
Jede Literaturquelle (= einzeln zugängliche Studie) erhält eine fortlaufend vergebene ID.

b) Jahr
Das Jahr der Erscheinung (publizierter Quellen) oder Erstellung (unpublizierter Quellen).

c) ...

Kodiervariablen: Stichprobenmerkmale

d) Sample_ID
Jede Stichprobe (= separat ausgewertete Probandengruppe) erhält eine fortlaufend vergebene ID.

e) N
Stichprobengröße (bei widersprüchlichen Angaben das N, auf dem die ES-Berechnung beruht oder aus Korrelationstabelle übernehmen).

f) % fem
Anteil Frauen in der Stichprobe in Prozent.

g) …

Kodiervariablen: Effektstärkenmerkmale

h) ES_ID
Jede Effektstärke erhält eine fortlaufend vergebene ID.

i) aV
mögliche Konsequenzen von Emotionsarbeit; hier die Kategoriennummer zuordnen.
Liegen sowohl Gesamt- als auch Facettenscores vor, werden Scores auf Ebene der genannten Kategorien bevorzugt, liegen mehrere Indikatoren einer Kategorie vor, werden nach HS composite scores gebildet:

1) emotional exhaustion(auch Burnout-Gesamtscore)	2) depersonalization
3) personal accomplishment	4) psychological strain
5) psychosomatic complaints	6) job satisfaction
7) organizational attachment (auch revers: Kündigungsabsicht)	8) task performance
9) emotional performance (friendliness, emotional display)	9) customer satisfaction

j) aV_Name
im Freitext wie in der Originalarbeit verwendet übernehmen

k) ES_r
Korrelationskoeffizient zw. uV und Emotionsarbeit-Facette eintragen (2 Nachkommastellen).

l) …

Abbildung 2. Auszug aus einem beispielhaften Kodiermanual korrespondierend zu Abbildung 3.

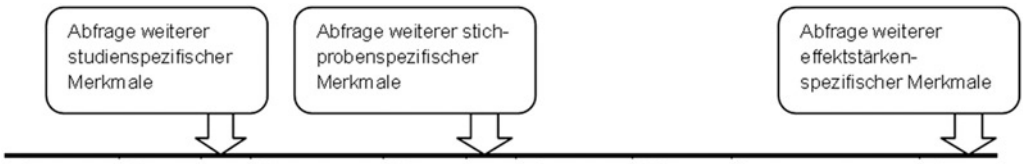
Kodiertabelle

In die Kodiertabelle werden die im Manual spezifizierten Informationen übertragen. Mit der aus Primärstudien vertrauten „flachen" Datenstruktur (jede Zeile entspricht einer Primärstudie) stößt man bei der Organisation der Daten später schnell an organisatorische Grenzen (vgl. Wilson, 2009). Um dem hierarchischen Aufbau der Daten gerecht zu werden (z. B. je drei Effektstärken genestet in zwei Stichproben genestet in einer publizierten Studie), bietet sich ein hierarchischer Aufbau der Datenstruktur an. Dies kann entweder in relationalen Datenbanken (z. B. MS Access) oder über eine sequentiellen Auflistung (s. Abb. 3) eines Tabellenkalkulationsprogramms (z. B. MS Excel) bzw. SPSS umgesetzt werden. Hier wird für jede kodierte Effektstärke eine Zeile ausgefüllt, wobei dann nicht einfach alle Effektstärken innerhalb der Tabelle gemittelt werden dürfen, sondern für die Kalkulation gezielt entnommen werden.

Im gezeigten Beispiel sind Kodierergebnisse aus drei Primärstudien abgebildet, die Zusammenhänge zwischen Surface Acting und möglichen Konsequenzen berichten. Zu allen Variablen-Kategorien (aufgeführt unter Punkt h der Abb. 2) waren jeweils eigenständige Metaanalysen geplant. Studie 1 enthält zwei unterschiedliche Stichproben, die den gleichen Zusammenhang berichten, wegen Unabhängigkeit der Daten aber beide in eine metaanalytische Auswertung eingehen dürfen. Zu Stichprobe 2 liegen jedoch zwei Effektstärken vor, die Unterfacetten von „organizational attachment" darstellen: sie wurden unter ES_ID 04 zusammengefasst, damit diese Stichprobe nicht mit doppeltem Gewicht in die Integration eingeht.

Der Kodierprozess

Für alle am Kodierprozess beteiligten Personen ist entsprechend dem Vorschlag von Stock (1994) ein Training und gleichzeitiger Optimierungsprozess erforderlich, bei



A	B		D	E		G	H	I	J
Stud_ID	Jahr	...	Samp_ID	N	...	ES_ID	aV	aV_Name	ES_r
01	1998	...	01	245	...	01	6	Job satisfaction	.26
01	1998	...	02	316	...	02	7_1	Emotional attachment	.34
01	1998	...	02	316	...	03	7_2	Turnover thoughts (rev.)	.32
01	1998	...	02	316	...	04	7	Comb: Org. attachment	.33
02	2005	...	03	81	...	05	7	Org. attachment	.4

Anmerkung: Erläuterung der Abkürzungen in Abbildung 2.

Abbildung 3. Ansicht einer hierarchischen Tabellenstruktur anhand von Kodieritems einer Metaanalyse zu Konsequenzen von Emotionsarbeit.

dem die Kodierer einerseits ein gemeinsames Verständnis des Manuals erarbeiten und gleichzeitig das Manual iterativ durch Probekodierungen und Feedbackschleifen optimiert wird. Kodieren mehrere Personen gleichzeitig, ist ein regelmäßiger Austausch bezüglich Ergänzungen oder Abweichungen vom Manual notwendig. Während des Kodierens sollte weiterhin auch das Flussdiagramm zur Literaturrezeption gepflegt werden, so dass nachvollziehbar ist, welche Studie aufgrund welcher Selektionskriterien ausgeschlossen wurde.

Interrater-Übereinstimmung

Die Qualität des Kodierprozesses sollte bei allen Metaanalysen berichtet werden; dies erfolgt i. d. R. über die Interrater-Übereinstimmung. Im besten Fall werden alle, alternativ aber wenigstens ein bestimmter Prozentsatz von Studien (in der Praxis meist 10 – 30 %; als Rat von Lipsey & Wilson, 2001: 20 – 50 Studien) von zwei Personen unabhängig kodiert und Reliabilitätswerte zu den ergebnisrelevanten Items berechnet. Dazu bieten sich Interrater-Maße an, die für zufällige Übereinstimmung korrigieren, etwa Cohens Kappa (Cohen, 1960) für kategoriale Variablen. Soll zusätzlich das Maß an Abweichung in die Reliabilität eingehen, kann das gewichtete Kappa verwendet werden (Cohen, 1968). Führt eine unbalancierte Zellenverteilung zur Verzerrung dieser Angaben bietet sich Andrés und Marzos Delta an (2004). Für die Reliabilitätsberechnung von kontinuierlichen Variablen werden verschiedene Intra-Klassen-Korrelationen (ICC) vorgeschlagen (Shrout & Fleiss, 1979), wobei in den meisten Metaanalysen der ICC 2.1 das passende Maß sein wird: zwei Personen aus einer Anzahl dazu fähiger Personen kodieren dieselben x Studien. Wenn die Voraussetzungen einer Normalverteilung verletzt werden, muss der ICC allerdings mit Vorsicht interpretiert werden (Jones, Johnson, Butler & Main, 1983), und Krippendorfs Alpha als flexible Alternative ist vorzuziehen (De Swert, 2012). Benchmarks zur Interpretation der Werte sind nicht speziell für den Bereich der Literaturkodierung vorhanden, so dass vor allem auf Daumenregeln zurückgegriffen wird: Landis und Koch (1977) etwa bezeichnen Cohens Kappa Werte ab .4 als moderat, ab .6 als substantiell und ab .8 als sehr gut. Für den ICC 2.1 geben Cicchetti und Sparrow (1981) Daumenregeln zur Evaluation der Werte an. Die kodierten Daten, bei denen keine Übereinstimmung vorliegt, werden dann im Team diskutiert und im Konsens bestimmt.

Datenextraktion und Effektstärkenberechnung

Trotz aller Bemühungen bspw. der APA zur Standardisierung des Berichtens empirischer Befunde (APA, 2009) lassen sich die erforderlichen Informationen in den Primärstudien oftmals nicht auf den ersten Blick und in der gewünschter Form identifizieren. Im folgenden Abschnitt werden Hilfsmittel und Vorgehensweisen präsentiert, um dennoch möglichst valide und vollständige Datensätze kodieren zu können. Eine kritische Entscheidung dieser Phase betrifft die Genauigkeit, mit der eine Effektstärke

bestimmt werden muss: Dürfen Schätzungen Eingang in die Metaanalyse finden? Werden d-Effektstärken kodiert, muss in Verbindung mit der konzeptionellen Studiendesignfrage außerdem entschieden werden, welche Art von Varianzbestimmung zur Berechnung des Effekts verwandt wird (Morris, 2008).

Bestimmung der Effektstärke

Enthält die Studie relevante Effektstärken als Werte in Tabellen oder Text (was bei Korrelationen meist der Fall ist), können sie direkt in die Kodiertabelle übertragen werden. Liegen stattdessen andere Angaben (Mittelwerte und Standardabweichung oder Teststatistiken) vor, sollten diese in die Tabelle eingehen und erst im zweiten Schritt die Effektstärke daraus berechnet werden. Die meisten Effektstärken können ineinander überführt oder aus univariaten Teststatistiken abgeleitet werden, z. B. mit Hilfe des Online Calculators von Wilson (2010). Wie multivariate Teststatistiken zur Bestimmung eines univariaten Effekts genutzt werden sollten, ist noch strittig. Während Peterson und Brown (2005; angewendet z. B. in Lux, Crook & Woehr, 2010) recht simple Schätzmethoden zur Bestimmung einer Effektstärke aus Beta-Koeffizienten vorstellen, schlagen Aloe und Becker (2012) einen detaillierten Index mit Variationen vor. Bernard und Abrami (2007) bieten eine dreistufige Präzisionshierarchie für die Ausgangswerte zur Bestimmung von Effektstärken an, die zur Bestimmung von Moderationseffekten oder für Sensitivitätsanalysen hilfreich ist. Sehr präzise können Effektstärken aus direkten Angaben von Mittelwerten und Standardabweichungen, Korrelationskoeffizienten oder äquivalenten statistischen Testverfahren berechnet werden. Als ausreichend präzise beurteilen die Autoren Ausgangwerte, anhand derer Effektstärken geschätzt werden können, wie etwa multivariate Teststatistiken. Als unpräzise klassifizieren sie Angaben wie „$p < .05$“.

