Z Erziehungswiss (2014) (Suppl) 17:79–99
DOI 10.1007/s11618-014-0510-9

ZfE

Aufbereitung von Evidenz für bildungspolitische und pädagogische Entscheidungen: Metaanalysen in der Bildungsforschung

Hans Anand Pant

Zusammenfassung: In diesem Beitrag wird aufgezeigt, welche Möglichkeiten und Grenzen Metaanalysen haben, um wissenschaftliche Evidenz für die Wirksamkeit pädagogischer und bildungspolitischer Maßnahmen zu gewinnen. Es wird unter Bezug auf Berliners Arbeiten (2002) argumentiert, dass in der Bildungsforschung, anders als in Teilen der bio-medizinischen Forschung, die Kontextbedingungen von Lehr-Lernsituationen nur sehr eingeschränkt kontrollierbar sind. Daher eigenen sich Metaanalysen im Bildungsbereich eher für die Aufdeckung und Beschreibung von Kontextbedingungen, die die Wirkungen von Maßnahmen und Programmen beeinflussen, und weniger für einen strengen Wirksamkeitsnachweis selbst. Vor diesem Hintergrund wird die bisher umfangreichste Forschungsbefundsynthese im Bildungsbereich, John Hatties Meta-Metaanalyse *Visible Learning* (2009), kritisch auf ihre Validität geprüft. An Hatties Beispiel wird gezeigt, dass eine bessere Verständigung über die Standards von Metaanalysen und eine Kommunikation des Geltungsanspruchs ihrer Befunde für die pädagogische und bildungspolitische Praxis notwendig erscheint.

Schlüsselwörter: Metaanalyse · Evidenz · Bildungspolitik · Schulleistungen · Steuerungswissen · Unterrichtsqualität

Abstract: This article discusses the promise and pitfalls of meta-analysis as a basis for evidence-based policy and practice in education. Referring to David Berliner's work (2002) it argues that educational research as opposed to research in the bio-medical field is much more restricted by the presence of powerful context effects and interactions, making rigorous experiments the exception as a basis for meta-analytical evidence. Therefore, an effectiveness-oriented use of meta-analysis of successful approaches of teaching and schooling is seen as more appropriate than an efficacy-oriented one. Finally, John Hattie's (2009) mega-analysis *Visible Learning* is discussed and evaluated against these caveats for the use of meta-analytical evidence in informing educational policy and practice.

Der Beitrag ist die modifizierten Fassung eines Textes, der in dem Tagungsband Bundesministerium für Bildung und Forschung (Hrsg.). (im Druck). Evidenzbasierte Bildungspolitik – Voraussetzungen und Hindernisse der Nutzung wissenschaftlichen Wissen. Berlin: BMBF, erscheint.

H. A. Pant (✉)
Institut zur Qualitätsentwicklung im Bildungswesen (IQB), Humboldt-Universität, Berlin, Deutschland
E-Mail: hansanand.pant@hu-berlin.de

Keywords: Meta-analysis · Evidence-based education · Educational policy · Academic achievement · Evidence-based governance · Quality of teaching

Im Zuge der Neuen Steuerung im Bildungswesen kommt der Generierung von empirischer Evidenz auf den verschiedenen Ebenen des Bildungssystems eine immer größer werdende Bedeutung zu (Campbell und Levin 2009; Coe 2009; Schildkamp et al. 2012; Slavin 2008; Wiseman 2010). Im Kern geht es darum, ob und wie es gelingt, Evidenz hinsichtlich der Wirksamkeit pädagogischer und bildungspolitischer Maßnahmen zur Verfügung zu stellen, um diese in datengestützte Entwicklungskreisläufe der Unterrichts-, Schul- und Bildungsqualität einspeisen zu können. Jede evidenzbasierte Entscheidung für eine bestimmte Maßnahme muss ihrerseits auf ihre Effekte hin überprüft werden und das Evaluationsergebnis in den Entwicklungskreislauf einfließen. Soll eine Lehrkraft ihren Unterricht stärker auf individuelle Förderung umstellen, um die Leistungsschwächeren an das Niveau der Lerngruppe heranzuführen? Welche erprobten Ansätze und Programme der Förderung stehen zur Verfügung? Trägt Schulinspektion dazu bei, dass *Failing Schools* (d. h. Schulen mit wiederholt weit unterdurchschnittlichen Leistungen) wieder den Anschluss finden? Welches Sprachförderkonzept soll eine Landesregierung flächendeckend implementieren, um soziale und zuwanderungsbezogene Disparitäten schulischer Leistungen zu reduzieren?

