Z Erziehungswiss (2014) (Suppl) 17:55–78
DOI 10.1007/s11618-014-0509-2

ZfE

Möglichkeiten und Grenzen systematischer Evidenzkumulation durch Forschungssynthesen in der Bildungsforschung

Andreas Beelmann

Zusammenfassung: Im Beitrag wird eine Übersicht zu Funktionen, Potentialen und Grenzen systematischer Forschungssynthesen gegeben. Einleitend werden zunächst die Rationale und bedeutsame Entwicklungen dieses Forschungsansatzes skizziert, wobei der Schwerpunkt auf Meta-Analysen als eine prominente Form der Forschungssynthese liegt. Im Anschluss erfolgt eine Erörterung wichtiger Funktionen, indem der Mehrwert gegenüber Einzeluntersuchungen und empirischen Primärstudien herausgestellt wird. Dabei geht es u. a. um die Potentiale, empirische Evidenzen umfassender und valider abzubilden, wissenschaftliche Artefakte und forschungsmethodische Einflüsse besser kontrollieren zu können sowie um einen verbesserten Transfer von wissenschaftlichen Produkten in Praxis, Gesellschaft und Politik. Anschließend werden mögliche Probleme und Grenzen von Forschungssynthesen wie etwa Publikationsverzerrungen, Konfundierungen von bedeutsamen Ergebnismoderatoren, die Rolle der forschungsmethodischen Qualität bei der Selektion von Primärstudien und anderes diskutiert. Möglichkeiten und Grenzen werden abschließend an Beispielen aus der Bildungsforschung (z. B. Evaluation von Förder- oder Präventionsprogrammen, Lern-/Lehrforschung) illustriert.

Schlüsselwörter: Meta-Analyse · Forschungssynthese · Wissenschaft -Praxis-Transfer · Empirische Bildungsforschung · Publikationsverzerrungen · Prävention · Lernförderung

Potentials and limits of the systematic accumulation of evidence via systematic research synthesis within educational research

Abstract: This paper provides an overview of the functions, potentials, and limits of systematic research reviews. It first describes the rationale and gives a short history of systematic reviews focusing mainly on meta-analysis as the most prominent form of research synthesis. We than discuss several functions by highlighting the additional benefits of research syntheses compared to primary research and single studies. Potentials such as being able to picture empirical evidence more comprehensively and validly, a better control of artifacts and methodological bias, and a better transfer of research results to practice, society, and policymaking are reported. These

A. Beelmann (✉)
Abteilung für Forschungssynthese, Intervention und Evaluation, Institut für Psychologie,
Friedrich-Schiller-Universität Jena,
Jena, Deutschland
E-Mail: andreas.beelmann@uni-jena.de

potentials are contrasted with the problems and limits of research synthesis such as publication bias, methodological bias, and confounds between study characteristics. Finally, the potentials and limits of research synthesis are illustrated with examples from recent educational research (e.g., evaluation of preventive interventions, research on teaching and learning).

Keywords: Meta-analysis · Research synthesis · Translational research · Empirical educational research · Publication bias · Prevention · Promotion

1 Einleitung

Als vor fast 40 Jahren die erste Meta-Analyse von Smith und Glass (1977) zur Wirksamkeit von Psychotherapie veröffentlicht wurde, war den Autoren wahrscheinlich nicht klar, welchen weiteichenden Einfluss diese Publikation auf die sozial- und naturwissenschaftliche Forschung haben würde. Seitdem sind etwa 20.000 Meta-Analysen erschienen und jedes Jahr kommen 500 bis 1.000 neue Einträge in den einschlägigen Literaturdatenbanken hinzu. Vor allem im Bereich der Evaluation sozial- und verhaltenswissenschaftlicher Interventionen sind Meta-Analysen weit verbreitet und tragen maßgeblich dazu bei, dass Evidenzen der wissenschaftlichen Forschung auf systematische Weise integriert und zur Verfügung gestellt werden.

Meta-Analysen können als eine prominente Form der Forschungssynthese betrachtet werden. Sie sind darauf ausgerichtet, empirische Forschungsergebnisse zu einem Thema quantitativ und unter Nutzung von statistischen Verfahren zusammenzufassen. Der Begriff der Forschungssynthese ist weiter gefasst und beinhaltet alle systematischen Ergebnisbilanzen, wobei (anders als bei Meta-Analysen) nicht notwendigerweise eine statistische Integration der Daten erforderlich ist (wie z. B. in klassischen narrativen Literaturreviews). Der Begriff der integrativen Forschung wird in der Regel noch breiter verwandt und bezieht zum Beispiel eine Zusammenfassung und Integration theoretischer Modelle mit ein (vgl. dazu insbesondere Cooper 1982, 1988). Im Folgenden wird sich die Darstellung allerdings allein wegen ihrer großen Verbreitung in der sozialwissenschaftlichen Forschung vorwiegend auf meta-analytische Arbeiten beziehen, während der Begriff der Forschungssynthese verwandt wird, wenn alle systematischen Ergebnisintegrationen gemeint sind.

Seit der klassischen Arbeit von Smith und Glass (1977) sind Forschungssynthesen und speziell Meta-Analysen zu sehr unterschiedlichen Themen der psychologischen und pädagogischen Forschung erschienen (vgl. bereits Lipsey und Wilson 1993). Das wohl prominenteste Beispiel aus der aktuellen Bildungsforschung stammt von Hattie (2009), der die Ergebnisse von über 800 Meta-Analysen zu Bedingungen und Einflussfaktoren des Lehrens zusammenfassend dargestellt hat. Organisationen wie die Cochrane Collaboration in der Medizin und seit einigen Jahren auch die Campbell Collaboration in den Sozial- und Verhaltenswissenschaften (vgl. Lösel 2009) sowie das am US-Department of Education angesiedelte „What Works Clearinghouse“ (WWC) in der Bildungsforschung haben es sich zur Aufgabe gestellt, einer breiten Öffentlichkeit aus Wissenschaft, Politik und Gesellschaft qualitativ hochwertige Forschungsüberblicke zu gesellschaftlich relevanten und praktisch bedeutsamen Themen zur Verfügung zu stellen. Diese Gesellschaften haben mittlerweile hunderte von Meta-Analysen und Ergebnisbilanzen über ein-

schlägige Datenbanken im Internet publiziert (vgl. z. B. www.campbellcollaboration.org). Sie regen zudem kontinuierliche Aktualisierungen der Arbeiten an, um fortlaufend neue Erkenntnisse in bereits bestehendes Wissen zu integrieren.

Methodik und Praxis von Forschungssynthesen haben sich in den letzten Jahrzenten beträchtlich weiter entwickelt. Einer anfänglichen und zum Teil noch immer anzutreffenden unkritischen Euphorie, derartige Arbeiten seien frei von Problemen der Objektivität und Reliabilität, ist mittlerweile der Gewissheit gewichen, dass auch Forschungssynthesen eine Vielzahl von inhaltlichen Fragen und methodischen Problemen aufwerfen. Infolge dieser Diskurse wurden auch die Verfahren systematisch verbessert, was sich an zwei Entwicklungslinien darstellen lässt. Erstens ist es zu einer weitgehenden Systematisierung von Forschungssynthesen gekommen. Vor allem Harris Cooper (vgl. Cooper 1982, 2010; Cooper und Hedges 2009a) hat dafür plädiert, integrative Forschungsarbeiten insgesamt als eine strukturierte Abfolge von Arbeitsschritten zu verstehen, um eine hohe Qualität und maximale Transparenz des Vorgehens sicher zu stellen. Danach können fünf Forschungsphasen differenziert werden, die von der Formulierung einer Fragestellung bis hin zur Interpretation und Dokumentation der Ergebnisse reichen(vgl. Tab. 1). Auf jede dieser Phasen lassen sich charakteristische Aufgaben und bestimmte Problemstellungen nennen, die eine integrative Forschungsarbeit kennzeichnen. Sie haben mittlerweile auch in methodischen Standards ihren Niederschlag gefunden (vgl. APA Standards 2008).

Eine zweite Entwicklungslinie betrifft die Konstruktion verbesserter statistischer Analyse- und Integrationsmethoden speziell im Rahmen von Meta-Analysen. Dort wird die Zusammenfassung empirischer Forschungsergebnisse unter Nutzung bestimmter Verfahrenstechniken und statistischen Methoden organisiert, die einen möglichst vollständigen und eine zumeist große Studienzahl umfassenden Überblick gewährleisten sollen. Kern dieses Prozesses ist die Berechnung einheitlicher und zusammenfassender Ergebnisparameter (Effektstärken, integrierte Signifikanztests), die es nicht nur erlauben, quantitative Gesamtbilanzen zu erstellen, sondern auch die Möglichkeit eröffnen, den Einfluss bestimmter Studienmerkmale (z. B. Art der Intervention, Intensität der Programme, Merkmale der Stichprobe u. a.) auf die Studienergebnisse zu untersuchen.

Im Rahmen dieser Ergebniszusammenfassung haben sich seit den ersten meta-analytischen Arbeiten zahlreiche Innovationen ergeben. Hunter und Schmidt (2004) haben etwa gezeigt, dass auch größere Ergebnisunterschiede zwischen an sich vergleichbaren Studien auf methodische Einflussgrößen wie Stichprobenfehler oder die Reliabilität der Messinstrumente zurückgeführt werden können, die bei der Integration der Befunde berücksichtigt werden müssen. Andere Autoren wie Hedges und Kollegen (vgl. Hedges und Olkin 1985) haben statistische Analysemodelle entwickelt, bei denen vor allem die Ergebnisvariabilität untersucht wird. Ziel dieser Verfahren ist zunächst die Zusammenfassung aller Primärstudien-Stichproben zu einer Art aggregierter Gesamtstichprobe, mit deren Hilfe ein Populationsparameter als mittlerer Befund aller Einzelergebnisse geschätzt wird. Dazu werden alle Studienergebnisse zunächst in Form von Effektstärken standardisiert und dann jeweils mit dem Inversen ihrer Varianz, die im Wesentlichen aus dem Stichprobenumfang der jeweiligen Studien geschätzt wird, gewichtet und sodann über Studien hinweg aggregiert. Somit gehen „größere“ Studien mit einem höheren Gewicht in die Ergebniszusammenfassung ein. Anschließend erfolgt die Prüfung der Ergebnisva-