Unabhängigkeit der integrierten Effektstärken

Um die Unabhängigkeit einzelner Werte zu gewährleisten, darf pro Stichprobe und metaanalytischer Effektbestimmung nur eine Effektstärke eingehen. Das Ausgangsmaterial aus Primärstudien liegt jedoch nicht immer einheitlich vor, etwa wird der Zusammenhang zwischen Burnout und Surface Acting in einer Studie mit drei Skalen untersucht und in Form von drei Korrelationen berichtet, in einer anderen Studie liegt jedoch nur eine Korrelation zum Gesamtscore Burnout vor. Um Effektstärken aus einer Stichprobe nicht mehrfach in eine metaanalytische Auswertung aufzunehmen, müssen im Kodiermanual Regeln zum Umgang mit abhängigen Effektstärken formuliert werden, z. B. welche Alternative präferiert wird, wenn Gesamtscore und Dimensionen vorliegen, oder ob Studien auszuschließen sind, die keine Dimensionswerte berichten (siehe Punkt i Abb. 2). Ein alternatives Vorgehen ist die Integration von Facetten eines Konstrukts zu zusammengesetzten Korrelationen. Sind Analysen auf verschiedenen Hierarchieebenen relevant empfiehlt es sich, auf der niedrigsten Hierarchiestufe (d. h. hier auf Ebene der Facetten) zu kodieren und später den Gesamtwert und die zugehörige gewichtete Reliabilität zu berechnen. Für die Berechnung von zusammengesetzten Korrelationen stellen Hunter und Schmidt (2004, Kap. 10) Formeln bereit, die neben der Effektstärke r auch die Berechnung eines gewichteten Reliabilitätswerts nach Mosier erlauben. Die vorgeschlagenen Gewichtungsmethoden von Cheung und Chan (2004) greifen geringe Unzulänglichkeiten der SH-Formeln auf und verbessern diese. Geyskens und Kollegen (2009) berichten über ihre Stichprobe von 69 Metaanalysen aus dem Management-Bereich, dass 54 % keine Aussagen zu dem Umgang mit abhängigen Effektstärken machten, danach waren das Bilden eines einfachen Mittelwertes (22 %) und die Berechnung zusammengesetzter Effekte nach Hunter und Schmidt (15 %) die verbreiteten Vorgehensweisen.

Sich überschneidende oder sogar gleiche Stichproben sind im Kodierprozess nicht immer einfach zu identifizieren, wenn sie in verschiedenen Quellen berichtet werden. Hinweise auf abhängige Stichproben geben die Namen der Autoren sowie die Anzahl und Geschlechtsverteilung der Versuchspersonen, auf deren Basis Wood (2008) eine standardisierte Prozedur für das Erkennen „doppelter“ Datensätze vorschlägt.

Metaanalytische Integration

Artefaktkorrektur nach Hunter und Schmidt

Hunter und Schmidt (2004) gehen davon aus, dass metaanalytische Techniken im Regelfall genutzt werden, um eine Annäherung an den „wahren“ Effekt zu berechnen, den Effekt auf Ebene der Konstrukte also, der möglichst frei von Verzerrungen und Messfehlern (Artefakten) ist. Da Primärstudienergebnisse immer mit Artefakten behaftet sind, die eine Abweichung vom wahren Populationsmittelwert bedingen, können und sollten diese im Laufe der metaanalytischen Integration korrigiert werden. Der unsystematische Stichprobenfehler (within study variation) geht darauf zurück, dass sich jede Stichprobe zufällig von einer möglichen anderen Stichprobenzusammensetzung unterscheidet. Er hat umso weniger Einfluss, je mehr Messwerte in einen Mittelwert eingehen, da deren Abweichungen vom wahren Populationseffekt in positiver und negativer Richtung sich aufheben. Indem die Effektstärken an der Stichprobengröße gewichtet werden, wird dieses Artefakt korrigiert: größere Stichproben weisen eine größere Präzision der Effektstärke auf und gehen deshalb mit mehr Gewicht in die Integration ein und bilden die um Stichprobenfehler korrigierte mittlere Effekt-

stärke. Auch die Varianz des mittleren Effekts wird korrigiert: die Gesamtvarianz lässt sich zerlegen in Populationsvarianz und auf Artefaktvarianzanteile. Wird der auf den Stichprobenfehler zurückgehende Varianzanteil von der Gesamtvarianz subtrahiert, erhält man die um diesen Fehler korrigierte Varianz des mittleren Effekts. Diese beiden Angaben (mittlerer Effekt und Varianz) bilden die Ausgangsbasis der SH-Metaanalyse und sind Ergebnisse der einfachsten Anwendungsart, der Bare-Bones-Methode. Der Stichprobenfehler ist das Artefakt mit dem größten Einfluss auf die Studienvariabilität (illustriert z. B. bei Schmidt, Pearlman, Hunter & Shane, 1979) und wird bei allen gängigen Verfahren standardmäßig verwendet.

Folgende Artefaktkorrekturen werden darüber hinaus laut der Untersuchung von Geyskens und Kollegen (2009) bei Metaanalysen im Management-Bereich häufig vorgenommen: Korrektur des Messfehlers in der abhängigen (aV) und unabhängigen (uV) Variable (49 % der untersuchten Metaanalysen) und direkte oder indirekte Einschränkung oder Erweiterung der Varianz (in aV oder uV; 22 % der Metaanalysen) sowie die Dichotomisierung von stetig verteilten Variablen (in aV und uV; 7 % der Metaanalysen). Sogenannte „einfache“, unabhängige Artefakt-Korrekturwerte werden bestimmt und in Form eines Korrekturfaktors *A* multiplikativ miteinander verknüpft, um dann die um Artefakte bereinigte mittlere Effektstärke und zugehörige Varianz zu erhalten.

Welche Art von Artefaktkorrektur für welchen Anwendungsfall?

Generell vertreten Hunter und Schmidt (2004) die Ansicht, so viele Korrekturen wie möglich anzuwenden, da der unverzerrte Zusammenhang auf Konstruktebene dem wahren Populationswert entspricht, den Forscher auch generalisieren wollen. Kritiker sehen aber gerade in einer routinemäßig angewandten Korrektur von Messartefakten eine gefährliche Scheinlösung für tatsächlich vorliegende Validitätsprobleme bei Messinstrumenten (Borsboom & Mellenberg, 2002). In der Praxis haben sich für verschiedene Forschungsfragen unterschiedliche Standards etabliert (vgl. Geyskens et al., 2009). Bei Validierungsstudien für Verfahren der Eignungsfeststellung etwa wird nicht für die Unreliabilität der uV korrigiert, dafür wird auf Form (direkt oder indirekt) und Höhe der Varianzeinschränkung (im Original: range restriction) geachtet.

Bei der Korrektur der Varianzeinschränkung soll berücksichtigt werden, dass in manchen Situationen die Varianz der uV eingeschränkt ist, weil nur für einen Teil der Gesamtstichprobe Daten für die aV vorliegen (Hunter, Schmidt & Le, 2006; Le & Schmidt, 2006). Im Fall der Personalauswahl liegen i. d. R. nur Erfolgskriterien für diejenigen Bewerber vor, die aufgrund ihrer guten Prädiktorwerte eingestellt wurden – die Varianz des Prädiktors ist demnach eingeschränkt, was die Höhe der Korrelation mit dem Kriterium negativ beeinflusst. Bei der Annahme der direkten Varianzeinschränkung geht man davon aus, dass die Varianz allein durch die Ausprägung des Prädiktors eingeschränkt wurde (Thorndike Fall II, Thorndike, 1949), während bei der indirekten Varianzeinschränkung angenommen wird, dass die Ausprägung des Prädiktors nur eine von mehreren Einflussgröße auf die Varianzeinschränkung war (Thorndike Fall III, Thorndike, 1949). Übertragen auf das Beispiel der Personalauswahl wird mit den beiden Formen der Varianzeinschränkung zwischen Situationen unterschieden, in denen die Auswahl allein auf den in der Metaanalyse untersuchten Prädiktor, z. B. Intelligenz in Form eines Intelligenztests zurückgeht (direkte Varianzeinschränkung) oder in denen sich die Bewerberauswahl auf mehrere Konstrukte stützt (z. B. Gewissenhaftigkeit, Intelligenz, Teamorientierung,…), von denen der in der Metaanalyse untersuchte Prädiktor (z. B. Intelligenz) nur einer ist (indirekte Varianzeinschränkung). Indirekte Varianzeinschränkung liegt auch vor, wenn die Auswahl der Bewerber zwar nicht auf dem zu untersuchenden Prädiktor (z. B. Intelligenz) beruht, das genutzte Auswahlkriterium aber mit dem zu untersuchenden Prädiktor korreliert ist (z. B. Abschlussnoten).