1 Evidenzbasierung: Grundlagen- oder Anwendungsforschung?

Um solche konkreten Fragen beantworten zu können, benötigt man Wissen darüber, *was wirkt* oder funktioniert (*„What works?“*, Ruthven 2011; Slavin 2008). Die *What-works*-Frage weist dabei jedoch unterschiedliche Aspekte auf. Unter einem eher *grundlagenforschungsorientierten* Aspekt sollen die *kausal zurechenbaren Effekte* von pädagogischen Maßnahmen oder bildungspolitischen Programmen auf definierte Zielgrößen, wie beispielsweise die schulische Leistungsentwicklung, identifiziert werden. Maßgeblich hierfür ist ein expliziter Theoriebezug („Wie ist der kausale Wirkmechanismus einer Intervention oder eines Programms theoretisch begründet?“) und der Rekurs auf empirische Studien, die möglichst rigorosen methodischen Standards für intern und extern valide Wirksamkeitsnachweise genügen.[1] Für den Bereich der pädagogischen Psychologie geben Winne und Nesbit (2010) einen umfassenden Überblick darüber, welche belastbaren *psychologisch-theoriegeleiteten* Befunde zu den kognitiven, meta-kognitiven und motivationalen Wirkfaktoren auf Schulleistungen vorliegen. Unter *anwendungsorientierter* Perspektive geht es hingegen stärker um die Frage, *ob* sich neue Maßnahmen oder Programme auch in komplexen pädagogischen und bildungspolitischen Handlungsfeldern mit ihren je spezifischen Konstellationen von Kontextfaktoren als wirksam erweisen und ob sich die Implementation einer Innovation vor dem Hintergrund der bestehenden Praxis „lohnt“. Es geht also um die Gewichtung und Nuancierung der drei „klassischen E's“ der Evaluationsforschung: *Efficacy (Wirksamkeit), Effectiveness (Wirkung), Efficiency (Effizienz)* (Wortman 1983). Während der Nachweis der *Wirksamkeit* einer Maßnahme unter hochgradig kontrollierten und standardisierten Bedingungen erfolgt, zeigen sich die Wirkun-

gen erst unter typischen Realbedingungen (z. B. in Unterrichtssituationen; Wortman 1983, S. 230). *Effizienz* schließlich bezieht als Evaluationskriterium die Kosten-Nutzen-Bilanz einer Intervention oder eines Programms bei gegebener *Wirkung* ein.

Für den Bereich medizinischer Maßnahmen und Therapien hat sich eine Hierarchie der Evidenzquellen, quasi ein Stufenmodell des Wirksamkeitsnachweises, etabliert. Ganz oben in dieser Hierarchie stehen Studien mit *randomisiertem Kontrollgruppendesign* und, sozusagen als Krönung, die systematische Befundintegration solcher und anderer methodisch hochwertiger Studien in Form von *Metaanalysen*. Evidenzbasierung im Sinne grundlagenforschungsorientierter Qualitätskriterien ist also primär an Befunden zur *Wirksamkeit* einer Intervention interessiert; soll Evidenz dagegen Aspekte des Anwendungsnutzens mitberücksichtigen, dann sind Informationen zu den *Wirkungen* und u. U. zur *Effizienz* mindestens gleichrangige Ziele einer wissenschaftlichen Forschungsbefundintegration.