Tab. 1: Übersicht zu Forschungsphasen und Aufgaben bei Forschungssynthesen

Forschungsphase	Aufgaben/Voraussetzungen/Problemstellungen
Formulierung einer Forschungsfrage und Festlegung von Auswahlkriterien	Vorliegen eines Bedarfs an integrativer Arbeit (empirisch bereits mehrfach geprüfte Fragestellung)
	Definition zentraler inhaltlicher Konstrukte (ggf. auch unter Abgrenzung zu verwandten Konzepten)
	Präzise Formulierung von inhaltlich und methodisch begründeten Auswahlkriterien
Literatursuche und Studienauswahl	Identifikation möglichst aller nach den Auswahlkriterien relevanten Untersuchungen
	Suche und Beschaffung über multiple Kanäle (Datenbanken, zentrale Journale, bereits verfügbare Reviews, Suche in bereits identifizierten Studien, etc.)
	Identifikation und Beschaffung „grauer“ und unpublizierter Literatur, um Selektionen vorzubeugen (siehe Publikationsverzerrungen)
Kodierung nach inhaltlichen und methodischen Merkmalen	Erarbeitung eines einheitlichen Auswertungsplans zur Erfassung inhaltlich und methodisch relevanter Studienmerkmale (z. B. Art der Intervention, Intensität, Art des Untersuchungsdesigns, Stichprobengröße, sozialer Kontext der Untersuchung)
	Informationsbeschaffung auch bei unvollständiger Dokumentation in den Primärstudien (ergänzende Studien, Autorenkontakt)
	Überprüfung und Sicherstellung einer hohen Kodierqualität, insbesondere bei Merkmalen, die hohe Inferenz verlangen (z. B. Implementationsqualität der Intervention, methodische Qualität) durch Training und statistischer Prüfung
Systematische Dokumentation und Auswertung der Befunde	Detaillierte Dokumentation der ausgewerteten Ergebnisse der integrierten Primärstudien
	Berechnung einheitlicher Ergebnisparameter (z. B. Effektstärken wie *d*-Index, Korrelation, Odds Ratio)
	Aggregation der Befunde unter Berücksichtigung der Stichprobengröße der Primärstudien (unterschiedliche Berechnungsmodelle)
	Identifikation signifikanter Ergebnismoderatoren durch Meta-Varianz oder Meta-Regressionsanalysen
Interpretation und Dokumentation der Ergebnisse	Interpretation der mittleren Ergebnisse und signifikanter Ergebnismoderatoren
	Möglichst anschauliche Darstellung der Ergebnisbilanzen (Umwandlung von Effektgrößen in inhaltlichen Aussagen)
	Detaillierte Dokumentation der meta-analytischen Methoden und Primärliteratur (einschließlich Auswahlkriterien und Auswahlprozesse)

rianz auf Homogenität. Dabei wird die Frage beantwortet, ob die aufgetretene Variabilität der Befunde mit dem kumulierten Stichprobenfehler erklärt werden kann oder weitere Varianzquellen inhaltlicher oder methodischer Art vorliegen müssen. Abhängig von der Art der Inferenz, die vorgenommen werden soll, bieten sich ähnlich wie bei varianzanalytischen Methoden zwei unterschiedliche Rechenmodelle an. Das Fixed-Effect-Model

(FEM) geht davon aus, das den integrierten Studien eine gemeinsame Populationseffektstärke zugrunde liegt, während beim Random-Effect-Model (REM) die Populationseffektstärke als Zufallsvariable geschätzt wird und somit eine größere Generalisierbarkeit ermöglicht. Die Implikationen dieser beiden statistischen Modelle sollen hier nicht weiter diskutiert werden (vgl. dazu Hedges und Vevea 1998; Schulze 2007), in der meta-analytischen Praxis kann ihre Verwendung jedoch erhebliche Auswirkungen auf die mittleren Ergebnisschätzungen und die Identifikation von Moderatoren haben (vgl. z. B. Lösel und Beelmann 2003). In der Regel sollte unter der Bedingung signifikanter Heterogenität der Ergebnisse das REM verwendet werden, während homogene, d. h. eine nicht über den Stichprobenfehler hinausgehende Variabilität der Befunde das FEM nahe legen. Die Wahl des Aggregationsmodells ist jedoch auch von inhaltlichen Überlegungen, zum Beispiel zur *inhaltlichen* Heterogenität der Fragestellung, abhängig.

In beiden Modellen ist die nachfolgende Analyse von Zusammenhängen zwischen inhaltlichen und methodischen Merkmalen (d. h. die Analyse von Moderatoren) möglich und sinnvoll. Dazu besteht grundsätzlich die Möglichkeit kategorialer oder kontinuierlichen Auswertungen über stichprobengewichtete Varianz- oder Regressionsanalysen. Typischerweise werden kategoriale Auswertungsprozeduren für dichotome oder qualitative Variablen (e.g., Interventionstyp, Geschlecht der Stichprobe, Forschungsdesign), kontinuierliche Auswertungsprozeduren für metrische Variablen (Programmintensität, Alter der Stichprobe, methodische Qualität der Studie) verwendet, analog zu ungewichteten Analysen ist dies jedoch nicht zwingend erforderlich. Gewichtete Meta-Regressionsanalysen dienen zudem der Bestimmung der relativen Varianzstärke von Moderatoren, weil selbstverständlich zwischen den einzelnen Merkmalen mehr oder weniger große Konfundierungen vorliegen (vgl. dazu Abschn. 2.2.3). Weiterführende Informationen zu den statistischen Verfahren von Meta-Analysen finden sich z. B. bei Cooper (2010), Cooper et al. (2009) sowie Lipsey und Wilson (2001).

1.1 Potentiale und Nutzen von Forschungssynthesen

Welche Nützlichkeitsannahmen liegen nun Forschungssynthesen im Vergleich zu Primärstudien zugrunde oder anders gefragt: Was ist der spezielle Mehrwert einer derartigen Ergebniszusammenfassung? Zunächst ist erstens ein forschungspragmatischer Nutzen festzustellen. Adressaten von Forschungsergebnissen, seien es nun Wissenschaftler selbst, politische Entscheidungsträger, die interessierte Öffentlichkeit, Medien oder die sozialwissenschaftliche und pädagogische Berufspraxis, fällt es zunehmend schwerer, relevante Forschungen und die daraus erwachsenen Erkenntnisse zu überblicken. Allein an der Entwicklung der Publikationszahlen wird deutlich, dass es heute selbst für Experten schwierig ist, auf dem Laufenden zu bleiben und bedeutsame Neuentwicklungen zu registrieren. Forschungssynthesen weisen daher einen für verschiedene Adressatengruppen beträchtlichen Nutzen im Hinblick auf die niedrigschwellige Bereitstellung aktueller Forschungsbilanzen auf. Dieser Vorteil impliziert jedoch zugleich eine hohe Verantwortung. Forschungssynthesen werden der Aufgabe, zuverlässige Bilanzen zu erstellen, nur in dem Maße gerecht, wie ein zuverlässiges und vollständiges oder mindestens repräsentatives Gesamtbild zu einer Forschungsfrage erstellt wird.

Neben dem pragmatischen Nutzen von Forschungssynthesen liegen zweitens wissenschaftstheoretisch begründete Nutzenargumente vor. So ist die Primärforschung von mindestens zwei ernsthaften Problemen betroffen, die Forschungssynthesen über den reinen Aspekt der Zusammenfassung hinaus rechtfertigen. Zum einen weist jede wissenschaftliche Studie idiosynkratrische Elemente auf, d. h. sie findet in einem bestimmten Setting statt, wurde durch eine bestimmte Forschergruppe durchgeführt und untersucht eine bestimmte Stichprobe etc., so dass es oftmals schwierig ist, die Ergebnisse ohne weiteres über Kontexte, Personen, Zeit und anderes zu generalisieren. Zum anderen ist jede Studie mit forschungsmethodischen Restriktionen verbunden, weil die Kontrolle sämtlicher Validitätsgefährdungen zumindest in der Anwendungsforschung illusorisch und de facto unmöglich ist. Diese Probleme werden – so die Annahme – durch systematische Zusammenfassungen inhalts- und strukturgleicher Studien abmildert, ausgeglichen oder systematisch kontrolliert (z. B. bei der Gegenüberstellung von Ergebnissen aus Längsschnittstudien mit hohem vs. geringem Stichprobenausfall).

Schließlich besteht drittens speziell in Meta-Analysen die Möglichkeit, thematisch vergleichbare Studien anhand variierender Merkmale (z. B. hinsichtlich unterschiedlicher Altersgruppen oder Schulformen) quantitativ gegenüber zu stellen und damit (indirekt) Einflussfaktoren zu untersuchen, die im Rahmen von Primärstudien nicht oder nur unter sehr hohem Aufwand untersucht werden können. Dazu gehören etwa die Rolle unterschiedlicher forschungsmethodischer Merkmale (z. B. hinsichtlich des Untersuchungsdesigns), der Vergleich einer größeren Anzahl unterschiedlicher Interventionsansätze (z. B. unterschiedliche Unterrichtsformen) oder die Gegenüberstellung von Ergebnissen, die in verschiedener Studiensettings (z. B. in unterschiedlichen Schultypen) erzielt wurden. Damit tragen Forschungssynthesen nicht nur zu einer Differenzierung von Forschungsergebnissen bei, sondern u. U. auch zu verbesserten Möglichkeiten der Theoriebildung über die untersuchten Zusammenhänge (Cook et al. 1992). Allerdings sind der Testung von Hypothesen sowie insgesamt der Theorieprüfung in Forschungssynthesen auch Grenzen gesetzt, die vor allem damit zu haben, dass Studienmerkmale nicht zufällig variieren (d. h. nicht randomisiert über Studien vorliegen), sondern in der Regel systematisch miteinander zusammenhängen und konfundiert sind (Beelmann und Lipsey in Druck). Eine solche zufällige Verteilung wäre aber wichtig, um theoretische Annahmen unabhängig zu prüfen. Die Konfundierungsproblematik ist allerdings auch schon bei der Ergebniszusammenfassung ein wichtiges Thema, auf das noch einzugehen sein wird (vgl. Abschn. 2.2.3).

Die genannten Potentiale sowie der Anspruch, eine möglichst komplette Übersicht zu einem Themengebiet zu erstellen, ist zugleich mit der Annahme verbunden, dass Forschungssynthesen und speziell meta-analytische Ergebnisse eine insgesamt höhere Validität aufweisen als die Befunde einzelner Primärstudien. Cook (1991) hat diese Annahme am Validitätskonzept von Cook und Campbell (1979; Shadish et al. 2002) verdeutlicht. Danach erhöht sich die statistische Validität durch eine verbesserte statistische Aussagekraft, die durch Aggregation großer Datenmengen resultiert. Die interne Validität wird durch die Möglichkeit verbessert, systematische Verzerrungen durch forschungsmethodische Einflussgrößen zu untersuchen und ggf. zu korrigieren. So kann beispielsweise der Frage nachgegangen werden, ob die Art des Untersuchungsdesign (z. B. randomisiertes Experiment vs. Quasi-Experiment) einen Einfluss auf die empirischen Resultate hat.