Über diese Einteilung hinausgehend haben Sackett und Yang (2000; aktualisiert von Yang, Sackett & Nho, 2004) eine erweiterte Typologie mit Korrekturformeln erstellt, in der sie Szenarien u. a. danach unterscheiden, ob die Selektion anhand des Prädiktors, des Kriteriums oder einer Drittvariable getroffen wird und für welche Gruppen Prädiktorwerte und Varianzangaben bekannt sind. Van Iddekinge und Ployhart (2008) geben bezogen auf den Bereich der Personalselektion ebenfalls einen Überblick über empfehlenswerte Korrekturstrategien. Ein Problem fehlender Daten tritt bei der Korrektur um Varianzeinschränkung häufig auf: für die Korrekturformeln wird die Standardabweichung der Prädiktorwerte sowohl *vor* als auch *nach* der Selektion benötigt, um das Selektionsverhältnis *Ux* zu bestimmen (Hunter et al., 2006). Da eine uneingeschränkte Prä-Selektions-Messung allerdings selten ist, werden oftmals Populationsnormen des Prädiktors herangezogen. Auch wenn dieses Vorgehen nicht ohne Kritik bleibt, zeigen empirische Studien, dass diese Normvarianzen akzeptable Schätzer für z. B. Persönlichkeitsskalen (Ones, Viswesvaran & Schmidt, 2003) und Fähigkeitstests (Hoffmann, 1995; Sackett & Ostgaard, 1994) bilden.

Beinhalten die einzelnen Primärstudien genug Informationen, um Werte für die Artefaktkorrektur zu bestimmen, sollte jede Effektstärke um den spezifischen Betrag des Artefakts in dieser Studie korrigiert werden. Bei diesem Vorgehen (*individual artifact correction*) können einzelne fehlende Artefaktangaben durch den Mittelwert der betreffenden Artefaktverteilung ersetzt werden (z. B. auch bei 30 % fehlender Angaben, vgl. Judge, Piccolo, Podsakoff, Shaw & Rich, 2010). Können Werte für die

Artefaktausprägung jedoch nur in einer Minderheit der Studien bestimmt werden, wird der Anspruch aufgegeben, jede Effektstärke studienspezifisch zu korrigieren. Stattdessen werden aus den jeweils vorhandenen Angaben einer jeden Artefaktart Verteilungen gebildet, deren Mittelwert dann je einen Korrekturfaktor darstellt. Sind gar keine Artefaktangaben aus Primärstudien zugänglich, ist es üblich und zulässig, diese aus anderen Forschungsarbeiten oder Manualen zu übernehmen. Oft herangezogene meta-analytisch bestimmte Reliabilitätskoeffizienten sind z. B. die Reliabilität der Ein-Item-Abfragen zu Arbeitszufriedenheit oder Arbeitsleistung ($r = .7$; Wanous & Hudy, 2001) oder die Validität von berufsbezogener Leistungseinschätzung durch Vorgesetzte ($r = .52$) und Kollegen ($r = .42$; beide Werte aus Viswesvaran, Ones & Schmidt, 1996).

Wahl der Software

Welche Software für die Berechnung einer Metaanalyse besonders geeignet ist, hängt von den spezifischen Zielsetzungen und dem Umfang der Auswertung ab. Ein von Schmidt und Le vertriebenes Softwareprogramm[1] (2004; rezensiert von Roth, 2007) entspricht exakt dem Vorgehen von Hunter und Schmidt (2004) und erlaubt komfortabel die Artefaktkorrektur der Effektstärken (auch um indirekte Varianzeinschränkung). Alternativ dazu stehen Basisrechnungen in Skripten oder Makros für SPSS, Stata oder Excel-Blättern zur Verfügung (Wilson, 2010). Sollen der Mittelwertintegration keine umfangreichen Sensitivitätsanalysen folgen, reichen alle bisher angesprochenen Möglichkeiten aus. Weiterführende Analysen oder die Erstellung von Grafiken lassen sich damit allerdings nur eingeschränkt durchführen, so dass man im Baukasten-Stil auf verschiedene Anwendungen für spezifische Analysen zurückgreifen muss.

Für das Programm SAS liegen detaillierte Skripte inklusive Ausreißer-Überprüfung und Moderator-Modellierung vor (Arthur, Bennet & Huffcutt, 2001). Darüber hinaus besteht die Möglichkeit, vorhandene Skripte an den eigenen Bedarf anzupassen, z. B. das statistisch sehr ausgereifte und flexible R-Paket *metafor* (Viechtbauer, 2010) zu nutzen und die nicht standardmäßig vorgesehene Artefaktkorrektur selbst zu ergänzen (z. B. in Goertz, Hülsheger & Maier, in review). Sind keine Artefaktkorrekturen vorgesehen stehen deutlich mehr Software-Alternativen zur Auswahl (für einen Vergleich siehe Bax, Yu, Ikeda & Moons, 2007). Comprehensive Meta Analysis (Borenstein, Hedges, Higgins & Rothstein, 2005; rezensiert von Pierce, 2008) bietet als kostenpflichtiges Programm z. B. deutlich mehr Komfort und Auswertungsoptionen als andere Alternativen, erlaubt aber weder Artefaktkorrekturen noch Eingriffe in die Berechnungsformeln, die für eine längsschnittliche Effektbestimmung notwendig wären. Die Möglichkeiten, auf Formeln der Effektberechnung und -integration selbst Einfluss zu nehmen, sind bei kommerziellen Programmen generell eingeschränkt, bei Anwendungen mit austauschbaren Skripten (z. B. R-Project, SPSS, SAS) flexibel anpassbar.

Signifikanz und Generalisierbarkeit des Effekts

Mit den zu errechnenden Ergebnissen sollen die Ausgangsfragen der Metaanalyse beantwortet werden: Erstens, gibt es einen bedeutsamen Effekt in der integrierten Stichprobe von Studien, der von Null verschieden ist? Nach SH-Vorgehen wird das Konfidenzintervall, das Null nicht enthalten darf, zur Überprüfung der Signifikanz herangezogen. Dieses Intervall des Standardfehlers um den Mittelwert (meist auf 95 % Niveau) wird vor der Korrektur des Stichprobenfehlers gebildet und gibt Auskunft über die Sicherheit, mit der das Vorliegen eines Effekts angegeben werden kann (Whitener, 1990). Zweitens, wie ist die Größe des Effekts zu bewerten? Die mittlere, um Stichprobenfehler korrigierte Effektstärke (r oder d) und die zusätzlich um Artefakte korrigierte Effektstärke (ρ [= roh] oder Δ [= delta]) lassen sich nach Cohens Daumenregeln (1988; klein, mittel und groß) beurteilen, sollten jedoch im Zusammenhang mit vergleichbaren Konstrukten oder Interventionen später diskutiert werden. Drittens, kann dieser in der Stichprobe gefundene Effekt auf die Population aller vergangenen und zukünftigen Studien zum gleichen Zusammenhang oder Unterschied generalisiert werden? Ist die metaanalytisch generierte mittlere Effektstärke also ein akzeptabler Schätzer für Studien außerhalb der aufgenommenen Primärstudien? Der Anlehnung an Bayesische Modelle und dem Verständnis als Modell zufälliger Effekte folgend wird im SH-Ansatz ein zusätzliches Intervall des um Artefakte korrigierten Standardfehlers um den Mittelwert gebildet. Dieses Kredibilitätsintervall (konventionell auf 80 % Niveau) wird zur Beurteilung der Homogenität der integrierten Primärstudien genutzt und gibt an, in welchen Grenzen sich 80 % der Populationseffekte der zugrundeliegenden Verteilung von Primärstudien befinden (vgl. Whitener, 1990). Ist dieses Intervall weit (Daumenregel: weiter als $r = .11$; Koslowsky & Sagie, 1993), sollte, wie im nächsten Abschnitt beschrieben, nach moderierenden Einflüssen gesucht werden, um die heterogene Gesamt-

[1] Bei der Verwendung des Programms sind einige Besonderheiten zu beachten: in deutschen Sprachregionen funktioniert die Darstellung nicht, deshalb sollte die Regionseinstellung auf ein englischsprachiges Land geändert werden. Das Programm verarbeitet keine Null-Effekte, stattdessen sollte 0.001 eingegeben werden. Bei der Ergebnisausgabe können Varianzanteile z. T. negative Werte annehmen, oder Varianzanteile über 100 % werden erklärt; diese Werte sollten auf 0 bzw. 100 % gesetzt werden (zur weiteren Erklärung siehe Steel und Kammeyer-Mueller, 2002). Die Eingabe ist mühevoll, bei größeren Datenmengen bietet sich „Copy and Paste" der Daten aus vorhandenen Tabellen und die Direkteingabe über die txt-Dokumente am Speicherort des Programmes an. In ihrem Standardwerk von 2004 geben Hunter und Schmidt detaillierte Anweisungen zur Handhabung des Programms.

stichprobe in homogenere Subgruppen zu teilen, die einen mittleren Effekt mit geringerer Varianz aufweisen. Enthält das Kredibilitätsintervall Null, sollte der Effekt nicht generalisiert werden, auch wenn ein signifikanter mittlerer Effekt errechnet wurde (Whitener, 1990). Die zugrunde liegende Verteilung von Primärstudieneffekten enthält dann zu verschiedene Subpopulationen, um eindeutig auf bedeutsame Effekte in zukünftigen Situationen schließen zu können. Andere Ansätze nutzen zur Klärung dieser Frage den Q-Test (Viechtbauer, 2007b). Viertens, wie genau kann die Schätzung der mittleren Effektstärke geleistet werden, bzw. wie variabel ist der Effekt in der Studienpopulation? Dazu geben mehrere Kennwerte Auskunft. Ein spezifisch nach SH-Ansatz berechneter Wert, neben dem Kredibilitätsintervall, betrifft den Anteil der durch Stichprobeneffekte und andere Artefakte erklärten Varianz an der Gesamtvarianz. Je höher der Wert, desto weniger Varianz zwischen den integrierten Studien geht auf „wahre" Effektstärkenunterschiede zurück und desto mehr Generalisierbarkeit über Situationen hinweg ist gegeben. Eine Daumenregel (Hunter & Schmidt, 2004) besagt, dass 75 % erklärter Varianz auf eine homogene Stichprobe hinweisen, bei einem geringeren Wert bedingen moderierende Variablen Heterogenität zwischen den Studieneffekten, was die Interpretation des Effekts erschwert. Metaanalyse-Ansätze anderer Autoren geben zur Bewertung der Stichprobenhomogenität I^2-Werte oder Kennwerte der Q-Statistik an (Higgins & Thompson, 2002; Huedo-Medina, Sánchez-Meca, Marín-Martínez & Botella, 2006, Viechtbauer, 2007a). Durch diese Kennwerte können SH-Ergebnisse sinnvoll ergänzt werden, wenn die Metaanalyse über das Feld der AO-Psychologie hinaus Leser finden soll.