Während im medizinischen Bereich die Orientierung an einem Evidenzverständnis, das auf Wirksamkeitsnachweise abzielt, noch als plausibel gelten kann[2], wird dies für den pädagogischen Bereich teilweise vehement in Zweifel gezogen (z. B. Bellmann und Müller 2011). Berliner (2002) benennt drei grundsätzliche Charakteristika des Bildungsbereichs, die eine Übernahme des medizinischen Konzepts der Evidenzbasierung erschweren: die Wirkung multipler Kontexteinflüsse in realen Lehr-Lernsituationen *(Power of Contexts)*, die Allgegenwart von Wechselwirkungen *(Ubiquity of Interactions)*und die geringe „Halbwertszeit" der Befunde empirischer Bildungsforschung *(Decade x Findings Interactions)*. Unter den ersten beiden Charakteristika versteht Berliner das nach seiner Ansicht unentwirrbare und experimentell unkontrollierbare Mehrebenengefüge von Wirkfaktoren und Interaktionen in institutionellen Lehr-Lern-Kontexten, wie z. B. dem Klassenzimmer. Auf welche *Konstellation* von sozialer und ethnischer Schülerzusammensetzung, Lehrerkompetenz, Schulleitungshandeln, materieller Ausstattung der Schule, administrativem Unterstützungssystem etc. trifft die Implementation einer konkreten Interventionsmaßnahme? Was genau wirkt dann noch bzw. verhindert das Wirksamwerden einer Maßnahme? Mit geringer „Halbwertszeit" der Befunde in der Bildungsforschung meint Berliner (2002), dass selbst solide gewonnene empirische Evidenz aus der Bildungsforschung mitunter allein schon aufgrund des raschen sozialen Wandels der Kontextbedingungen von Bildungserwerb obsolet werden kann. So hat möglicherweise empirische Evidenz zu den Effekten von integrativer vs. nicht-integrativer Beschulung von Kindern mit sonderpädagogischem Förderbedarf auf deren Kompetenzerwerb in zehn Jahren kaum noch Gültigkeit, da sich das Ausmaß und die Konzepte integrativer Unterrichtsformen in naher Zukunft deutlich verändern.

Die Komplexität des Kontexts und der kontextbezogenen Interaktionen machen es nach Berliner nahezu unmöglich, den Effekt einzelner interessierender Faktoren, z. B. den eines bestimmten Sprachförderprogramms, zu isolieren und soweit generalisierbar zu machen, dass Effekte für eine konkrete Schule bzw. eine angebbare Konstellation von konkreten Kontextmerkmalen abschätzbar sind.

Nach Berliners (2002) Argumentation ist sowohl die interne als auch die externe Validität von randomisierten Kontrollgruppenansätzen angesichts dieser Barrieren bedroht; *Evidenz-basierte Bildungspolitik* nach dem Muster und den Standards der Evidenz-ba-

sierten Medizin wäre somit ein kaum einlösbares Versprechen (vgl. auch Chatterji 2007; Davies et al. 2008; Ong-Dean et al. 2011; Song und Herman 2010).

In der aktuellen Diskussion um die Qualitätsstandards von sozial- und verhaltenswissenschaftlicher Forschung im deutschsprachigen Raum wird versucht, die Dichotomie und den anscheinenden Widerspruch zwischen *Erkenntnis- und Nutzeninteresse* bei der Generierung von Evidenz abzuschwächen. So schlagen Brüggemann und Bromme (2006) als Ergebnis einer Diskussion der Begutachtungskriterien von DFG-Forschungsanträgen vor, zwischen den Polen der „reinen" Grundlagenforschung und der Anwendungsforschung eine „Anwendungsorientierte Grundlagenforschung" zu platzieren. Die Anwendungsorientierte Grundlagenforschung bleibt nach Brüggemann und Bromme (2006) zwar den grundlagenforschungsorientierten Kriterien der Generalisierbarkeit und einer stringenten Theorieanbindung verpflichtet, intendiert aber *zugleich* „dezidiert praktische Effekte bzw. praktischen Nutzen und zieht sie auch zur Begründung und Beschreibung ihrer Fragestellungen heran" (S. 113). Die Autoren konzedieren, dass im Rahmen Anwendungsorientierter Grundlagenforschung einer (experimentellen) Kontrollierbarkeit Grenzen gesetzt sind und dass die experimentelle Isolation einzelner Variablen sogar ein unangemessenes Zielkriterium zur Erforschung komplexer Zusammenhänge darstellen könne. Weiterhin plädieren Brüggemann und Bromme (2006) dafür, dass in der Anwendungsorientierten Grundlagenforschung auch das zweite Qualitätskriterium der „reinen" Grundlagenforschung, die „stilreine" Bezugnahme auf möglichst nur eine Theorie, gelockert werden müsse. Bei anwendungsorientierten Fragestellungen dürften auch gleichzeitig Theorieelemente und entsprechende Konstrukte aus verschiedenen „(…) Diskursen herangezogen werden, um dem jeweiligen Problem gerecht zu werden" (S. 115).