Die Konstruktvalidität erhöht sich, weil – technisch gesprochen – unterschiedliche Operationalisierungen der Konstrukte gegenüber gestellt werden können und sich somit die Generalisierbarkeit der Aussagen verbessert. Bei Zusammenfassungen von Interventionsstudien können beispielsweise unterschiedliche (alternative) Interventionsverfahren im Hinblick auf eine große Bandbreite von Erfolgskriterien miteinander verglichen werden. Eine verbesserte Generalisierbarkeit gilt auch für die externe Validität zum Beispiel im Hinblick auf unterschiedliche Studiensettings und Personengruppen. Forschungssynthesen führen somit zu zuverlässigeren Kausalinterpretationen durch kumulative Evidenzen aus heterogenen Replikationen oder anders gesagt: Durch die wiederholte Überprüfung und Bestätigung empirischer Forschungsergebnisse.

1.2 Funktionen von Forschungssynthesen und Potentiale gegenüber der Primärforschung

Die genannten Überlegungen leiten über zur weiterführenden Frage, welche besonderen Funktionen aus den genannten Potentialen abgeleitet werden können. Dazu gibt Tab. 2 zunächst eine Übersicht zusammen mit den sich daraus resultierenden Vorteilen gegenüber empirischen Primärstudien.

Prominentester Zweck ist die bereits ausgeführte Zusammenfassung und Integration von vorhandenen wissenschaftlichen Erkenntnissen und die Bereitstellung einer „State of the Art"- Bilanz (*Integrative Funktion* im engeren Sinne). Innerhalb dieser Funktion spielen Forschungssynthesen eine besondere Rolle beim Konzept der Evidenzbasierung, das seit gut einer Dekade auch in der Bildungsforschung prominent vorgetragen wird.

Tab. 2: Übersicht zu Funktionen von Forschungssynthesen und Vorteile gegenüber der Primärforschung

Funktion	Beschreibung	Vorteil gegenüber der Primärforschung
Integrationsfunktion i.e.S	Zusammenfassung bestehender Evidenzen auf Basis einer Gesamtbilanz	Höhere Reliabilität und Validität der Befunde
Historische Funktion	Beschreibung des historischen Verlaufs eines Forschungsfeldes	Weitgehende Vollständigkeit zugrundeliegender Forschungsarbeiten
Allokationsfunktion	Steuerung von Ressourcen je nach bisherigen Forschungsergebnissen und –volumen	Weitgehende Vollständigkeit zugrundeliegender Forschungsarbeiten
Strukturierungsfunktion	Neu-Strukturierung eines Forschungsfeldes	Weitgehende Vollständigkeit zugrundeliegender Forschungsarbeiten
Kontrollfunktion	Überprüfung von Fehlern, Unterlassungen, problematischen Auswertungsprozeduren etc.	Bessere Vergleichsmöglichkeiten und breitere Datenbasis
Transferfunktion	Übertragung von wissenschaftlichen Befunden in praktische oder politische Handlungsfelder	Handlungsempfehlungen auf Basis tatsächlich bestehender Evidenzen

Mit diesem Konzept wird der Anspruch formuliert, professionelle Handlungsstrategien und -empfehlungen auf geprüftes Wissen aus empirischen Untersuchungen zu stützen. So wird verlangt, dass etwa Förderprogramme nur dann eingesetzt werden sollen, wenn ihre Effektivität in empirischen Untersuchungen mit Kontrollgruppendesign als hinreichend bestätigt gilt. Das Konzept der Evidenzbasierung ist mittlerweile weit verbreitet und wird auch in der sozialwissenschaftlichen und pädagogischen Praxis intensiv diskutiert. Bislang fehlen allerdings anerkannte Konkretisierungen und allgemein verbindliche Standards. Insbesondere ist unklar, welche Befunde genau vorliegenden müssen, um den Status einer evidenzbasierten Handlungsempfehlung zu erreichen. Angesichts dieser Unklarheiten können Forschungssynthesen stärker als Einzelstudien einen signifikanten Beitrag zur Umsetzung evidenzbasierter Konzepte beitragen, weil tatsächlich auf Basis einer *Gesamtbilanz bestehender Evidenzen* argumentiert wird. Damit wird eine nachvollziehbare Interpretation erleichtert, während die Ergebnisse von Einzeluntersuchungen nicht selten erheblich variieren können, so dass es im Einzelfall immer möglich sein wird, unterschiedliche Positionen mit entsprechenden Befunden zu stützen. Die Möglichkeit, bestimmte Standpunkte durch selektive Auswahl von Studienergebnissen zu belegen, entfällt somit oder ist zumindest deutlich eingeschränkt.

Weitere Funktionen systematischer Ergebnisüberblicke betreffen beispielsweise Aussagen im Hinblick auf die *historische Entwicklung* eines Forschungsfeldes etwa zu bestimmten Forschungsparadigmen, Paradigmenwechsel, kulturspezifischen Differenzierungen, zur Entwicklung der Forschungsqualität und anderes mehr (historische Funktion). Desweiteren können durch Forschungssynthesen Forschungsfelder umfassend beschrieben, Forschungsumfänge skizziert und spezifische Forschungslücken identifiziert werden, die zum Beispiel für *Allokationsentscheidungen* etwa bei der Forschungsförderung genutzt werden können (Allokationsfunktion). Viertens beinhalten Forschungssynthesen immer eine *Neustrukturierung des Forschungsfeldes*, um aktuell verfügbare Evidenzen einheitlich auswerten zu können. Dies trägt in aller Regel auch zur weiteren Entwicklung eines Forschungsfeldes bei (Strukturierungsfunktion). Fünftens kann über die systematische Zusammenfassung von Primärstudienergebnissen die wissenschaftliche Praxis nach definierten Standards überprüft oder die mangelnde Rezeption vorhandener Erkenntnisse, Plagiate, Widersprüche, der Einfluss von Forscher- oder Autoreninteressen, tendenziöse Auswertungsmethoden identifiziert werden (*Kontrollfunktion*). So konnte etwa Eisner und Humphreys (2011) in einer Reanalyse von meta-analytischen Daten zur Wirkung von Elterntrainingsprogrammen zeigen, dass finanzielle Interessenkonflikte von Autoren systematisch mit höheren Effektstärken einhergingen. Diese Unterschiede waren beträchtlich und legten weiterführende Überlegungen zur Unabhängigkeit von persönlichen Interessen und Forschungsergebnissen bzw. Publikationsgewohnheiten nahe.

Schließlich eignen sich Forschungssynthesen vermutlich besonders gut dazu, die allenthalben thematisierte Kluft zwischen Wissenschaft auf der einen und Praxis und Politik auf der anderen Seite zu verringern, weil ein höherer Rezeptionsgrad im Vergleich zu Einzelstudien angenommen werden kann (*Transferfunktion*). Diese Funktion können Forschungssynthesen allein deshalb besser als Einzelstudien leisten, weil eine halbwegs vollständige Rezeption der Einzelforschung für Akteure der Praxis oder Politik durch das wachsende Forschungsvolumen nicht mehr zu leisten ist. Mit der Erstellung einer repräsentativen Gesamtbilanz werden indes günstige Voraussetzungen für eine erleichterte

Tab. 3: Adressatengruppen mit ihren problematischen Interessen beim Transfer von wissenschaftlichen Befunden

Gruppe	Transferprobleme	„Problematische“ Interessen
Professionelle Praktiker in sozialwissenschaftlichen und pädagogischen Berufen	Übertragung mittlerer Ergebnisse auf den Einzelfall	Möglichst praxisnahe Beantwortung alltagstypischer Problemstellungen
Personen aus dem administrativ-politischen Bereich auf lokaler Ebene (z. B. Schul- und Jugendämter)	Wissenschaftliche Befunde konkurrieren mit anderen Interessen, Handlungszielen und einer etablierten Verwaltungsdynamik	Darstellung von Handlungsbereitschaft, Umsetzung von Ergebnissen vor dem Hintergrund begrenzter Ressourcen
Öffentlichkeit, Gesellschaft, über Medien adressiert	Hohe Komplexität von Forschungsergebnissen hinderlich, zum Teil geringe mediale Verwertbarkeit von Forschung	Medienwirksame Aufbereitung von Wissenschaft, auf „Verkauf“ ausgerichteter Nutzen von Erkenntnissen und Befunden
Akteure der Privatwirtschaft, die sich in sozialwissenschaftlichen oder pädagogischen Handlungsfeldern engagieren wollen	Wissenschaftliche Befunde widersprechen möglicherweise dem Marketing	Günstige Außendarstellung und mediale Verwertbarkeit der Aktivitäten, Umsetzung der Firmen „philosophie“
Entscheidungsträger auf Bundes- und Landesebene in der Bildungspolitik	Unterschiedliche Zielstruktur von Wissenschaft und Politik (Wahre Aussagen vs. Verhandlung multipler Interessen, Medienwirkung)	Verwendung von Erkenntnissen, Wissen und Befunden zu Durchsetzung eigener Interessen (symbolische Nutzung)

Rezeption sowie eine direkte oder instrumentelle Nutzung (Rich 1977) von Forschungsergebnissen geschaffen.

Beim Transfer wissenschaftlicher Erkenntnisse treten allerdings verschiedene Probleme auf, die vor allem mit unterschiedlichen und zum Teil problematischen Interessen der beteiligten Akteure zu tun haben (Jonas und Beelmann 2009). So existieren auf der Adressatenseiten von Forschungsergebnissen im Vergleich zur Absenderseite (Forschung) typischerweise nicht die gleichen Beweggründe für die Nutzung wissenschaftlicher Ergebnisse und den sich daraus ergebenen Handlungsoptionen. Tabelle 3 verdeutlicht exemplarisch die unterschiedlichen Interessenlagen verschiedener Akteursgruppen sowie die assoziierten Transferprobleme.