Moderator- und Sensitivitätsanalysen

Die folgenden Schritte dienen dazu, den errechneten Effekt auf moderierende oder verzerrende Einflüsse zu prüfen, indem theoretisch hergeleitete Drittvariablen, extreme Werte und überproportionale Gewichte identifiziert und ihr Einfluss auf den Mittelwert und die Streuung der Effekte bestimmt werden. In Metaanalysen mit geringem k ($>$ ca. 25) stellt sich dabei das Problem, dass gerade bei geringer Studienanzahl Verzerrungseffekte durch Moderatorvariablen und Extremwerte einen großen Einfluss auf die Ergebnisse haben, die gebräuchlichen metaanalytischen Testverfahren allerdings wenig Power bei geringer Stichprobengröße aufweisen (Cafri, et al., 2010; Hedges & Pigott, 2004). Eine gute Vorgehensweise bei allen vorgestellten Verfahren ist unserer Meinung, potentielle Gefährdung durch Verzerrungen und die Power des Verfahrens bei gegebenem k gegeneinander abzuwägen, und im Zweifelsfall Ergebnisse alternativer Auswertungsstrategien (z. B. mit und ohne Ausreißer) ebenfalls zu berichten.

Moderatoranalysen

Die ermittelten metaanalytischen Effektstärken sollen möglichst präzise Schätzer für den wahren Effekt in der Population darstellen. Üblicherweise ergeben sich schon theoriegeleitet Annahmen darüber, dass Drittvariablen den Effekt moderieren (z. B. bei Nguyen & Ryan, 2008; laut Aytug, Rothstein, Zhou & Kern, 2011; übrigens eine besonders transparent beschriebene Metaanalyse) und der Effekt in Subpopulationen untersucht werden sollte. Weisen das breite Kredibilitätsintervall und geringe Anteile erklärter Varianz durch Artefakte auf Heterogenität der Stichprobe hin (Koslowsky & Sagie, 1993; Whitener 1990), empfehlen Hunter und Schmidt (2004) a posteriori Subgruppen anhand kategorialer Variablenausprägungen zu bilden oder stetig verteilte Merkmale an einem Cut-Off-Wert zu kategorisieren und die jeweiligen Mittelwerte (sollten sich voneinander unterscheiden), Kredibilitätsintervalle (sollten kleiner werden) und den Anteil der erklärten Varianz (sollte zunehmen) der jeweiligen Subgruppe zu interpretieren (vgl. auch Cortina, 2003). Während die Autoren in einer älteren Auflage (Hunter & Schmidt, 1990) noch den z-Test von Steiger (1980) zur Testung signifikanter Mittelwertsunterschiede zwischen den Subgruppen heranziehen (angewendet z. B. bei Ilies et al., 2007; Podsakoff, LePine & LePine, 2007; für einen alternativen Signifikanztest siehe Aguinis & Pierce, 1998), findet sich in der aktuellen Version (2004) keine Referenz mehr zu diesem Vorgehen.

Der Erfolg von Moderatoranalysen kann danach beurteilt werden, ob die Subgruppenbildung erfolgreich im Hinblick auf Generalisierungsaussagen ist, ob sich Subgruppenausprägungen bedeutsam unterscheiden und wie gut alle untersuchten Moderatoren gemeinsam die Vorhersage des Haupteffekts verbessern (Hunter & Schmidt, 2004; Viswesvaran & Sanchez, 1998; Whitener, 1990). Da es innerhalb des SH-Ansatzes gegenwärtig keine Empfehlung für eine Quantifizierung des Erfolgs von Moderatoranalysen gibt, muss auf Techniken anderer Anwendungstraditionen zurückgegriffen werden. Aguinis und Pierce (1998) bauen auf der Q-Statistik (Hedges & Olkin, 1985) auf und propagieren ein dreistufiges, an den SH-Ansatz angelehntes Vorgehen um ein Maß für die Variabilität der Effektstärke innerhalb und über Subpopulationen hinweg zu ermitteln (im Vergleich: Aguinis, Sturman & Pierce, 2008). Dieser schnitt bei Monte-Carlo Analysen im Vergleich mit Steigers z-Test zur Überprüfung dichotomer Moderatoren bei kleinen Stichproben ($k < 40$) besser ab (Marín Mártinez & Sánchez Meca, 1998).

Vom heutigen Stand der metaanalytischen Methoden aus gesehen, sollten Moderatoren jedoch nicht sequentiell überprüft, sondern multiple Moderatoren simultan be-

rücksichtigt werden. Dazu ist man in der SH-Tradition auf eine hierarchische Subgruppenbildung angewiesen, wobei die Aussagekraft mit zunehmend kleinerer Anzahl von Primärstudien in einer Sub-Metaanalyse schwächer wird (Cortina, 2003). Eine Moderatoranalyse mit der WLS-Metaregressionstechnik (weighted-least-sqares) bietet den Vorteil, dass mehrere kontinuierliche und kategoriale Moderatoren gleichzeitig überprüft, ihr Einfluss quantifiziert und die Gesamterklärungskraft des gemischten Modells (mixed model = Modell zufälliger Effekte mit Moderatoren) bestimmt werden können (Overton, 1998). Steel und Kammeyer-Mueller (2002) demonstrierten in Bezug auf kontinuierliche Moderatoren, dass die WLS-Technik als einzige unter vielen Methoden trotz vorliegender Multikollinearität von Moderatoren stabile Befunde bei guter Power aufweist. Das R-Paket metafor (Viechtbauer, 2010) bietet die entsprechenden Befehle, um diese gemischten Modelle zu spezifizieren und zu überprüfen (z. B. Jin & Rounds, 2012), eine ausführliche Beschreibung (allerdings für *z*-transformierte Verteilungen) findet sich auch bei Hedges und Pigott (2004).

Der Umgang mit extremen Werten

Ähnlich wie bei Primärstudien ist auch bei Metaanalysen der Umgang mit extremen Werten und Techniken der Ausreißeranalyse strittig. Hunter und Schmidt (2004) kritisieren generell den Ausschluss von Extremwerten, da unergründlich bleibt, ob diese durch Datenfehler oder große Stichprobenfehler zustande kommen. Wäre Letzteres der Fall und Extremwerte würden ausgeschlossen, würde die mittlere Effektstärke verzerrt. Traditionelle Verfahren außerhalb der Metaanalyse, die einen Cutoff-Wert für die Werteverteilung anlegen, greifen bei metaanalytischen Verfahren zu kurz, da variierende Stichprobengrößen nicht berücksichtigt werden. Ein spezifisches Metaanalyse-Outlier-Verfahren sind die SAMD-Statistiken (sample-adjusted meta-analytic deviancy, Huffcutt & Arthur, 1995), die jedoch (bezogen auf Korrelationskoeffizienten, nicht auf *d*-Werte) bei mittlerer Effektstärke und $k < 20$ systematisch mehr Outlier mit geringer als mit hoher Stichprobengröße identifizieren (Beal, Corey & Dunlap, 2002). Trotzdem empfehlen Geyskens und Kollegen (2009), im Sinne von Sensitivitätsanalysen SAMD-Analysen durchzuführen und Ergebnisse ohne und mit Outlier zu berichten (im Gegensatz zur untersuchten Stichprobe von 69 Metaanalysen, von denen 84 % keine Aussage zum Umgang mit Extremwerten machte). Empfehlenswert ist dieses Vorgehen auch, wenn einzelne Studien aufgrund deutlich abweichender Stichprobengröße mit proportionalem Übergewicht in die Integration eingehen (vgl. Nguyen & Ryan, 2008). Alternativ propagieren Fuller und Hester (1999), diese proportional übergewichteten Effekte nicht komplett auszuschließen, sondern ihr ursprüngliches Gewicht durch den Mittelwert zu ersetzen. Einen graphischen Überblick über Gewicht und Höhe der einzelnen Effektstärken erlaubt ein Forest-Plot (s. Abb. 4; Anzures-Cabrera & Higgins, 2010; Neyeloff, Fuchs & Moreira, 2012), der entweder die aufgenommenen Studieneffekte (z. B. bei Overstreet, Cegielski & Hall, 2013) oder mehrere metaanalytische Subgruppenergebnisse (z. B. bei Hülsheger & Schewe, 2011) untereinander darstellt und damit einen Vergleich ermöglicht.

Der Publikationsbias

Eine häufig formulierte Kritik an Metaanalysen ist die Gefahr des Publikationsbias, der besagt, dass Studien mit signifikanten Effekten aus verschiedenen Gründen höhere Chancen haben, publiziert zu werden. Somit bestünde die Gefahr, einen positiv verzerrten mittleren Wert zu erhalten, wenn unpublizierte Studien nicht ebenfalls in der Metaanalyse aufgenommen werden. Sutton (2009) stellt statistische und graphische Methoden vor, um die Verzerrungsgefahr im Nachhinein zu untersuchen. Der oft verwendete Fail-Safe-N-Test gibt an, wie viele Studien mit Nulleffekten unpubliziert in den Schubladen von Forschern liegen müssten, um den mittleren Effekt auf einen bestimmten Wert (z. B. kleiner als $r = .1$ oder .0) zu verringern (Dickersin, 2005). Die graphische Darstellung des Trim-and-Fill-Plots (Duval, 2005), dessen Umsetzung z. B. in Excel oder mit metafor in R möglich ist, erlaubt das Entdecken von fehlenden Studien in einem ausgewogenen Verhältnis von Effektgröße und Versuchspersonenanzahl und somit die Identifikation eines Publikationsbias. Auch die Untersuchung des Publikationsstatus als Moderator der Effektgröße (publizierte vs. graue Literatur) kann Aufschluss über eine Verzerrung geben. Banks, Kepes und McDaniel (2012) weisen auf die Problematik der geringen Power dieser traditionellen Methoden hin und schlagen Alternativen vor. Ergebnisse aktueller Studien im Bereich der angewandten Psychologie (Dalton, Aguinis, Dalton, Bosco & Pierce, 2012) legen jedoch nahe, den Einfluss des Schubladeneffekts auf Metaanalysen nicht überzubewerten, da die (oft aus Korrelationstabellen) entnommenen Effekte nicht zwangsweise hypothesenrelevant in der jeweiligen Primärstudie sind.