Damit zeichnet sich im deutschsprachigen Raum eine etwas andere Stoßrichtung in der Frage der Evidenzbasierung in pädagogischen und bildungspolitischen Handlungsfeldern ab als in den USA (Fischer et al. 2005). Dort dominiert nach den Empfehlungen des U.S. Department of Education im Zuge der *No Child Left Behind*-Gesetzgebung (U.S. Department of Education 2002) eine klare Orientierung am strengen Wirksamkeitsmodell von Evidenz, verbunden mit der Präferenz für randomisierte Kontrollgruppenstudien (U.S. Department of Education 2005) und entsprechend darauf aufbauenden systematischen Reviewverfahren. So wurden im Auftrag des US Department of Education zahlreiche Metaanalysen durchgeführt, um die Effekte der Sonderbudgetierung von Schulen mit einer sozial benachteiligten Schülerschaft auf die Leistungsentwicklung zu überprüfen (Zimmer et al. 2007). Auf Initiative des US Department of Education wurde zudem 2002 das *What Works Clearinghouse* (WWC) eingerichtet. Das WWC ist eine zentrale Datenbasis, die Politik, Wissenschaft und vor allem pädagogischer Praxis über Metaanalysen und andere systematische Reviewverfahren abgesicherte Informationen bereitstellen soll, welche Maßnahmen und Programme im Bildungsbereich effektiv sind (What Works Clearinghouse 2011).

Welche Implikationen hat nun die aufgezeigte Diskussion für die Nutzung und Bewertung von *Metaanalysen* als methodischem Instrument der Evidenzerzeugung im Bildungsbereich? Im folgenden Abschnitt wird zunächst der Ansatz der Metaanalyse in seinen Grundzügen skizziert. Dabei wird auf diejenigen Aspekte im Prozess einer Metaanalyse fokussiert, die auf der statistisch-methodischen Ebene eine Präferenz für *Wirksamkeits*- vs. *Wirkungs*-Kriterien von Evidenz reflektieren. Vor diesem Hintergrund

wird anschließend die gegenwärtig wohl einflussreichste Forschungsbefundsynthese im Bildungsbereich, *Visible Learning* von John Hattie (2009; 2012), kurz vorgestellt und kritisch diskutiert. Der Schlussabschnitt bilanziert Nützlichkeit und Begrenztheit metaanalytischer Ansätze für die Evidenzgenerierung in pädagogischen und bildungspolitischen Handlungsfeldern.

2 Metaanalytisches Vorgehen: Spielraum für Entscheidungen

Beelmann und Bliesener (1994, S. 211) definieren den metanalytischen Ansatz wie folgt: „Die Metaanalyse kann als Sammlung konzeptioneller und methodischer Verfahren verstanden werden, mit deren Hilfe empirische Daten zu einer festgelegten Fragestellung in einer quantitativ orientierten Weise zusammengefasst werden".

Metaanalysen sind ein Spezialfall *systematischer* Übersichtsarbeiten. Diese sind dadurch gekennzeichnet, dass sie im Unterschied zu sogenannten *narrativen* Übersichtsarbeiten versuchen, sämtliche für einen interessierenden Effekt existierenden Untersuchungen (Primärstudien) zu berücksichtigen und relevante Merkmale und Befunde der Studien (z. B. theoretischer Ansatz, Untersuchungspopulation, Interventionstyp, Zielkriterien, Studiendesign) systematisch zu erfassen. Darüber hinaus werden die Suchstrategie in Literaturdatenbanken und anderen Quellen sowie die Ausschlusskriterien für Primärstudien protokolliert, so dass die Datenbasis eines systematischen Reviews für Dritte überprüfbar wird.