Nun wäre es wohl vermessen anzunehmen, dass Forschungssynthesen diese Interessenkonflikte kompensieren oder gar lösen könnten, zumal Transferprobleme oft vielschichtiger sind, als dies hier angedeutet werden kann. Beispielsweise setzt eine Umsetzung von wissenschaftlichen Erkenntnissen zwingend ein entsprechendes Interesse voraus, diese bei Handlungsentscheidungen überhaupt mit berücksichtigen zu wollen, und diese Voraussetzungen sind etwa bei politischen Entscheidungsträgern insbesondere bei einstellungskonträren Ergebnissen nicht unbedingt anzunehmen. Gleichwohl– und dies wäre ein genuiner Vorteil von Forschungssynthesen – dürften die verschiedenen Adressaten nicht mehr (oder nur unter erhöhten Argumentationsaufwand) in der Lage sein, wissenschaftliche Ergebnisse zu ignorieren oder sie symbolisch (Rich 1977), d. h. selektiv zur Stärkung

eigener Interessen zu nutzen, was unter Rückgriff auf ausgewählte Einzelstudien oftmals gelingt. Insofern ist der Vorteil integrativer Befunde aus Forschungssynthesen bei Transferproblemen vor allem darin zu sehen, dass bestimmte, auf Relativierung oder gar Diskreditierung wissenschaftlicher Befunde ausgerichtete Argumentationslogiken (z. B. „es gibt unterschiedliche Positionen", „andere Studien bestätigen das nicht") auf Seiten der Adressaten nicht mehr oder nur erschwert zu bedienen sind.

2 Probleme und Grenzen der Aussagekraft von Forschungssynthesen

Wie bei jeder wissenschaftlichen Methodik so bestehen bei Forschungssynthesen nicht nur besondere Potentiale, sondern auch verschiedene Probleme, die mit Grenzen ihrer Aussagekraft einhergehen (Cooper und Hedges 2009b). Vier Aspekte gefährden die grundsätzliche Validität integrativer Ergebniszusammenfassungen: Publikationsverzerrungen, die geringe oder unterschiedliche methodische Qualität der Primäruntersuchungen, Konfundierungen zwischen Ergebnismoderatoren und unangemessene Dateninterpretationen.

2.1 Publikationsverzerrungen

Ein erster Kritikpunkt bezieht sich auf die Auswahl der integrierten Studien. Speziell Meta-Analysen verfolgen die Absicht, möglichst *alle* verfügbare Evidenz zusammenzutragen oder zumindest eine repräsentative und besonders aussagekräftige Auswahl aller Untersuchungsergebnisse zu einer Fragestellung zu berücksichtigen. Publikationsverzerrungen beschreiben systematische Verletzungen dieser Annahmen, die durch eine eingeschränkte Studienauswahl oder eine geringere Publikationswahrscheinlichkeit bestimmter Studienergebnisse zustande kommen.

Eine Gefährdung durch Publikationszerrungen kann auf drei Arten einstehen. Erstens beschränken sich viele Forschungssynthesen auf die Auswertung veröffentlichter Publikationen und verzichten auf unveröffentlichte Forschungsberichte, Dissertationen und andere „graue" Literatur oder sie schränken ihre Literatursuchen so stark ein, dass bestimmte Dokumente mit einer geringeren Wahrscheinlichkeit identifiziert werden (z. B. Beschränkung auf Datenbankrecherchen, in denen bestimmte Journale oder Beiträge in Herausgeberbänden nicht registriert werden). Die meta-analytische Erfahrung zeigt, dass es zu einer systematischen Überschätzung der mittleren Effektstärke kommt, wenn unpublizierte Quellen (wie etwa Dissertationen und Projektberichte) unberücksichtigt bleiben. Daher gehören die Verwendung multipler Suchstrategien sowie systematische Anstrengungen, unpublizierte Studien zu beschaffen, heute zum Standard guter meta-analytischer Forschung. Den Anspruch allerdings, alle bislang durchgeführten Untersuchungen zu einer Thematik zusammenzufassen, konnte bislang wohl keine Forschungssynthese tatsächlich einhalten. Dies ist allein aufgrund *echt* unpublizierter und nicht weiter dokumentierter Arbeiten wahrscheinlich unmöglich. Die Literatursuche in Forschungssynthesen steht am Ende auch unter Kosten-Nutzen-Gesichtspunkten, das heißt der Suchaufwand muss in einem angemessenen Verhältnis zur Identifizierungsrate stehen. Konkrete oder spezifische Richtlinien liegen allerdings nicht vor, sodass es jedem

Autor überlassen bleibt, auf weitere Literatursuchen zu verzichten und damit das Risiko der Nichtentdeckung relevanter Studien einzugehen.

Eine besondere Form der Einschränkung der Studienauswahl liegt in der Begrenzung auf bestimmte Publikationssprachen (zumeist Englisch), weil die Beschaffung anderssprachiger Forschungen schwierig ist und ihre Rezeption entsprechende Sprachkenntnisse voraussetzt (sogenannter *language bias*). Da Englisch in zunehmendem Maße die Wissenschaftssprache ist, sind die zu erwartenden Ausfälle zumindest bei aktuellen Forschungsarbeiten allerdings gering. Dennoch zeigen Meta-Analysen, die auch andere Publikationssprachen berücksichtigen, dass damit ein nennenswerter Ausfall in Kauf genommen wird (Beelmann und Schulz in Vorbereitung). Da die Publikationssprache hoch mit dem Durchführungssetting von Studien korreliert (in Deutsch publizierte Studien sind mit hoher Wahrscheinlichkeit auch im deutschsprachigen Raum durchgeführt worden), ergeben sich bei Berücksichtigung unterschiedlicher Publikationssprachen zudem bessere Möglichkeiten für kulturvergleichende Untersuchungen. Daher sollte soweit möglich auch anderssprachige Literatur einbezogenen werden, zumal es sich bei der Publikationssprache um ein pragmatisches und inhaltlich wohl kaum zu begründendes Selektionskriterium handelt.

Ein zweites, weniger leicht zu lösendes Problem im Kontext von Publikationsverzerrungen besteht in der Tendenz, dass bestimmte Forschungsergebnisse nicht zugänglich gemacht werden. Typischerweise resultiert ein derartiges Publikationsverhalten aus dem Umstand, dass erwartungskonträre Ergebnisse erzielt wurden, z. B. in Interventionsstudien nicht die erhofften Effekte auftraten. Im Ergebnis resultiert eine geringere Publikationswahrscheinlichkeit nicht hypothesenkonformer Ergebnisse im Vergleich zu signifikanten und hypothesenkonformen Befunden. In Meta-Analysen führt dies zumeist zu einer negativen Korrelation zwischen der Stichprobengröße und den Befunden (Effektstärken), da in Studien mit geringem Stichprobenumfang die Ergebnisse eindeutiger sein müssen (z. B. ein Interventionseffekt ausgeprägter), um statistisch abgesichert zu werden. Treten so in „kleinen" Studien nur kleine (nicht-signifikante) Effekte auf, ist offenbar das Publikationsinteresse geringer als wenn die gleichen Effekte in „großen" Studien auftreten und allein aufgrund der Stichprobengröße signifikant werden.

Derartige Verzerrungen sind insbesondere in der Interventionsforschung nicht zu unterschätzen und können gravierende Auswirkungen auf die Aussagekraft von Meta-Analysen haben. Rothstein (2008) und Kollegen (Rothstein et al. 2005) haben sich daher eingehend mit Publikationsverzerrungen auseinander gesetzt. Um ihr Ausmaß näher abzuschätzen, wird heute zumeist eine grafische Darstellung vorgeschlagen, bei der die Präzision der Effektstärken (abgeleitet aus der Stichprobengröße oder davon direkt abhängige Kennwerte wie der Standardfehler der Effektstärke) gegen die ermittelten Ergebnisparameter abgetragen werden (sogenannter „funnel-plot"). Dabei müsste sich – ohne Publikationsverzerrungen – eine weitgehend symmetrische Verteilung um den mittleren Effekt herum ergeben, wobei aufgrund von Stichprobenfehlern mit einer geringen Effektstärkenvariation bei Studien mit großen Stichproben und einer hohen Effektstärkenvariation bei Studien mit kleinen Stichproben gerechnet wird. Ist die Verteilung dann nicht symmetrisch, fehlen also Studien in bestimmten Bereichen der Verteilung, kann von mehr oder weniger systematischen Verzerrungen durch unterschiedliche Publikationswahrscheinlichkeiten ausgegangen werden.

Auf diesen Analysen basierend sind im Anschluss unterschiedliche Strategien möglich, um Publikationsverzerrungen bei der Beurteilung der Ergebnisse zu berücksichtigen. Da der Rezeption unveröffentlichter Literatur Grenzen gesetzt sind (mangelnde Verfügbarkeit, keine oder unzureichende Dokumentation), hat bereits Rosenthal (1979) die Berechnung des sogenannten „fail-safe-*N*" vorgeschlagen. Dieser Index gibt die hypothetische Zahl von Studien mit Nulleffekten an, die existieren müsste, um einen integrativen Befund soweit zu inflationieren, dass er statistisch bedeutungslos würde. Das „fail-safe-*N*" erlaubt damit eine Einschätzung, ob eine verzerre Studienauswahl die ermittelten Befunde grundsätzlich gefährden könnte.

Duval und Mitarbeiter (Duval 2005; Duval und Tweedie 2000) haben weiterführend die sogenannte „Trim and Fill"-Analyse vorgeschlagen. Ausgangspunkt dieser Methode ist wiederum ein Funnel-plot und die Identifikation von Studienergebnissen, die keine Entsprechung in der Effektstärken-Verteilung aufweisen (z. B. eine Studie mit geringem Stichprobenumfang und hoher Effektstärke, der ein negatives Äquivalent auf der anderen Seite der Verteilung um den mittleren Effekt fehlt). Zu diesen Studien werden nun Verteilungsäquivalente angenommen (Studienergebnisse, die es nach den Verteilungsannahmen eigentlich geben müsste) und das durchschnittliche meta-analytische Ergebnis neu berechnet. Schließlich wird beurteilt, ob es sich um Publikationsverzerrungen handelt, die die ursprünglichen Befunde substanziell beeinträchtigen können (vgl. Beispiel unter Abschn. 2.3.1).

Die „Trim-and-fill"-Analyse beruht allerdings auf der Annahme, dass nicht publizierte Studienergebnisse in jedem Fall mit Publikationsverzerrungen zu tun haben. Die höhere Effektivität in Studien mit geringen Stichprobenumfängen kann aber zum Beispiel auch mit inhaltlichen Merkmalen der Studien kovariieren (z. B. in Interventionsstudien eine bessere Implementationsqualität bei Studien mit geringem Stichprobenumfang) und somit andere Ursachen haben als jene, die typischerweise bei Publikationsverzerrungen angenommen werden (zu weiteren Methoden siehe Rothstein 2008; Rothstein et al. 2005).