Weiterführende Analysen

Viele Metaanalysen geben nicht nur integrierte mittlere Effektstärken an, sondern nutzen diese ermittelten Werte, um etwa Wirkmechanismen zu analysieren. So ist die Überprüfung metaanalytischer multivariater Zusammenhänge anhand von Strukturgleichungsmodellen (SEM) als MASEM verbreitet und das Vorgehen auf statistisch einfachem Weg ausführlich beschrieben (Shadish, 1996; Viswesvaran & Ones, 1995). Da die Kombination von SEM und Metaanalysen statistisch weiterentwickelt wurde, sind aktuelle TTSEM-Ansätze (two-stage-MASEM, z. B. Beretvas & Furlow, 2006; Cheung, 2013) in der

Lage, Herausforderungen wie das Vorliegen von Korrelations- statt Kovarianzmatrizen und der Homogenitätsüberprüfung zufriedenstellend zu lösen. Landis (2013) vergleicht beide Ansätze und gibt Anwendungsempfehlungen.

Ergebnispräsentation und Interpretation

Die Ergebnisse einer Metaanalyse können je nach Zielsetzung, Zeitschrift und Adressaten mit verschiedenen Schwerpunkten dargestellt werden. In allen Fällen ist die saubere Dokumentation der Vorgänge und getroffenen Entscheidungen wichtig, denn auch Metaanalysen sollten, wie jede empirische Studie, das Kriterium der Replizierbarkeit erfüllen. Die APA hat, auf Vorarbeiten der Cochrane Collaboration beruhend, „meta-analytic reporting standards" formuliert (MARS; APA Working Group on JARS, 2008), die auch im aktuellen Publication Manual (APA, 2009) zu finden sind. Gegliedert nach Inhalten, wie eine publizierte Metaanalyse selbst, stellt diese Liste dreierlei Hilfen dar: eine universelle Gliederungsvorlage ebenso wie eine Checkliste vor dem Einreichen von Manuskripten für Verfasser und Bewertungskriterien für Leser und Beurteiler von psychologischen Metaanalysen. Die Arbeiten von Aytug und Kollegen (2011) und Geyskens und Kollegen (2009) offenbaren, dass ein Großteil der Metaanalysen in unserem Fachbereich das Vorgehen eher intransparent als replizierbar berichtet. Im Folgenden stellen wir Strategien für vollständiges Berichten und erleichternde Interpretation von metaanalytischen Ergebnissen vor. Anschließend gehen wir auf typische Herausforderungen bei der Verbreitung sowohl in wissenschaftlichen Zeitschriften als auch in der (organisationalen) Praxis ein.

Tabellen und Abbildungen

Jenseits von Signifikanztests helfen Abbildungen und Tabellen dabei, die Höhe von Effektstärken, ihre Streuung und Verschiedenheit (voneinander oder von Null) zu erkennen. Die Hauptergebnisse einer Metaanalyse mit Artefaktkorrektur lassen sich gut in Tabellenform abbilden und umfassen meist folgende Werte oder Äquivalente (vgl. Hülsheger & Schewe, 2011): k (Anzahl der Effektstärken), N (Anzahl der Versuchspersonen aller k), die nur um Stichprobengröße korrigierte gemittelte Effektstärke (rc oder dc) und ihre Varianz, die um alle relevanten Artefakte korrigierte Effektstärke (ρ oder Δ) und ihre Varianz, das 95 % Konfidenzintervall um ρ, das 80 % Kredibilitätsintervall um ρ und den Anteil der durch Artefakte erklärten Varianz. Der letzte Wert wird außerhalb der SH-Methode nicht berechnet, so dass es sinnhaft erscheint, universell genutzte Kennzahlen für das Ausmaß der Heterogenität zwischen Studien zusätzlich zu berichten, z. B. das Maß I^2 (Higgins & Thompson, 2002). Als Interpretationshilfe für das Einordnen der Effektstärke und ihrer Variabilität bieten Carlson und Ji (2011) hilfreiche Beispiele sowie AO-psychologische spezifische Mittelwerte für metaanalytische Effekte ($\rho = .18$, $SD\rho = .106$). Wie bei anderen multifaktoriellen Untersuchungen üblich, sollte eine Korrelationstabelle enthalten sein, wenn mit Regressionen oder Strukturgleichungsmodellen mehrere Konstrukte metaanalytisch in Beziehung gesetzt werden (z. B. Joseph & Newman, 2010).

Bax, Ikeda, Fukui, Yaju, Tsuruza und Moons (2008) untersuchten gebräuchliche metaanalytische Abbildungen hinsichtlich ihrer Objektivität bei der Bewertung von Studienheterogenität und eines Publikationseffekts. Forest-Plots (wie in Abb. 4), die die Effektstärken der enthaltenen Primärstudien oder metaanalytisch integrierten Subgruppen und deren Variabilität als Balken untereinander darstellen, erlauben eine gute Beurteilung der Heterogenität und können z. B. mit Excel erstellt werden (Neyeloff et al., 2012). Graphische Darstellungen der Trim-and-Fill-Technik eignen sich, um die Gefahr eines verzerrenden Publikationseffekts abzuschätzen, ebenso das (allerdings schwerer anzuwendende) Copas selection model (Anzures-Cabrera & Higgins, 2010).

Exemplarische Herausforderungen beim Schreiben für wissenschaftliche Zeitschriften

Während andere empirische Studien nicht ohne Hypothesen auskommen, sind viele frühe Metaanalysen zwar auf Basis einer Forschungsfrage, jedoch ohne explizite Hypothesen verfasst worden (Bsp. Cotton & Tuttle, 1986). Inzwischen liegen zu vielen Forschungsbereichen explorative Metaanalysen vor und die Anforderungen von Gutachtern haben sich geändert (vgl. Humphrey, 2011). Statt einfach „nur" Ergebnisse zusammenzufassen, sollte die Intention der eigenen Arbeit in Forschungsfrage und Hypothesen explizit genannt werden.

Die Anforderung, mit dem Methodenteil einer Metaanalyse die Replikation zu ermöglichen ist mit gängigen Beschränkungen der Textlänge kaum vereinbar. Deshalb ist es üblich, bestimmte Detailinformationen zum Recherche- und Kodierprozess nicht im Manuskript zu benennen, sondern auf die Verfügbarkeit dieser Aufzeichnungen und Dokumente hinzuweisen (auf Anfrage vom Erstautor erhältlich) oder sie als Online-Supplement zur Verfügung zu stellen (z. B. bei Klug & Maier, in press). Aytug und Kollegen (2011) nehmen einen Abgleich der Dokumentationsvollständigkeit laut APA mit der Realität und Besonderheiten von AO-psychologischen Metaanalysen vor und legen somit fachspezifische Anforderungen für Verfasser vor.

Auch wenn metaanalytische Methoden seit rund 40 Jahren weiterentwickelt werden (vgl. Schmidt & Hunter, 2003), hält sich Kritik zur Aussagekraft von Metaanalysen. Aguinis und Kollegen (2011) gehen unter anderem der An-

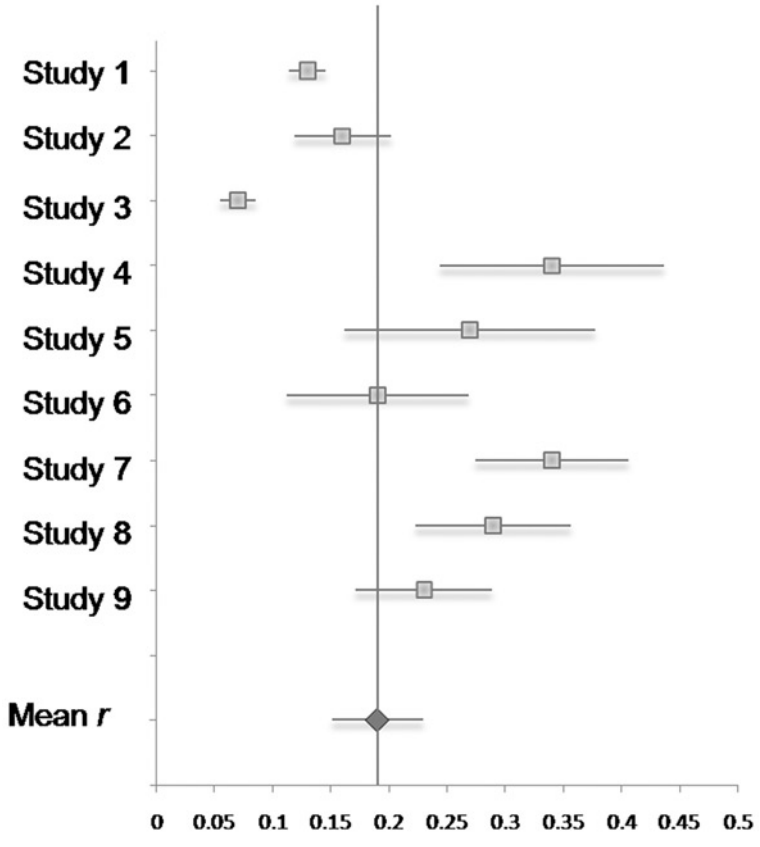


Abbildung 4. Beispielhafter Forest-Plot von neun Primärstudieneffektstärken und ihren Konfidenzintervallen sowie dem metaanalytisch integrierten Mittelwert *r*.

nahme nach, dass eine metaanalytische Effektstärke kausale Zusammenhänge belegt. Tatsächlich weisen Metaanalysen ein quasi-experimentelles Studiendesign auf (da Studieneigenschaften den Studien nicht randomisiert zugeordnet werden) und können nur die Aussagen liefern, zu denen die Ausgangsstudien Informationen bieten. Bei vorliegenden querschnittlichen Korrelationen aus Primärstudien erlaubt die Metaanalyse deshalb auch nur Aussagen über die *Voraussetzung* einer kausalen Beziehung, also die Korrelation aus uV und aV (vgl. Shadish, 1996). Etwas präzisere Hinweise auf Kausalität liefern metaanalytische Pfadanalysen in denen die angenommene Einflussrichtung und der reverse Pfad getestet werden (z. B. Nielsen & Einarsen, 2012). Kausale Schlussfolgerungen werden am stärksten gestützt, wenn die integrierten Studien längsschittliche oder (quasi-)experimentelle Designs aufweisen. In Forschungsarbeiten werden metaanalytische Effekte als stärkere Belege als z. B. gemeinsam interpretierte Primärstudienergebnisse angesehen (vgl. Combs, Ketchen, Crook, Russell & Roth, 2011), wobei auch ein Maß für die Variabilität des Effekts berichtet werden sollte.