Als zentrales differenzierendes Merkmal von Metaanalysen gegenüber anderen Formen systematischer Literaturübersichten gilt die quantifizierende Integration bzw. Aggregation von Primärstudienergebnissen (Cooper et al. 2009). Ziel der Aggregation ist es, am Ende eine Maßzahl zu haben, die eine hohe Verdichtung der vorgefundenen empirischen Evidenz zur Wirksamkeit bzw. Wirkung von Interventionen, Maßnahmen oder Programmen darstellt. Dazu müssen zunächst die Befunde aller Primärstudien (z. B. zur Wirksamkeit von Sprachförderprogrammen auf Lesekompetenz), die in der Regel mit unterschiedlichen Messinstrumenten gewonnen wurden, auf ein einheitliches quantitatives Maß gebracht werden: die *Effektgröße* (auch: Effektstärke, abgek. *ES*). Als Effektgrößen bezeichnet man definierte Maße des Zusammenhangs zwischen zwei interessierenden Merkmalen bzw. Maße des Effektes einer unabhängigen auf eine abhängige Variable. Der *d*-Index als bekanntestes Effektgrößenmaß (Cohen 1988) ermittelt die Mittelwertsdifferenz zwischen zwei betrachteten Gruppen (z. B. Interventions- und Kontrollgruppe), relativiert an der gepoolten Standardabweichung beider Gruppen oder der Streuung der Kontrollgruppe. Diese Maße drücken hinsichtlich der abhängigen Variablen (z. B. Lesekompetenz) den *Verteilungsabstand* von Interventions- und Kontrollgruppe (z. B. Teilnehmer *vs.* Nichtteilnehmer an einer Sprachfördermaßnahme) in Einheiten der Standardabweichung aus. Die meisten Effektgrößenmaße, wie die *d*-Indizes, *r*-Korrelationsmaße, bivariaten Regressionskoeffizienten *b*, lassen sich leicht ineinander überführen, so dass sich z. B. Befunde aus Studien mit Kontrollgruppendesigns (*d*-Maße), Korrelationsstudien (*r*-Maße) oder einfachen Regressionsdesigns (*b*-Maße) in einer Metaanalyse zu einer gemittelten Effektgröße verrechnen lassen.

2.1 Evidenz für Efficacy oder Effectiveness: Weichenstellungen im metaanalytischen Prozess

Wenn Beelmann und Bliesener (1994, S. 211) von Metaanalysen als einer „Sammlung konzeptioneller und methodischer Verfahren" sprechen, so ist damit angedeutet, dass von der Forscherin bzw. dem Forscher an etlichen Stellen des metaanalytischen Prozesses Entscheidungen gefordert werden.

Im Folgenden sollen einige dieser Entscheidungen als Weichenstellungen beschrieben werden. Je nach Entscheidungsrichtung orientiert sich eine Metaanalyse eher am Ziel eines *Wirksamkeits*nachweises, d. h. an Standards der Grundlagenforschung mit dem Ziel *inferenzstatistischer Hypothesen- und Theorietestung*. Anders verlaufende Entscheidungen kennzeichnen dagegen einen primär *explorativen* Anspruch an das Verfahren und seine Befunde. Dieser Anspruch kann als *Wirkungs*-orientiert bezeichnet werden, da – in Berliners (2002) Sinn – versucht wird, die moderierenden Einflüsse von Kontextbedingungen in komplexen Handlungsfeldern nicht zu kontrollieren, sondern sie zu modellieren.

Aus Platzgründen können hier nicht alle Entscheidungspunkte im Metaanalyseprozess beschrieben werden (vgl. hierzu ausführlich Cooper et al. 2009). Daher konzentriert sich die Darstellung auf Weichenstellungen an zwei Stellen: a) die *Such- und Auswahlstrategie* der Primärstudien (d. h. der ursprünglichen Untersuchungen) und b) die Frage, wie die Effektgrößen der Primärstudien statistisch *aggregiert* werden. Die methodischen Ausführungen werden anhand von Beispielen publizierter Metaanalysen aus dem Bereich der Schul- und Unterrichtsforschung illustriert.