Ein drittes, noch schwieriger zu kontrollierendes Problem von Publikationsverzerrungen besteht darin, dass Autoren dazu neigen, nicht alle ermittelten Befunde, sondern nur signifikante oder besonders günstige Effekte zu berichten. Solange unerwünschte Ergebnisse zwar erwähnt, aber keinen exakten Statistiken vorliegen, können Umrechnungen praktiziert werden, um eine verzerrte oder unvollständige Ergebnisberichterstattung zu kompensieren. Problematischer ist jedoch der Fall, wenn Ergebnisse einfach ausgelassen werden. Gorman (2005) hat zum Beispiel für die Drogenprävention und sogenannten Life-skill-Programmen zeigen können, dass Autoren zu einer verzerrten und selektiven Berichterstattung in Publikationen neigen. Für den Fall, dass die Autoren aber nichtsignifikante Ergebnisse überhaupt nicht berichten, lassen sich zumeist keine leichten Kompensationen finden. Hier bleibt dem Meta-Analytiker allein die Option, die Projektberichte sehr aufmerksam zu lesen und ggf. zusätzliche Informationen und Publikationen zu nutzen, um mögliche Auslassungen abzuschätzen. In gravierenden Fällen könnte auch der vollständige Ausschluss entsprechender Studien erwogen werden. Nach Eisner (2009) werden solche Verzerrungen durch nicht-berichtete hypothesendiskonforme Befunde nicht nur auf Ebene der Autoren, sondern auch auf Ebene von Gutachtern und Herausgebern begünstigt. Kritisch schätzt dieser Autor in diesem Zusammenhang den Einfluss von Interessenkonflikten ein, wenn z. B. Programm-Entwickler ihre Interventio-

nen anbieten und bewerben, was mit positiven Evaluationsergebnissen sicher erfolgversprechender ist. Publikationsverzerrungen können somit einen beträchtlichen Einfluss auf Ergebnisbilanzen haben und sind insofern bei Forschungssynthesen stets zu überprüfen.

2.2 Geringe oder unterschiedliche methodische Qualität der Primärstudien

Schon früh ist gegen Meta-Analysen eingewandt worden, dass ihre Ergebnisse schweren Verzerrungen unterliegen, wenn die Integration ohne Berücksichtigung der methodischen Qualität der Primärstudien erfolgt (Eysenck 1978). Der Zusammenhang zwischen der methodischen Qualität einer Studie und ihren Ergebnissen ist allerdings eine empirische Frage und setzt zunächst voraus, dass sich die methodische Qualität reliabel beurteilen lässt. Dies erweist sich allerdings bei genauerer Betrachtung als relativ schwierig, da eine Reihe von methodischen Merkmalen zur Qualität beitragen (z. B. Designmerkmale, Art und Reliabilität der Datenerhebung, Stichprobenselektion, etc.). Grundsätzlich ist zunächst festzustellen, dass in der meta-analytischen Praxis eine Tendenz besteht, ohnedies nur methodisch vergleichsweise hochwertige Studien auszuwählen. Viele Analysen zur Interventionsforschung beschränken sich zum Beispiel allein auf randomisierte Kontrollstudien, die per se einen hohen methodischen Standard garantieren. Ob diese Auswahl immer gerechtfertigt ist, ist eine offene Frage. Verschiedene Autoren zeigen zum Beispiel, dass auch gute Quasi-Experimente zuverlässige Daten erzeugen können (Heinsman und Shadish 1996). Unabhängig von der Frage der methodischen Mindestkriterien bestehen in der meta-analytischen Literatur zur Kodierung der methodischen Qualität zwei grundsätzliche Ansätze: Methodische Globalratings oder die Beurteilung von methodischen Einzelparametern. Der Ansatz der methodischen Globalratings ist insbesondere durch Sherman und Mitarbeiter (vgl. Farrington 2003; Farrington et al. 2002) propagiert worden. Diese Autoren plädieren mit ihrer „Maryland-Scale“ dafür, die methodische Qualität vor allem am Forschungsdesign zu beurteilen. Danach bekommen randomisierte Experimente die höchste Kodierung, Quasi-Experimente eine mittlere und unkontrollierte Studien ein geringes Rating. Abgesehen davon, dass in Meta-Analysen zumeist eine Einschränkung auf bestimmte Forschungsdesigns stattfindet und insofern nur sehr wenig Varianz bei den integrierten Studien durch die Maryland-Scale erzeugt wird, ist der Ansatz der methodischen Globalratings auch relativ unspezifisch. Andere Arbeiten bewerten daher sehr unterschiedliche Kriterien und fassen diese Aspekte zu methodischen Summenscores zusammen oder berechnen Zusammenhänge zwischen zahlreichen Einzelkriterien und den Effektstärken (z. B. McLeod und Weisz 2004; Valentine 2009). Je nach Ergebnis sind im Anschluss unterschiedliche Strategien möglich. Einige Autoren plädieren dafür, Effektstärken hinsichtlich der methodischen Qualität zu korrigieren (Hunter und Schmidt 2004). Andere Autoren berechnen regressionsanalytisch korrigierten Effektstärken, je nachdem ob methodische Einflussgrößen einen signifikanten Einfluss hatten oder nicht (vgl. z. B. Lipsey und Wilson 1998).

Unabhängig von der Art der methodischen Beurteilung von Studien und der Frage, ob Studienergebnisse korrigiert werden sollten, finden sich empirisch bislang recht inkonsistente Befunde zum Zusammenhang von methodischen Merkmalen und Ergebnisparametern. Gleichwohl klären methodische Merkmale beträchtliche Varianzanteile von Studienergebnissen auf, wie Wilson und Lipsey (2001) in einer systematischen Auswer-

tung von Meta-Analysen zeigen konnten. Mit etwa einem Viertel der Ergebnisvarianz zeigte sich dabei eine ähnliche Größenordnung wie für inhaltliche Merkmale. Daraus folgt, dass die Einflüsse methodischer Merkmale der Primärstudien auf meta-analytische Ergebnisse jeweils mit zu berücksichtigen sind, da ansonsten massive Fehlschlüsse hinsichtlich inhaltlicher Interpretationen drohen.

2.3 Konfundierungen von Moderatoren und Grenzen der Aussagekraft

Über Potentiale von Meta-Analysen, inhaltliche und methodische Moderatoren eines Zusammenhangs zu ermitteln, wurde bereits gesprochen. Diesen Potentialen stehen andererseits beträchtliche Risiken gegenüber, meta-analytische Daten zu differentiellen Ergebnisanalysen zu nutzen. Dies hängt vornehmlich damit zusammen, dass Ergebnismoderatoren in der Regel systematisch konfundiert sind und somit eigentlich nicht unabhängig voneinander untersucht werden können. Zum Beispiel zeigen Analysen zur Präventionsforschung, dass der Präventionstyp (universell vs. gezielt) systematisch mit dem Alter der Zielgruppen kovariiert. Bei jüngeren Gruppen wird vornehmlich populationsbezogen, also universell, bei älteren Gruppen vornehmlich mit Risikogruppen, also gezielt, gearbeitet. Einfache Gegenüberstellungen nach dem Alter kommen dann zu höheren Effektstärken bei älteren Zielgruppen (was eigentlich der Präventionslogik widerspricht), allerdings nur, weil gezielte Präventionsmaßnahmen insgesamt höhere Wirkungen aufweisen. Ein Rückschluss auf die moderierende Wirkung von Alter wäre danach nicht zulässig (vgl. Beelmann und Raabe 2009).

Das Problem der Konfundierung ist in nahezu allen Meta-Analysen gegeben, weil sich wissenschaftliche Studien aufeinander beziehen und bestimmte Kombinationen von Studienmerkmalen aus inhaltlichen Gründen nicht gleich wahrscheinlich sind. Eine Berücksichtigung dieser Konfundierungen bei der Datenanalyse und Interpretation ist daher unerlässlich, stößt aber an ihre Grenzen, wenn die Zahl potentieller Moderatoren groß und die Zahl der Primärstudien relativ gering ist. Eine simple Analyse kategorialer Variablen (z. B. Vergleich der Ergebnisse hinsichtlich zweier Altersgruppen oder Schulformen) führt somit oft zu verzerrten Ergebnissen, weil zahlreiche Variablen zeitgleich beteiligt sind (vgl. Lipsey 2003). Meta-Regressionsmodelle bieten Möglichkeiten, die jeweils varianzstärksten Moderatoren auszuwählen. Zudem erlauben diese Analysen die Korrektur von Effektstärken im Hinblick auf methodische und inhaltliche Konfundierungen. Die Einschätzung der relativen Bedeutung von Variablen sowie die Berechnung korrigierter Effektstärken auf Basis eines Meta-Regressionsmodells ist ein bedeutsamer Vorteil im Vergleich zu Primärstudien. Wie bei kategorialen Vergleichen sind allerdings auch Grenzen vorhanden, z. B. bei einer geringen Anzahl integrierter Primärstudien, einer hohen Anzahl konfundierter Merkmale oder grundsätzlich geringer Varianzaufklärung durch die betrachteten Moderatoren.

2.4 Interpretation der Ergebnisse

Ein letztes Problem betrifft die Interpretation meta-analytischer Ergebniszusammenfassungen. Eine systematische Datenkollektion und elaborierte statistische Auswertungsmodelle suggerieren zwar eindeutige Interpretationen, vernachlässigen jedoch, dass

empirische Daten (gleich auf welchem Abstraktionsniveau) immer Interpretationsspielräume bieten. Da aber Meta-Analysen stärker als die Ergebnisse der Primärforschung zur Information von politischen Entscheidungsträgern und der Gesellschaft genutzt werden sollten, existieren verschiedene Möglichkeiten, die klassischen Effektstärkenparameter in anschaulichere Maße umzuwandeln. Relativ bekannt ist bei Evaluationsdaten die Umrechnung in prozentuale Verbesserungsraten, wie sie bereits im Binominal Effect Size Display von Rosenthal und Rubin (1982) zum Ausdruck kommt (z. B. $d=0{,}80$ entspricht einer Korrelation von $r=0{,}37$ und damit einer Verbesserungsrate von 37 %). Cohen (1988) hat vorgeschlagen, Effektstärken in klein ($d=0{,}2$, $r=0{,}1$), mittel ($d=0{,}5$, $r=0{,}2$) und groß ($d=0{,}8$, $r=0{,}4$) einzuteilen. Diese Daumenregel wird in der meta-analytischen Forschung häufig zitiert, hat allerdings den Nachteil, dass die Effektstärkeninterpretation ohne Berücksichtigung inhaltlicher Überlegungen vorgenommen wird. So haben verschiedene Autoren gezeigt, dass zum Beispiel kleine Effekte etwa bei geringer Programmintensität durchaus eindrucksvoll sein können (Ellis 2010; Prentice und Miller 1992). Zudem hängt die Interpretation immer von empirischen Erwartungen und speziell von den erhobenen Variablen ab. Wir konnten zum Beispiel in einer Evaluationsstudie zeigen, dass eine kleine Effektstärke für ein Eltern- und Kindertraining ($d=0{,}20$) immerhin einer Reduktion von 72 % an Kindern mit problematischen Verhalten in der ersten beiden Schulklassen entsprach (vgl. Lösel et al. 2009). Insofern führt die oben genannte Daumenregel zumeist zu wenig reflektierten Überlegungen und sollte durch stärker inhaltsbezogene Interpretationen mindestens ergänzt werden (vgl. Hill et al. 2008).