Zur Zukunft von Metaanalysen in der Forschung und Praxis

Bedeutsam für die Anwender der psychometrischen Metaanalyse wird sein, den Anschluss an methodische Weiterentwicklungen in Bezug auf allgemeine Modelle zufälliger Effekte nicht zu verpassen. Zum Teil werden Erweiterungen der Methodik gut angenommen, z. B. die Nutzung von Metaregression zur Moderatoranalyse (Viechtbauer, 2007a) oder kombinierte Meta-Strukturgleichungsansätze (Landis, 2013). Andere Weiterentwicklungen sind aufgrund der Komplexität oder der benötigten Programme weniger verbreitet. Die mehrebenenbezogene Auswertung bietet Flexibilität beim Umgang mit abhängigen Effektstärken wie multiplen Messzeitpunkten oder gruppenbezogenen Auswertungen (Hox, 2010, Kap. 11; angewendet z. B. bei Seibert, Wang & Courtright, 2011). Bayesische Verfahren (Brannick & Hall, 2003; Louis & Zeltermann, 1994) vergleichen eine a-priori bestimmte Effektstärke mit dem berechneten mittleren Effekt und erlauben präzise Berechnungen der Varianz des mittleren Effekts (Steel & Kammeyer-Mueller, 2007). Neben der Öffnung für Weiterentwicklungen der anderen metaanalytischen Schulen bleibt die Herausforderung bestehen, SH-spezifische Kennwerte und Vorgehensweisen methodisch zu erklären, wenn diese auch an Leser außerhalb des AO-psychologischen Bereichs gerichtet ist.

Anwendern in der Praxis liefern Metaanalysen relevante Informationen zu der Richtung, Stärke, und Variabilität eines zu erwartenden oder vorhandenen Effekts (vgl. Le, Oh, Shaffer & Schmidt, 2007). Metaanalytische

Verfahren basieren auf der Annahme der Effektgeneralisierung (Schmidt & Hunter, 2003), das heißt Menschen leisten, führen, folgen, mobben und verhalten sich generell an verschiedenen Arbeitsplätzen unterschiedlicher Organisationen sehr ähnlich. Der Erfolg von organisationalen Maßnahmen und Methoden kann demzufolge durch empirische Untersuchungen an anderen Arbeitsplätzen vorhergesagt werden. Eine Metaanalyse bietet im Vergleich zu einer einzelnen empirischen Studie eine reliablere Schätzung des Effekts und ein Kredibilitätsintervall, zwischen dessen Werten sich der Effekt mit einer angegebenen Wahrscheinlichkeit befindet. Diese Information ermöglicht evidenzbasierte Entscheidungen über die Durchführung von Maßnahmen, z. B. indem der metaanalytische ermittelte Effekt in Kosten-Nutzen-Rechnungen verwendet wird (vgl. Linden & Adams, 2007).

Für die praktische Planung einer Maßnahme sind die gefundenen Moderatoren einer Metaanalyse hilfreich bei der optimalen Konzeption. So fanden etwa Kluger und DeNisi (1996) in ihrer Metaanalyse zur Effektivität von Feedback-Interventionen, dass die positiven Effekte auf die Leistung sich in einen negativen Effekt verkehren, wenn das Feedback personenbezogene Merkmale betrifft anstatt auf aufgabenbezogene Details einzugehen. Gerade diese Informationen über Bedingungen, die die Stärke eines Effekts moderieren, sind selten überhaupt in Primärstudien oder Praxisberichten verfügbar. Metaanalysen können den Austausch zwischen Forschung und Praxis sinnvoll bereichern (vgl. Combs et al., 2011), allerdings müssen metaanalytische Ergebnisse den in der Praxis Handelnden auch zugänglich gemacht werden. Verglichen mit dem medizinischen Feld fehlen bisher Journals im AO-psychologischen Bereich, die wissenschaftliche Ergebnisse speziell für die Zielgruppe der in der Praxis Tätigen aufbereiten. Wiedergegebene Artikel im *Journal of Evidence Based Medicine* geben z. B. einen kurzen Einblick in den konzeptionellen Hintergrund der Untersuchungen und gehen dafür ausführlicher auf die praktischen Implikationen der Ergebnisse ein. Darüber hinaus können metaanalytische Techniken auch genutzt werden, um unternehmensweite Interventionen zu evaluieren und deren Nutzen zu bestimmen (vgl. z. B. Morrow, Jarrett & Rupinski, 1997; Shantz & Latham, 2011).

Fazit

Schon früh bei der Durchführung einer Metaanalyse müssen viele Entscheidungen getroffen und alternative Vorgehensweisen abgewogen werden. Bei der Suche nach der „besten" Vorgehensweise sehen sich Forscher mit einer Vielzahl spezifischer, oft ambivalent ausfallender methodischer Empfehlungen verschiedener Schulrichtungen konfrontiert. Für diese Entscheidungsprozesse haben wir Kriterien und Empfehlungen vorgelegt, die Forschenden eine Evaluation der vorhandenen Optionen erlauben und eine Durchführung anleiten sollen.

Um die Verbreitung und Verwendung metaanalytischer Erkenntnisse in der Praxis zu erhöhen, sollten neue Kommunikationswege zwischen Forschern und Praktikern entwickelt und von Vertretern beider Lager genutzt werden. Im Bereich der AO-Psychologie existiert bisher keine Zeitschrift, die Ergebnisse und Handlungsempfehlungen publizierter Studien, vor allem auch Metaanalysen, für die praktische Anwendung aufbereitet.

Literatur

ADEPT: APA Databases and Electronic Products Trainings Institute (American Psychological Association, Ed.). (2012). *Quick reference guide: PsycINFO.* Zugriff am 05. 08. 2013. Verfügbar unter: http://www.apa.org/pubs/databases/training/psycnet.pdf

Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A. & Dalton, C. M. (2011). Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. *Journal of Management, 37,* 5 – 38.

Aguinis, H. & Pierce, C. A. (1998). Testing moderator variable hypotheses meta-analytically. *Journal of Management, 24,* 577 – 592.

Aguinis, H., Sturman, M. C. & Pierce, C. A. (2008). Comparison of three meta-analytic procedures for estimating moderating effects of categorical variables. *Organizational Research Methods, 11,* 9 – 34.

Algera, J. A., Jansen, P. G. W., Roe, R. & Vijn, P. (1984). Validity generalization: Some critical remarks on the Schmidt-Hunter procedure. *Journal of Occupational Psychology, 57,* 197 – 210.

Aloe, A. M., & Becker, B. J. (2012). An effect size for regression predictors in meta-analysis. *Journal of Educational and Behavioral Statistics*, *37*, 278 – 297.

American Psychological Association (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Andrés, A. M. & Marzo, P. (2004). Delta: A new measure of agreement between two raters. *British Journal of Mathematical & Statistical Psychology, 57,* 1 – 19.

Anzures-Cabrera, J. & Higgins, J. P. T. (2010). Graphical displays for meta-analysis: An overview with suggestions for practice. *Research Synthesis Methods, 1,* 66 – 80.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63,* 839 – 851.

Arendt, J. (2007). How do psychology researchers find studies to include in meta-analyses? *Behavioral & Social Sciences Librarian, 26,* 1 – 23.

Arthur, W., Bennett, W., Edens, P. S. & Bell, S. T. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology, 88,* 234 – 245.

Arthur, W., Bennett, W. & Huffcutt, A. I. (2001). *Conducting meta-analysis using SAS.* Mahwah, N.J: Lawrence Erlbaum Associates.

Aytug, Z. G., Rothstein, H. R., Zhou, W. & Kern, M. C. (2011). Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analyses. *Organizational Research Methods, 15,* 103–133.

Banks, G. C., Kepes, S. & McDaniel, M. A. (2012). Publication bias: A call for improved meta-analytic practice in the organizational sciences. *International Journal of Selection and Assessment, 20,* 182–196.

Bax, L., Ikeda, N., Fukui, N., Yaju, Y., Tsuruta, H. & Moons, K. G. M. (2008). More than numbers: The power of graphs in meta-analysis. *American Journal of Epidemiology, 169,* 249–255.

Bax, L., Yu, L.-M, Ikeda, N. & Moons, K. G. M. (2007). A systematic comparison of software dedicated to meta-analysis of causal studies. *BMC Medical Research Methodology, 7,* 40.

Beal, D. J., Corey, D. M. & Dunlap, W. P. (2002). On the bias of Huffcutt and Arthur's (1995) procedure for identifying outliers in the meta-analysis of correlations. *Journal of Applied Psychology, 87,* 583–589.

Beretvas, S. N. & Furlow, C. F. (2006). Evaluation of an approximate method for synthesizing covariance matrices for use in meta-analytic SEM. *Structural Equation Modeling – A Multidisciplinary Journal, 13,* 153–185.