2.2 Entscheidungen bei der Such- und Auswahlstrategie von Primärstudien

Durch die Such- und Auswahlstrategie wird gleich zu Beginn des metaanalytischen Prozesses festgelegt, welcher Typus von Studien überhaupt zur Befundintegration „zugelassen" wird, insbesondere, wie homogen hinsichtlich inhaltsbezogener und methodischer Kriterien die Primärstudien sein sollen (Valentine 2009). Als die beiden hauptsächlichen Kontroversen gelten das Uniformitätsproblem *(„Äpfel-und-Birnen-Problem")* und das Problem der Artefaktkontrolle *(„Garbage-in-Garbage-out-Problem")*. Das *Äpfel-und-Birnen-Problem* bezieht sich auf die Frage, ob es überhaupt sinnvoll ist, heterogene Konstrukte als „gleichwertig" zu betrachten und deren Effekte quantitativ in einer aggregierten Effektgröße auszudrücken. So kritisierten beispielsweise McGee und Lomax (1990) die Metaanalyse von Stahl und Miller (1989). Diese hatten verschiedene Instruktionsmethoden des Erstleseunterrichts miteinander verglichen. Dabei integrierten sie die Ergebnisse von 117 Vergleichen, die jeweils den Effekt einer traditionellen analytischen Leselernmethode (Laut- und Buchstaben-bzw. Fibellernen) dem einer ganzheitlichen oder auf Spracherfahrungsansätzen beruhenden, synthetischen Methode gegenüberstellten. Bei letztgenannter Methode sollen Kinder das Lesen eigenaktiv anhand selbst gesprochener und (auch orthografisch falsch) hingeschriebener ganzer Wörter und Sätze erlernen. Eine gemittelte Effektgröße von $d = 0{,}09$ ließ die Autoren resümieren, „[…] [W]hole language/language experience approaches were not reliably different from basal reader approaches in their effects" (Stahl und Miller 1989, S. 94).

Die „*Äpfel-und-Birnen*"-Kritik von McGee und Lomax (1990) lautet im Kern, dass mit dem Etikett „whole language/language experience approaches" sowohl in historischer als auch theoretischer Sicht vollkommen unterschiedliche Konzepte als homogener Ansatz betrachtet wurden; dies hätte die mögliche Überlegenheit neuerer synthetischer Erstleselernmethoden maskiert. Zudem seien wichtige potenzielle Einflussgrößen, wie die Dauer der Lesezeit der Kinder, nicht beachtet worden, die bei den analytischen Methoden höher sei. Kritisiert wird also der *Verlust an Spezifität*, der dadurch entstehe, dass unterschiedlichste Interventionsansätze unter einem gemeinsamen Label (hier: Ganzheitliche Erstlesemethode) aggregiert würden, was die theoretische Relevanz der Analyse in Frage stelle (Eysenck 1995).

Wie oben ausgeführt, kennzeichnet „Strenge" hinsichtlich der theoretischen Bezugnahme und der Konstruktdefinition einen eher der Grundlagenforschung nahestehenden *Wirksamkeits*-Ansatz metanalytischer Evidenz. Wie etliche Autoren betonen, beschneidet jedoch eine rigorose Auswahlstrategie die Möglichkeit, Moderatoren der Wirksamkeit einer Maßnahme, wie zum Beispiel die theoretische Ausrichtung des Interventionsansatzes oder etwaige Umsetzungsvarianten bei der Implementation, zu identifizieren und hinsichtlich ihres differenziellen Effektes zu quantifizieren (zusammenfassend Cortina 2003).

Technisch betrachtet erfordern Moderatorenanalysen zunächst die Identifikation bestehender Unterschiede in den Effektgrößen der Primärstudien (Zwischenstudienvarianz). Ist diese Varianz substanziell, so kann der moderierende Effekt relevanter Randbedingungen entweder durch Bildung homogener Gruppen von Studien oder regressionsanalytisch bestimmt werden (Cortina und Pant 2009). Auf diese Weise können in Metaanalysen im Sinne der Cook-Campbell'schen Validitätskriterien (Shadish et al. 2002) sowohl wichtige Bausteine des Prozessverständnisses eines Interventionseffektes (interne Validität) als auch dessen Abhängigkeit von Operationalisierungsmerkmalen (Inhaltsvalidität) sowie den situativen und raum-zeitlichen Gegebenheiten einer Untersuchung (externe Validität) konkretisiert werden.