3 Forschungssynthesen in der Bildungsforschung: Ausgewählte Beispiele

3.1 Beispiel 1: Prävention von Verhaltensproblemen bei Kindern und Jugendlichen

Ein besonders umfangreiches Feld der meta-analytischen Forschung ist die Prävention von Verhaltensproblemen bei Kindern und Jugendlichen (vgl. Beelmann und Raabe 2009; Beelmann et al. 2014). Zahlreiche Meta-Analysen befassen sich mit der Wirksamkeit von Förderprogrammen für Kinder, Eltern, Lehrer oder kombinierter Maßnahmen. Sie machen zugleich deutlich, welche enormen Forschungsaktivitäten in den letzten 20 Jahren in diesem Bereich zu verzeichnen sind.

Neben zahlreichen wichtigen inhaltsbezogenen Ergebnissen konnten die vorhandenen Meta-Analysen dabei vor allem eine Reihe von Wirksamkeitsproblemen offenlegen. Ein Befund war etwa, dass Replikationsstudien zu an sich wirksamen Programmen selten jene Wirksamkeit erbrachten, die im ursprünglichen Setting der Erstuntersuchungen auftrat. Ein gutes Beispiel dafür ist das international weit verbreitete Olweus-Gewaltpräventionsprogramm (Olweus 2006), für das in internationalen Replikationen bislang nie die Wirksamkeit bestätigt werden konnte, die in Studien der Programm-Autoren erreicht wurde. Dieser Befund macht deutlich, dass eine simple Übertragung von Interventionsprogrammen, die in bestimmten sozialen Kontexten entwickelt und geprüft wurden, nicht ohne weiteres möglich ist.

Ergebnisse aus Meta-Analysen zeigten ferner, dass Implementationsparameter (z. B. die Mitarbeit der Adressaten, das Engagement der Programm-Administratoren, Rahmen-

bedingungen der Programm-Durchführung) über Programm-Inhalte hinaus ganz wesentlich an der Wirksamkeit von Präventionsprogrammen mitwirken. So haben beispielsweise Durlak et al. (2011) in einer Analyse von universellen Präventionsmaßnahmen in Schulen gezeigt, dass die Wirksamkeit beträchtlich nachließ, wenn Implementationsprobleme berichtet wurden. Diese Befunde waren besonders wertvoll, weil der Vergleich unterschiedlicher Implementationsbedingungen in Primärstudien nur unter sehr hohem Aufwand zu realisieren ist. Die Implementierung und die Implementationsbedingungen sollten somit stärker als bislang bereits in die Konstruktion von Präventionsmaßnahmen einfließen (vgl. Beelmann 2011).

Ein drittes Ergebnis ist nicht allein auf Meta-Analysen zur Präventionsforschung begrenzt, trifft dort jedoch in besonderem Maße zu. So zeigte sich, dass die Wirksamkeit von Intervention in hohem Maße von der Art der Effektkriterien abhängig war. Da aber Primärstudien in der Erhebung von Erfolgskriterien oftmals begrenzt sind, bieten Meta-Analysen hier deutliche verbesserte Möglichkeiten, die Bandbreite unterschiedlicher Wirkungen in inhaltlicher und erhebungsmethodischer Sicht abzubilden. So konnte u. a. gezeigt werden, dass Präventionsmaßnahmen ihre Wirkungen vor allem auf Wissensparameter entfalten, während die Effekte auf die eigentlichen Präventionsanlässe (z. B. Aggression, Suchtverhalten) deutlich geringer ausfielen (Beelmann et al. 2014). Solche Befunde stellen die Rationalität von Präventionsmaßnahmen erheblich in Frage und sollten idealerweise zur Weiterentwicklung der Programme führen.

Schließlich ergaben sich in verschiedenen Meta-Analysen massive Publikationsverzerrungen und ein Bias in eine positive Richtung, der vermutlich mit den besonderen Legitimationsbedingungen von Prävention zu tun hat (proklamierter Einsatz ohne Existenz eines aktuell vorhandenen Problems) und vermutlich massiv die Publikationswahrscheinlichkeit und damit die mittleren Befunde beeinflusst. In der Zusammenfassung der deutschsprachigen Präventionsforschung zeigte sich beispielsweise, dass praktisch keine negativen Befunde auftraten (Beelmann et al. 2014), was weder inhaltlich zu erwarten war noch den oben genannten Verteilungsannahmen der statistischen Integrationsmodelle entsprach. Nun sind auch tatsächliche oder vermeintliche Publikationsverzerrungen unterschiedlich zu interpretieren (z. B. selektive Berichterstattung vs. besonders gute Implementationsbedingungen in „kleinen“ Studien; s. o. sowie Beelmann 2011). Diese Ergebnisse begründen gleichwohl große Zweifel an der Validität der Befunde und sind darüber hinaus für die wissenschaftliche Publikationspraxis nicht besonders schmeichelhaft.

Abbildung 1 zeigt die entsprechende „Trim-and-Fill“-Analyse zur deutschsprachigen Präventionsforschung bei Kindern und Jugendlichen (vgl. Beelmann et al. 2014). Hier trat eine signifikante Asymmetrie der Studieneffekte auf (Egger-Test: $z = 7{,}10$, $p < 0{,}001$) und die anschließende „Trim-und Fill“-Analyse berechnete insgesamt 38 zusätzliche Studien zu den 160 empirisch vorhanden Studien (R0-Parameter). Diese Ergänzung würde die mittlere Effektstärke des Ausgangsdatensatzes ($d = 0{,}25$) substanziell reduzieren ($d = 0{,}12$), allerdings nicht soweit, dass diese korrigierte Effektstärke nicht mehr von Null verschieden ist, somit also keinen signifikanten Effekt ergeben hätte.

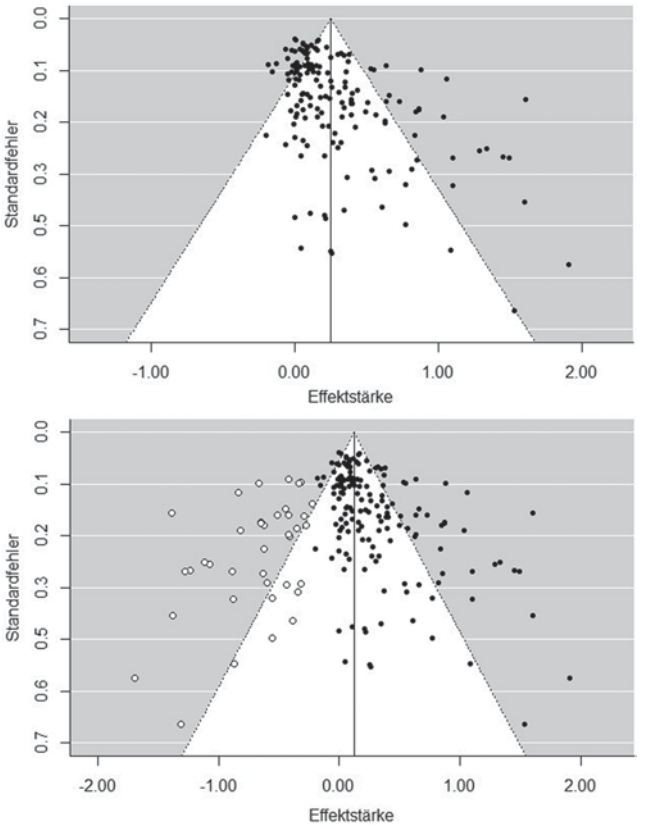

Abb. 1: Ergebnisse einer „Trim and Fill" –Analyse zu Ergebnissen der deutschen Präventionsforschung bei Kindern und Jugendlichen ohne (*oben*) und mit (*unten*) ergänzten Studienergebnissen (vgl. Beelmann et al. 2014). (Anmerkungen: ● = Studienergebnisse des Originaldatensatzes ($k = 160$); ○ = durch „Trim and Fill" ergänzte Studien (R0-Modell, $k = 38$). Weiß markierte Fläche entspricht jeweils dem 95 %igen Konfidenzintervall)

3.2 Beispiel 2: Effektivität von Förderprogrammen bei Kindern und Jugendlichen mit Lernstörungen

Grünke (2006) hat verschiedene Meta-Analysen zur Förderung von Kindern und Jugendlichen mit Lernstörungen zusammengetragen. Seine Synopse umfasste 26 Arbeiten aus den Bereichen der kognitiven, akademischen, sprachlichen, motivationalen und sozialen Förderung. Inhaltlich konnte durch diese Zusammenstellung vor allem gezeigt werden, dass bestimmte interventionsmethodische Vorgehensweisen zu präferieren sind. So ergab sich über die genannten Förderbereiche hinweg der relativ konstante Befund, dass ein strukturiertes, geplantes und interaktives Vorgehen (wie z. B. bei kognitiven Strategie- oder Selbstinstruktionstrainings) einer freien, entdeckenden, konstruktivistischen Vorgehensweise in den meisten Fällen überlegen war. Glaubt man Grünke (2006) ist dieses Ergebnis vor allem deshalb praxisrelevant, weil es nicht dem aktuellen Verbreitungsgrad von Förderprogrammen oder Lernstrategien und der vorherrschenden Interventionsmentalität in der Praxis entspricht.

Selbstverständlich stellt dieser Befund zunächst nur ein mittleres Ergebnis im genannten Anwendungsfeld dar. In einem konkreten Anwendungssetting müssen solche Handlungsheuristiken immer auf konkrete Zielgruppen oder Einzelfälle heruntergebrochen

werden. Gleichwohl liefern derartige Ergebnisse eine forschungsbasierte Grundorientierung auf einer höheren Abstraktionsebene, von der im Einzelfall nur mit guten Gründen abgewichen werden sollte. In diesem Sinne geht es in Forschungssynthesen und speziell in Meta-Analysen auch nicht nur um „mittlere Befunde", sondern im Sinne der o. g. Strukturierungsfunktion auch um die Entwicklung allgemeinerer Handlungsprinzipien oder neuerer, höherwertigere Handlungsheuristiken oder -theorien.