Bernard, R. & Abrami, P. C. (2007). *Statistical applications in meta-analysis: Extracting, synthesizing and exploring variability in effect sizes,* Center for the Studies of Learning and Performance, Concordia University, Canada. Zugriff am 05.08.2013. Verfügbar unter: http://www.ncddr.org/training/resources/mod2unit13_bernardabramidraft.pdf

Borenstein, M., Hedges, L., Higgins, J. & Rothstein, H. (2005) *Comprehensive Meta-Analysis Version 2* [Computer software]. Engelwood, NJ: Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P. & Rothstein, H. R. (Eds.). (2009). *Introduction to meta-analysis.* Chichester: Wiley.

Borenstein, M., Hedges, L. V., Higgins, J. P.T & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1,* 97–111.

Borsboom, D. & Mellenbergh, G. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence, 30,* 505–514.

Bortz, J. & Döring, N. (2006). Metaanalyse. In J. Bortz & N. Döring (Hrsg.), *Forschungsmethoden und Evaluation* (S. 671–700). Berlin: Springer.

Brannick, M. & Hall, S. (2003). Validity generalization from a Bayesian perspective. In K. R. Murphy (Ed.), *Validity generalization. A critical review* (pp. 339–364). Mahwah, NJ: Lawrence Erlbaum Associates.

Brannick, M. T., Yang, L.-Q. & Cafri, G. (2011). Comparison of weights for meta-analysis of r and d under realistic conditions. *Organizational Research Methods, 14,* 587–607.

Cafri, G., Kromrey, J. D. & Brannick, M. T. (2010). A meta-meta-analysis: Empirical review of statistical power, type I error rates, effect sizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research, 45,* 239–270.

Carlson, K. D. & Ji, F. X. (2011). Citing and building on meta-analytic findings: A review and recommendations. *Organizational Research Methods, 14,* 696–717.

Chan, M. E. & Arvey, R. D. (2012). Meta-analysis and the development of knowledge. *Perspectives on Psychological Science, 7,* 79–92.

Cheung, M. W.-L. (2013). Fixed- and random-effects meta-analytic structural equation modeling: Examples and analyses in R. *Behavior Research Methods,* advance online publication. doi: 10.3758/s13428–013–0361-y

Cheung, S. F. & Chan, D. K.-S. (2004). Dependent effect sizes in meta-analysis: Incorporating the degree of interdependence. *Journal of Applied Psychology, 89,* 780–791.

Cicchetti, D. & Sparrow, S. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency, 86,* 127–137.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70,* 213–220.

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

Combs, J. G., Ketchen, D. J., Crook, T. Russell & Roth, P. L. (2011). Assessing cumulative evidence within 'macro' research: Why meta-analysis should be preferred over vote counting. *Journal of Management Studies, 48,* 178–197.

Cooper, H. M., Hedges L.V. & Valentine J.C (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

Cortina, J. M. (2003). Apples and oranges (and pears, oh my!): The search for moderators in meta-analysis. *Organizational Research Methods, 6,* 415–439.

Cotton, J. & Tuttle, J. (1986). Employee turnover: A meta-analysis and review with implications for research. *The Academy of Management Review, 11,* 55–70.

Dalton, D. R., Aguinis, H., Dalton, C., Bosco, F. A. & Pierce, C. A. (2012). Revisiting the file drawer problem in meta-analysis: An assessment of published and nonpublished correlation matrices. *Personnel Psychology, 65,* 221–249.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origin and scope, and preventing harm. In H. R. Rothstein, A. Sutton & M. Borenstein (Eds.), *Publication bias in metaanalysis: Prevention, assessment and adjustments* (pp. 11–34). Chichester, UK: Wiley.

Dieckmann, N. F., Malle, B. F. & Bodner, T. E. (2009). An empirical assessment of meta-analytic practice. *Review of General Psychology, 13,* 101–115.

Duval, S. (2005). The trim and fill method. In H. R. Rothstein, A. Sutton & M. Borenstein (Eds.), *Publication bias in metaanalysis: Prevention, assessment and adjustments* (pp. 125–144). Chichester, UK: Wiley.

Federal Government of the U.S. (2012). *North American industry classification system: United States, 2012.* Lanham, Md: Bernan Press.

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods, 6,* 161–180.

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods, 10,* 444–467.

Field, A. P. & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology, 63,* 665–694.

Fried, Y. & Ferris, G. R. (1987). The validity of the Job Characteristics Model: A review and meta-analysis. *Personnel Psychology, 40,* 287–322.

Fuller, J. B. & Hester, K. (1999). Comparing the sample-weighted and unweighted meta-analysis: An applied perspective. *Journal of Management, 25,* 803–828.

Geyskens, I., Krishnan, R., Steenkamp, J. E. M. & Cunha, P. V. (2009). A review and evaluation of meta-analysis practices in management research. *Journal of Management, 35,* 393–419.

Glass, G. V., MacGaw, B. & Smith, M. L. (Eds.). (1981). *Meta-analysis in social research.* Beverly Hills: Sage.

Goertz, W., Hülsheger, U. R., & Maier, G. W. (in review). The validity of specific cognitive abilities for the prediction of training success in Germany: A meta-analysis. *Manuscript submitted for publication.*

Griffeth, R. W., Hom, P. & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *Journal of Management, 26,* 463–488.

Hall, S. M. & Brannick, M. T. (2002). Comparison of two random-effects methods of meta-analysis. *Journal of Applied Psychology, 87,* 377–389.

Hedges, L. V. & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9,* 426–445.

Hedges, L. V. & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504.

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

Higgins, J. P. T. & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics In Medicine, 21,* 1539–1558.

Hoffman, C. (1995). Applying range restriction corrections using published norms: Three case studies. *Personnel Psychology, 48,* 913–923.

Hofstede, G. H. (2001). Culture's consequences: Comparing values, behaviors, institutions and organizations across nations. Thousand Oaks, California: Sage.

Hox, J. J. (2010). *Multilevel analysis* (2nd ed.). Quantitative methodology series. London: Routledge Academic.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F. & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I2 Index? *Psychological Methods, 11,* 193–206.

Huffcutt, A. I. & Arthur, W. (1995). Development of a new outlier statistic for meta-analytic data. *Journal of Applied Psychology, 80,* 327–334.

Hülsheger, U. R., Anderson, N. & Salgado, J. (2009). Team-level predictors of innovation at work: A comprehensive meta-analysis spanning three decades of research. *Journal of Applied Psychology, 94,* 1128–1145.

Hülsheger, U. R., Maier, G. W., Stumpp, T. & Muck, P. M. (2006). Vergleich kriteriumsbezogener Validitäten verschiedener Intelligenztests zur Vorhersage von Ausbildungserfolg in Deutschland. *Zeitschrift für Personalpsychologie, 5,* 145–162.

Hülsheger, U. R. & Schewe, A. F. (2011). On the costs and benefits of emotional labor: A meta-analysis of three decades of research. *Journal of Occupational Health Psychology, 16,* 361–389.

Humphrey, S. E. (2011). What does a great meta-analysis look like? *Organizational Psychology Review, 1,* 99–103.

Hunter, J. E. & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings* (1st ed.). Thousand Oaks, California: Sage.

Hunter, J. E. & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, California: Sage.

Hunter, J. E., Schmidt, F. L. & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91,* 594–612.

Ilies, R., Nahrgang, J. D. & Morgeson, F. P. (2007). Leader-member exchange and citizenship behaviors: A meta-analysis. *Journal of Applied Psychology, 92,* 269–277.

Jin, J. & Rounds, J. (2012). Stability and change in work values: A meta-analysis of longitudinal studies. *Journal of Vocational Behavior, 80,* 326–339.

Jones, A. P., Johnson, L., Butler, M. & Main, D. (1983). Apples and oranges: An empirical comparison of commonly used indices of interrater agreement. *Academy of Management Journal,* 507–519.

Joseph, D. L. & Newman, D. A. (2010). Emotional intelligence: An integrative meta-analysis and cascading model. *Journal of Applied Psychology, 95,* 54–78.

Judge, T. A., Piccolo, R. F., Podsakoff, N. P., Shaw, J. C. & Rich, B. L. (2010). The relationship between pay and job satisfaction: A meta-analysis of the literature. *Journal of Vocational Behavior, 77,* 157–167.

Judge, T. A., Thoresen, C. J., Bono, J. E. & Patton, G. K. (2001). The job satisfaction-job performance relationship: A qualitative and quantitative review. *Psychological Bulletin, 127,* 376–407.

Kinicki, A. J., McKee-Ryan, F. M., Schriesheim, C. A. & Carson, K. P. (2002). Assessing the construct validity of the Job Descriptive Index: A review and meta-analysis. *Journal of Applied Psychology, 87,* 14–32.

Klug, H.J.P. & Maier, G.W. (in press). Linking goal progress and subjective well-being: A meta-analysis. *Journal of Happiness Studies.*

Kluger, A. N. & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254–284.

Koslowsky, M. & Sagie, A. (1993). On the efficacy of credibility intervals as indicators of moderator effects in meta-analytic research. *Journal of Organziational Behavior, 14,* 695–699.

Krampen, G., Fell, C. B. & Schui, G. (2012). Professionelle Publikationspräferenzen von Mitgliedern der Deutschen Gesellschaft für Psychologie (DGPs). *Psychologische Rundschau, 63,* 175–178.

Landis, J. R. & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Landis, R. S. (2013). Successfully combining meta-analysis and structural equation modeling: Recommendations and strategies. *Journal of Business and Psychology, 28,* 251–261.

Le, H., Oh, I. S., Shaffer, J. & Schmidt, F. (2007). Implications of methodological advances for the practice of personnel selection: How practitioners benefit from meta-analysis. *Academy of Management Perspectives, 21,* 6–15.

Le, H. & Schmidt, F. L. (2006). Correcting for indirect range restriction in meta-analysis: Testing a new meta-analytic procedure. *Psychological Methods, 11,* 416–438.

Lincoln, T. M., Suttner, C. & Nestoriuc, Y. (2008). Wirksamkeit kognitiver Interventionen für Schizophrenie. *Psychologische Rundschau, 59,* 217–232.

Linden, A. & Adams, J. L. (2007). Determining if disease management saves money: an introduction to meta-analysis. *Journal of Evaluation in Clinical Practice, 13,* 400–407.

Lipsey, M. W. & Wilson, D. B. (2001). *Practical meta-analysis.* Applied social research methods series: Vol. 49. Thousand Oaks, California: Sage Publications.

Louis, T. & Zeltermann, D. (1994). Bayesian approaches to research synthesis. In H. M. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 411–422). New York, NY: Russell Sage Foundation.

Lux, S., Crook, T. R. & Woehr, D. J. (2010). Mixing business with politics: A meta-analysis of the antecedents and outcomes of corporate political activity. *Journal of Management, 37,* 223–247.

Maier, G. & Woschée, R. (2007). *Test-retest and mean-level stability of organizational commitment: A meta-analysis.* Poster presented at the European Congress on Work and Organizational Psychology in Stockholm, May 2007.

Marín Martínez, F. & Sánchez Meca, J. (1998). Testing for dichotomous moderators in meta-analysis. *The Journal of Experimental Education, 67,* 69–82.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11,* 364–386.

Morrow, C. C., Jarrett, M. & Rupinski, M. (1997). An investigation of the effect and economic utility of corporate-wide training. *Personnel Psychology, 50,* 91–117.

National Research Council. (1992). *Combining information: Statistical issues and opportunities for research.* Washington, DC: National Academy Press.

Neyeloff, J. L., Fuchs, S. C. & Moreira, L. B. (2012). Meta-analyses and forest plots using a microsoft excel spreadsheet: step-by-step guide focusing on descriptive data analysis. *BMC Research Notes, 5,* 52–58.

Ng, T. W. H. & Feldman, D. C. (2012). Employee voice behavior: A meta-analytic test of the conservation of resources framework. *Journal of Organizational Behavior, 33,* 216–234.

Nguyen, H.-H. D. & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology, 93,* 1314–1334.

Nielsen, M. B. & Einarsen, S. (2012). Outcomes of exposure to workplace bullying: A meta-analytic review. *Work & Stress, 26,* 309–332.

O'Brien, K. E., Biga, A., Kessler, S. R. & Allen, T. D. (2010). A meta-analytic investigation of gender differences in mentoring. *Journal of Management, 36,* 537–554.

Ones, D. S., Viswesvaran, C. & Schmidt, F. L. (2003). Personality and absenteeism: A meta-analysis of integrity tests. *European Journal of Personality, 17*, 19–38.

Overstreet, R. E., Cegielski, C. & Hall, D. (2013). Predictors of the intent to adopt preventive innovations: a meta-analysis. *Journal of Applied Social Psychology, 43,* 936–946.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3,* 354–379.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., Fleishman, E. A., Levin, K. Y. et al. (2001). Understanding work using the occupational information network (ONET). *Personnel Psychology, 54,* 451–492.

Peterson, R. A. & Brown S. P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology, 90,* 175–181.

Pierce, C. A. (2008). Software Review: Comprehensive Meta-Analysis (Version 2.2. 027). *Organizational Research Methods, 11,* 188–191.

Podsakoff, N. P., LePine, J. A. & LePine, M. A. (2007). Differential challenge stressor-hindrance stressor relationships with job attitudes, turnover intentions, turnover, and withdrawal behavior: A meta-analysis. *Journal of Applied Psychology, 92,* 438–454.

Quintana, S. M. & Minami, T. (2006). Guidelines for meta-analyses of counseling psychology research. *Counseling Psychologist, 34,* 839–877.

Riketta, M. (2008). The causal relation between job attitudes and performance: A meta-analysis of panel studies. *Journal of Applied Psychology, 93,* 472–481.

Rosenthal, R. (1991). *Meta-analytic procedures for social research.* Applied social research methods series: Vol. 6. Newbury Park, Calif.: Sage Publ.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 118,* 183–192.

Roth, P. L. (2007). Software Review: Hunter-Schmidt Meta-Analysis Programs 1.1. *Organizational Research Methods, 11,* 192–196.

Sackett, P. R. & Ostgaard, D. J. (1994). Job-specific applicant pools and national norms for cognitive ability tests: Implications for range restriction corrections in validation research. *Journal of Applied Psychology, 79,* 680–684.

Sackett, P. R. & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85,* 112–118.

Schmidt, F. L. & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529–540.

Schmidt, F. L. & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1,* 199–223.

Schmidt, F. & Hunter, J. E. (2003). History, development, evolution, and impact of validity generalization and meta-analysis methods, 1975–2001. In K. R. Murphy (Ed.), *Validity generalization. A critical review* (Applied psychology series, pp. 31–65). Mahwah, NJ: Lawrence Erlbaum Associates.

Schmidt, F. & Le, H. (2004). *Hunter-Schmidt meta-analysis programs 1.1* [Computer software]: The University of Iowa.

Schmidt, F. L., Oh, I.-S. & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology, 62,* 97–128.

Schmidt, F. L., Pearlman, K., Hunter, J. E. & Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. *Personnel Psychology, 32*, 257–281.

Schulze, R. (2004). *Meta-analysis: A comparison of approaches.* Cambridge, Mass.: Hogrefe & Huber.

Schulze, R. (2007). Current methods for meta-analysis. *Zeitschrift für Psychologie, 215,* 90–103.

Seibert, S. E., Wang, G. & Courtright, S. H. (2011). Antecedents and consequences of psychological and team empowerment in organizations: A meta-analytic review. *Journal of Applied Psychology, 96,* 981–1003.

Shadish, W. R. (1996). Meta-analysis and the exploration of causal mediating processes: A primer of examples, methods, and issues. *Psychological Methods, 1,* 47–65.

Shantz, A. & Latham, G. (2011). The effect of primed goals on employee performance: Implications for human resource management. *Human Resource Management, 50,* 289–299.

Sharpe, D. (1997). Of apples and oranges, file drawers and garbage: Why validity issues in meta-analysis will not go away. *Clinical Psychology Review, 17,* 881–901.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Steel, P. D. & Kammeyer-Mueller, J. D. (2002). Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology, 87,* 96–111.

Steel, P. D. G. & Kammeyer-Mueller, J. (2007). Bayesian variance estimation for meta-analysis: Quantifying our uncertainty. *Organizational Research Methods, 11,* 54–78.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245–251.

Stock, W. (1994). Systematic coding for research synthesis. In H. M. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 125–138). New York, NY: Russell Sage Foundation.

Sutton, A. (2009). Publication bias. In H. M. Cooper, Hedges L.V. & Valentine J.C (Eds.), *The handbook of research synthesis and meta-analysis.* (2nd ed., pp. 435–452). New York: Russell Sage Foundation.

Swert, K. de. (2012). *Calculating inter-coder reliability in media content analysis using Krippendorff's Alpha,* University of Amsterdam. Verfügbar unter: http://www.polcomm.org/wp-content/uploads/ICR01022012.pdf

Thorndike, R. L. (1949). Personnel selection; test and measurement techniques. Oxford, England: Wiley.

U.S. Bureau of Labor Statistics. (2010). *2010 SOC User Guide.* Verfügbar unter: http://www.bls.gov/soc/soc_2010_user_guide.pdf

Van Iddekinge, C. H. & Ployhart, R. (2008). Developments in the criterion-related validation of selection procedures: A critical review and recommendations for practice. *Personnel Psychology, 61,* 871–925.

Van Iddekinge, C. H., Roth, P. L., Raymark, P. H. & Odle-Dusseau, H. N. (2012). The criterion-related validity of integrity tests: An updated meta-analysis. *Journal of Applied Psychology, 97,* 499–530.

Viechtbauer, W. (2007a). Accounting for heterogeneity via random-effects models and moderator analyses in meta-analysis. *Zeitschrift für Psychologie, 215,* 104–121.

Viechtbauer, W. (2007b). Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology, 60,* 29–60.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36,* 1–48. Verfügbar unter: http://www.jstatsoft.org/v36/i03/paper

Viswesvaran, C. & Ones, D. S. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. *Personnel Psychology, 48,* 865–885.

Viswesvaran, C., Ones, D. S. & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81,* 557–574.

Viswesvaran, C. & Sanchez, J. I. (1998). Moderator search in meta-analysis: A review and cautionary note on existing approaches. *Educational and Psychological Measurement, 58,* 77–87.

Wallace, D. S., Paulson, R. M., Lord, C. G. & Bond, C. F. (2005). Which behaviors do attitudes predict? Meta-analyzing the effects of social pressure and perceived difficulty. *Review of General Psychology, 9,* 214–227.

Wanous, J. P. & Hudy, M. J. (2001). Single-item reliability: A replication and extension. *Organizational Research Methods, 4,* 361–375.

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology, 75,* 315–321.

Wilson, D. (2009). Systematic coding. In H. M. Cooper, Hedges L.V. & Valentine J.C (Eds.), *The handbook of research synthesis and meta-analysis.* (2nd ed., pp. 159–176). New York: Russell Sage Foundation.

Wilson, D. (2010). *Webpage meta-analysis stuff,* George Mason University. Zugriff am: 05. 08. 2013. Verfügbar unter http://mason.gmu.edu/~dwilsonb/ma.html

Wood, J. A. (2008). Methodology for dealing with duplicate study effects in a meta-analysis. *Organizational Research Methods, 11,* 79–95.

Yang, H., Sackett, P. R. & Nho, Y. (2004). Developing a procedure to correct for range restriction that involves both institutional selection and applicants' rejection of job offers. *Organizational Research Methods, 7,* 442–455.

Eingegangen: 06. 08. 2013
Revision eingegangen: 31. 03. 2014

Dipl.-Psych. Anna F. Schewe
Prof. Dr. Günter W. Maier

Universität Bielefeld,
Abteilung für Psychologie
Arbeits- und Organisationspsychologie
Universitätsstr. 25
33615 Bielefeld
E-Mail: aschewe@uni-bielefeld.de

Dr. Ute R. Hülsheger

Department of Work and Social Psychology
Maastricht University
P.O. Box 616
6200 MD Maastricht
Niederlande
E-Mail: Ute.Hulsheger@maastrichtuniversity.nl