Was die Probleme und Grenzen der Aussagekraft von meta-analytischen Daten angeht, so werden bei Grünke (2006) ebenfalls zentrale Gefährdungen benannt. In methodenkritischen Reflexionen geht der Autor zum Beispiel darauf ein, dass der methodische Standard der integrierten Untersuchungen insgesamt dürftig ausfiel und inhaltlich zum Teil sehr unterschiedliche Verfahren in gemeinsamen Auswertungskategorien ausgewertet wurden. Ebenso spricht der Autor von erheblichen Konfundierungen und Abhängigkeiten, die differentielle Aussagen zur Wirksamkeit nur sehr bedingt zuließen. Dies wird allein dadurch deutlich, dass in manchen Meta-Analysen kategoriale Gegenüberstellungen vorgenommen wurden, bei denen einzelne Kategorien nur „einfach" besetzt waren (d. h. nur eine empirische Untersuchung vorlag). Wie in anderen Bereichen auch, sollten sich generalisierende Aussagen im Rahmen von Forschungssynthesen auf eine ausreichende Datenbasis beziehen. Welche Größenordnung dabei hinreichend ist, hängt auch von der Spezifität der konkreten Fragestellung und der Heterogenität eines Forschungsfeldes ab. Grundsätzlich werden natürlich Evidenzen und Erkenntnisse umso zuverlässiger, je breiter die Datenbasis ist. Da stellen Forschungssynthesen keine Ausnahmen dar.

3.3 Beispiel 3: Einflussfaktoren auf effektives Lernen

Eine besonders eindrucksvolle Zusammenstellung von meta-analytischen Arbeiten hat John Hattie (2009, deutsch: 2013; Pant 2014 in diesem Band) vorgelegt. Dieser Autor hat die Ergebnisse von über 800 Meta-Analysen zusammengetragen, die Einflussfaktoren effektiven Lernens untersucht hatten. Angesichts der Tatsache, dass in dieser Zusammenstellung mehrere hundert Meta-Analysen mit den Ergebnissen tausender empirischer Studien zusammengefasst wurden, kann an dieser Stelle selbstverständlich keine umfassende Auseinandersetzung mit dieser Arbeit erfolgen (vgl. dazu z. B. Beywl und Zierer 2013; Pant 2014 in diesem Band). Wir wollen uns daher allein auf die Frage beschränken, inwieweit die zusammengetragenen Befunde tatsächlich die Potentiale von Forschungssynthesen nutzen konnten und keine Grenzen der Aussagekraft überschritten wurden.

Zunächst wird man feststellen müssen, dass die Arbeit ein großes Publikum erreicht hat. Unzählige Diskussionen, Buchbesprechungen und Veröffentlichungen in den Printmedien zeugen davon, dass diese Ergebnisintegration ihre Wirkungen auf die Berufspraxis und Bildungspolitik nicht verfehlt hat, was nicht zuletzt mit einer adressatengerechten Präsentation in Barometerabbildungen und einfachen Ranglisten von Faktoren zu tun hat. Aber sind die Ergebnisse auch zuverlässig? Oder besser: Wurden bei der Zusammenstellung ggf. zu grobe Verallgemeinerungen produziert, die der Forschungsstand in dieser Form nicht hergibt?

In der Tat ist eine simple Gegenüberstellung meta-analytischer Befunde in der Regel wenig hilfreich, insbesondere wenn die Einflüsse methodischer Variablen und inhaltliche Konfundierungen nicht zugleich mit bedacht werden. Leider finden sich im Text nur sehr

wenige Hinweise darauf, ob und inwieweit forschungsmethodische Aspekte die gefundenen Ergebnisse in den zusammengetragenen Meta-Analysen beeinflusst haben. Dies gilt im Übrigen auch für methodische Aspekte des meta-analytischen Vorgehens selbst, denn selbstverständlich haben zum Beispiel die Art und Umfang der Literatursuche, das zugrundeliegende Integrationsmodell und anderes Auswirkungen auf die mittlere Effektstärke. Nun fasst Hattie (2013) jeweils Meta-Analysen zusammen, die zum Teil auf einer beträchtlichen Anzahl von Untersuchungen beruhen. Wir sollten daher erwarten, dass sich methodische Artefakte am Ende gegenseitig ausgleichen. Ist dies eine realistische Annahme? Nach allem, was insbesondere aus der Interventionsforschung bekannt ist, nein. Wir haben, wie bereits dargestellt, in vielen Studien je nach Interessenlage auch tendenziöse Darstellungen, die durch Auslassung, Wahl bestimmter methodischer Vorgehensweisen usw. entstehen können. Allein eine Auswertung zum Zusammenhang zwischen der mittleren Effektstärke in den analysierten Meta-Analysen und der Anzahl der integrierten Primärstudien vorzunehmen, wie es der Autor zum Thema Publikationsverzerrungen macht (Hattie 2013, S. 25), ist nicht ausreichend. Es wäre zu dokumentieren gewesen, inwieweit die einzelnen Meta-Analysen mögliche Publikationsverzerrungen untersuchten und bei der Interpretation ihrer Daten mitberücksichtigten. Im Grunde fehlt allgemein eine Dokumentation methodischer Analysen der einzelnen Meta-Analysen, denn nur bei vergleichbarer forschungsmethodischer Qualität oder zumindest bei etwa vergleichbaren methodischen Einflüssen auf die Ergebnisse wären die einzelnen Lernfaktoren im Rahmen meta-analytischer Daten vergleichbar gewesen. Dazu würde auch ein Verzeichnis der methodischen Selektionskriterien der einzelnen Meta-Analysen gehören, die zumindest eine grobe Information über mögliche validitätseinschränkenden Faktoren und Konfundierungen der Ergebnisse mit methodischen Merkmalen der Primärstudien gegeben hätte.

Was für methodische Merkmale versäumt wurde, gilt leider auch für inhaltliche Konfundierungen. Zwar werden die Daten zu den einzelnen Faktoren im Text differenziert diskutiert. Eine systematische Analyse von Konfundierungen zwischen inhaltlichen Variablen findet jedoch nicht statt, denn selbstverständlich sind die von Hattie (2013) zusammengetragenen inhaltlichen Faktoren nicht unabhängig, wenn man sich zum Beispiel allein die vielfältigen Zusammenhänge zwischen Schulfaktoren (z. B. Schultyp), Faktoren bei den Lehrpersonen (Ausbildung, Fachkompetenz) und den Unterrichtsmethoden (z. B. Lehrstrategien) vergegenwärtig. Das Problem der Unabhängigkeit der Datensätze beginnt aber bereits bei der Ergebniszusammenfassung, bei der die jeweils existierenden Überlappungen in den meta-analytischen Datensätzen offenbar unberücksichtigt blieben (d. h. die Gesamtzahl der Primärstudien zu einem Einflussfaktor ergibt sich aus der Summe der Primärstudien der einzelnen Meta-Analysen ohne zu überprüfen, ob nicht zumindest partiell dieselben Studien in die Arbeiten eingingen). Auf diese Weise wird ein Forschungsvolumen dokumentiert, das real nicht existiert und ggf. massive Fehleinschätzungen begünstigt.

Wir können also mit guten Gründen feststellen, dass wichtige Probleme und Grenzen von Forschungssynthesen in der Arbeit von Hattie (2013) nicht ausreichend berücksichtigt wurden. Aber entwerten diese Unterlassungen die grundsätzliche Validität der Aussagen? Angesichts der Fülle von Ergebnissen, Primärstudien und Meta-Analysen, die in die Zusammenstellung eingingen, ist wohl nicht zu erwarten, dass Aussagen, die sich auf

die mittlere Effektivität beziehen, grobe Fehleinschätzungen beinhalten. Gleichwohl ist davon ausgehen, dass nicht alle berichteten Ergebnisse vergleichbar zuverlässig sind und somit die vorgenommenen Gegenüberstellungen von Einflussfaktoren auf das Lernen in Ranglisten mit großer Vorsicht zu genießen sind. Dies kann im Einzelfall einen bestimmten Befund tatsächlich überbewerten und zu falschen Schlussfolgerungen führen.

Im Kern betreffen die genannten Probleme die grundsätzliche Frage, in welchem Ausmaß Komplexitätsreduktion bei Ergebniszusammenfassungen betrieben werden sollte und ab welchem Schwellenwert ein Ergebnis differenzielle Erkenntnisse derart reduziert, dass fehlerhafte Schlussfolgerungen abgeleitet werden. Führen Forschungssynthesen zu allgemeinen, gleichsam übergeordneten Erkenntnissen, werden Potentiale hervorragend genutzt. Werden – wie in diesem Beispiel – einfache Ranglisten produziert, sind Fehlschlüsse vorprogrammiert, denn der Teufel steckt – wie immer – im Detail. Notwendig wäre gewesen, die Ergebnisse intensiver im Hinblick auf die Grenzen forschungssynthetischer Arbeiten zu reflektieren und vielleicht ein anderes Dokumentationsformat zu wählen. Hattie (2013) gebührt aber mindestens der Verdienst, ein wichtiges Thema der Bildungsforschung mithilfe von Forschungssynthesen popularisiert und eine große Menge an Untersuchungsergebnisse strukturiert zur Verfügung gestellt zu haben.

4 Fazit/Schlussfolgerungen

Forschungssynthesen und speziell Meta-Analysen zielen auf die Erstellung einer repräsentativen Gesamtbilanz von Forschungsergebnissen zu einer Fragestellung unter gleichzeitiger Berücksichtigung von methodischen und inhaltlichen Faktoren, die zur Ergebnisvariabilität beitragen können. Mit dieser Zielsetzung liefern Forschungssynthesen einen wichtigen Beitrag zur Verbreitung und Nutzung sozialwissenschaftlicher Erkenntnissen. Die genannten Beispiele verdeutlichen jedoch zugleich Risiken, die bei einer unreflektierten Verwendung ihre Ergebnisse auftreten. Aus diesem Grund sollten folgende Fragen stets zur Beurteilung der Qualität von Forschungssynthesen gestellt werden:

1. Inwieweit wurde versucht, ein umfassendes Bild zu einer Forschungsfrage zu erstellen? Wurden alle relevanten Informationsquellen genutzt und mögliche Publikationsverzerrungen angesprochen oder geprüft?
2. In welchem Ausmaß waren die Ergebnisse durch methodische Merkmale der Primärstudien beeinflusst? Liegen entsprechende Auswertungen vor oder wurden ggf. auch Korrekturen vorgenommen?
3. Wurden bei der Identifikation von inhaltlichen und methodischen Einflussfaktoren mögliche Konfundierungen berücksichtigt?
4. Wurde versucht aus den aggregierten Daten übergeordnete Erkenntnisse (z. B. Interventionsprinzipien und Handlungsheuristiken) abzuleiten?
5. Wurden vor dem Hintergrund von Publikationsverzerrungen, methodischer Einflussgrößen und der Konfundierungsproblematik ggf. zu grobe Vereinfachungen des Forschungsstandes abgeleitet, die zu falschen Schlussfolgerungen führen können?

6. Wurden die Ergebnisse adressatengerecht vorgetragen, d. h. unterschiedlich dargestellt, um zu einem verbesserten Transfer beizutragen?

Werden Standards der integrativen Forschung eingehalten, Probleme ihrer Methodik bedacht und Grenzen der Aussagekraft berücksichtigt, liegen vielfältige Potentiale vor, insbesondere zum Transfer von Forschungsevidenzen in praktische Kontexte und politische Entscheidungsprozesse beizutragen. Eine Reihe von Arbeiten liefern hierzu auch in der Bildungsforschung bereits heute wichtige Beiträge. Für ihre Umsetzung müssen allerdings die potentiellen Adressaten gewillt sein, die erstellten Forschungsbilanzen zur Kenntnis zu nehmen und bei Entscheidungsprozessen mit zu berücksichtigen.

Literatur

APA (American Psychological Association) Publication and Communication Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology. Why do need them? What might they be? *American Psychologist, 63*(9), 839–851.

Beelmann, A. (2011). The scientific foundation of prevention. The status quo and future challenges of developmental crime prevention. In T. Bliesener, A. Beelmann, & M. Stemmler (Hrsg.), *Antisocial behavior and crime. Contributions of developmental and evaluation research to prevention and intervention* (S. 137–164). Cambridge: Hogrefe Publishing.

Beelmann, A., & Raabe, T. (2009). The effects of preventing antisocial behavior and crime in childhood and adolescence: Results and implications of research reviews and meta-analyses. *International Journal of Developmental Science, 3*(3), 260–281.

Beelmann, A., Pfost, M., & Schmitt, C. (2014). Prävention und Gesundheitsförderung bei Kindern und Jugendlichen: Eine Meta-Analyse der deutschsprachigen Wirksamkeitsforschung. *Zeitschrift für Gesundheitspsychologie, 22*(1), 1–14. doi:10.1026/0943-8149/a000104.

Beelmann, A., & Lipsey, M. W. (in Druck). Meta-analysis of effect estimates from multiple studies. In M. W. Lipsey & D. S. Cordray (Hrsg.), *Field experimentation: Methods for evaluating what works, for whom, under what circumstances, how, and why*. Thousand Oaks: Sage.

Beywl, W., & Zierer, K. (2013). Lernen sichtbar machen: Zur deutschsprachigen Ausgabe von „Visible Learning“. In J. Hattie (Hrsg.), *Lernen sichtbar machen* (S. VI–XXVI). Baltmannsweiler: Schneider.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). New York: Academic Press.

Cook, T. D. (1991). Meta-analysis: Its potential for causal description and causal explanation within program evaluation. In G. Albrecht & H. U. Otto (Hrsg.), *Social prevention and the social sciences* (S. 245–285). Berlin: de Gruyter.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.

Cook, T. D., Cooper, H. M., Cordray, D. S., Hartman, H., Hedges, L. V., Light, T. A., & Mosteller, F. (1992). *Meta-analysis for explanation: A casebook*. New York: Russell Sage Foundation.

Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research, 52*(2), 291–302. doi:10.3102/00346543052002291.

Cooper, H. M. (1988). Organizing knowledge synthesis: A taxonomy of literature reviews. *Knowledge in Society, 1,* 104–126.

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4. Aufl.). Thousand Oakes: Sage.

Cooper, H. M., & Hedges, L. V. (2009a). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges, & J. C. Valentine (Hrsg.), *Handbook of research synthesis and meta-analysis* (2 Aufl., S. 3–16). New York: Russell Sage Foundation.

Cooper, H. M., & Hedges, L. V. (2009b). Potentials and limitations. In H. Cooper, L. V. Hedges & J. C. Valentine (Hrsg.), *Handbook of research synthesis and meta-analysis* (2. Aufl., S. 561–572). New York: Russell Sage Foundation.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Hrsg.) (2009). *Handbook of research synthesis and meta-analysis* (2. Aufl.). New York: Russell Sage Foundation.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82*(1), 405–432. doi:10.1111/j.1467-8624.2010.01564.x.

Duval, S. (2005). The „trim and fill" method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Hrsg.), *Publication bias in meta-analysis: Prevention, assessment, and adjustment* (S. 127–144). Chichester: Wiley.

Duval, S. J., & Tweedie, R. L. (2000). Trim and fill: A simple funnel plot-based method for testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463.

Eisner, M. (2009). No effects in independent prevention trials: Can we reject the cynical view? *Journal of Experimental Criminology, 5,* 163–183. doi:10.1007/s11292-009-9071-y.

Eisner, M., & Humphreys, D. (2011). Measuring conflict of interest in prevention and intervention research. A feasibility study. In T. Bliesener, A. Beelmann, & M. Stemmler (Hrsg.), *Antisocial behavior and crime. Contributions of developmental and evaluation research to prevention and intervention* (S. 165–180). Cambridge: Hogrefe.

Ellis, P. D. (2010). *The essential guide to effect sizes. Statistical power, meta-analysis, and the interpretation of research results*. Cambridge: Cambridge University Press.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist, 33*(5), 517. doi:10.1037/0003-066X.33.5.517.a.

Farrington, D. P. (2003). Methodological quality standards for evaluation research. *Annals of the American Academy of Political and Social Science, 587*(1), 49–68. doi:10.1177/0002716202250789.

Farrington, D. P., Gottfredson, D. C., Sherman, L. W., & Welsh, B. C. (2002). The Maryland scientific methods scale. In L. W. Sherman, D. P. Farrington, B. C. Welsh, & D. L MacKenzie (Hrsg.), *Evidence-based crime prevention* (S. 13–21). London: Routledge.

Gorman, D. M. (2005). Drug and violence prevention: Rediscovering the critical rational dimension of evaluation research. *Journal of Experimental Criminology, 1,* 39–62. doi:10.1007/s11292-004-6461-z.

Grünke, M. (2006). Zur Effektivität von Fördermethoden bei Kindern und Jugendlichen mit Lernstörungen. Eine Synopse vorliegender Metaanalysen. *Kindheit und Entwicklung, 15*(4), 239–254. doi:10.1026/0942-5403.15.4.239.

Hattie, J. (2009). *Visible learning*. A synthesis of over 800 meta-analyses relating to achievement. Abington: Routledge.

Hattie, J. (2013). *Lernen sichtbar machen*. Deutschsprachige Ausgabe von „Visible learning". Baltmannsweiler: Schneider.

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

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

Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers form randomized experiments? *Psychological Methods, 1,* 154–169. doi:10.1037/1082-989X.1.2.154.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Developmental Perspectives, 2*(3), 172–177. doi:10.1111/j.1750-8606.2008.00061.x.

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

Jonas, K. J., & Beelmann, A. (2009). Einleitung: Begriffe und Anwendungsperspektiven. In A. Beelmann & K. J. Jonas (Hrsg.), *Diskriminierung und Toleranz. Psychologische Grundlagen und Anwendungsperspektiven* (S. 19–40). Wiesbaden: Verlag für Sozialwissenschaften.

Lipsey, M. W. (2003). Those confounded moderators in meta-analysis: Good, bad, and ugly. *The Annals of the American Academy of Political and Social Science, 587,* 69–81. doi:10.1177/0002716202250791.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis. *American Psychologist, 48*(12), 1181–1209.

Lipsey, M. W., & Wilson, D. B. (1998). Effective intervention for serious juvenile offenders: A synthesis of research. In R. Loeber & D. P. Farrington (Hrsg.), *Serious and violent juvenile offenders: Risk factors and successful interventions* (S. 313–345). Thousand Oaks: Sage.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks: Sage.

Lösel, F. (2009). Förderung einer evidenz-basierten Politik durch systematische Forschungssynthesen. Die Campbell Collaboration. *Psychologische Rundschau, 60*(4), 246–247. doi:10.1026/0033-3042.60.4.246.

Lösel, F., & Beelmann, A. (2003). Effects of child skills training in preventing antisocial behavior: A systematic review of randomized evaluations. *Annals of the American Academy of Political and Social Science, 587,* 84–109. doi:10.1177/0002716202250793.

Lösel, F., Stemmler, M, Jaursch, S., & Beelmann, A. (2009). Universal prevention of antisocial development. Short- and long-term effects of a child- and parent-oriented program. *Monatsschrift für Kriminologie und Strafrechtsreform, 92*(2–3), 289–307.

McLeod, B. D., & Weisz, J. R. (2004). Using dissertations to examine potential bias in child and adolescent clinical trials. *Journal of Consulting and Clinical Psychology, 72*(2), 235–251. doi:10.1037/0022-006X.72.2.235.

Olweus, D. (2006). *Gewalt in der Schule. Was Lehrer und Eltern wissen sollten – und tun können*. Bern: Huber.

Pant, H. A. (2014). Aufbereitung von Evidenz für bildungspolitische und pädagogische Entscheidungen: Metaanalysen in der Bildungsforschung. In R. Bromme & M. Prenzel (Hrsg.), *Von der Forschung zur evidenzbasierten Entscheidung: Die Darstellung und das öffentliche Verständnis der empirischen Bildungsforschung* (in diesem Heft). Sonderheft der *Zeitschrift für Erziehungswissenschaft*. Wiesbaden: Springer VS.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112*(1), 160–164. doi:10.1037/0033-2909.112.1.160.

Rich, R. F. (1977). Use of social science information by federal bureaucrats. Knowledge for action versus knowledge for understanding. In C. H. Weiss (Hrsg.), *Using social research in public policy making* (S. 199–233). Lexington: Lexington Books.

Rosenthal, R. (1979). The „file drawer problem“ and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641. doi:10.1037/0033-2909.86.3.638.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*(2), 166–169. doi:10.1037/0022-0663.74.2.166.

Rothstein, H. R. (2008). Publication bias as a threat to the validity of meta-analytic results. *Journal of Experimental Criminology, 4,* 61–81. doi:10.1007/s11292-007-9046-9.

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

Schulze, R. (2007). Current methods for meta-analysis: Approaches, issues, and developments. *Zeitschrift für Psychologie, 215*(2), 90–113. doi:10.1027/0044-3409.215.2.90.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*(9), 752–760. doi:10.1037/0003-066X.32.9.752.

Valentine, J. C. (2009). Judging the quality of primary research. In H. Cooper, L. V. Hedges, & J. C. Valentine (Hrsg.), *Handbook of research synthesis and meta-analysis* (2. Aufl., S. 129–146). New York: Russell Sage Foundation.

Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods, 6*(4), 413–429. doi:10.1037/1082-989X.6.4.413.