Analog zum Äpfel-und-Birnen-Problem wird darüber gestritten, ob man methodisch „schlechte" Studien von vornherein aus einer Metaanalyse ausschließen solle oder nicht, da diese die Qualität der Metaanalyse selbst kompromittieren könnten (Eysenck 1984; Oswald und Plonsky 2010). Um Ausschlusskriterien zu erhalten, werden z. B. Qualitätsscores pro Studie kodiert, die validitätseinschränkende Studienmerkmale erfassen sollen (Matt und Cook 2009). Ab einem willkürlichen Schwellenwert werden Studien mangels hinreichender methodischer Qualität aus der Metaanalyse ausgeschlossen (Schmidt et al. 2009a). Aus Sicht einer auf *Wirkungs*-Evidenz ausgerichteten Verwendung von Metanalysen ist, ähnlich wie beim Äpfel-und-Birnen-Problem, jedoch eher eine systematische Kodierung von validitätsrelevanten methodischen Studienmerkmalen vorzunehmen und deren moderierende Wirkung statistisch zu überprüfen.

So rücken beispielsweise Seidel und Shavelson (2007) Moderatorenanalysen ins Zentrum ihrer Metaanalysen zu den Effekten unterschiedlicher Faktoren des schulischen Lernens (wie z. B. Art der Klassenführung oder Kooperatives Lernen) auf den Lernerfolg von Schülerinnen und Schülern. Dabei möchten sie u. a. klären, ob sich methodische Faktoren, wie die Art des Studiendesigns (korrelativ vs. quasi-experimentell vs. experimentell), die Art der Operationalisierung der Einflussfaktoren (z. B. Messung von Klassenklima

über Lehrerfragebogen, Schülerfragebogen oder Videobeobachtung) und die Definition des Lernerfolgs (Status- vs. Veränderungsmessung), auf die feststellbaren Effektgrößen systematisch auswirken. Seidel und Shavelson (2007) finden erhebliche Unterschiede in den mittleren Effektgrößen in Abhängigkeit von derartigen methodischen Faktoren und raten dazu, die Sensitivität der Befunde von Primärstudien und von Metaanalysen, die dies nicht explizit betrachten, in der *Scientific Community* zu diskutieren.

2.3 Entscheidungen bei der Aggregation der Primärstudienergebnisse

Historisch hat sich die Metaanalyse als Befundintegration *bivariater* Zusammenhänge entwickelt (Beelmann und Bliesener 1994). Mit fortschreitendem Entwicklungsstand einer wissenschaftlichen Teildisziplin, wie z. B. der empirischen Schulleistungsforschung, werden aber bereits auf Primärstudienebene multivariate, prozesshafte, durch Wechselwirkungs-, Mediatoren- oder Mehrebeneneffekte gekennzeichnete Wirkmodelle entwickelt und geprüft. Deren differenzierende Aussagen lassen sich jedoch prinzipiell nicht in einer einzigen Effektstärke verdichten. Es entsteht das statistische Problem, dass bei allen Verfahren, die Partialkoeffizienten als Ergebnisstatistik verwenden (z. B. multiple Regressionsanalysen, Pfadmodelle, Faktorenanalysen), nur bei (möglichst exakter) Replikation des Variablensets eine Datenaggregierung sinnvoll ist. Anderenfalls würden Kennwerte zusammengefasst, die je nach Partialisierungsvariablen verschiedene Bedeutungen aufwiesen und damit inhaltlich unterschiedliche Zielgrößen schätzten (Becker 2009). Viele Primärstudien untersuchen zudem die Effekte einer pädagogischen Intervention nicht nur auf eine, sondern auf mehrere Zielgrößen (z. B. Wirksamkeit einer Maßnahme der Sprachförderung auf Verbesserung von Lesefähigkeit, Wortschatz und Orthografie), die dann zunächst multivariat zu einer Effektgröße verarbeitet werden müssen.

Metaanalysen sehen sich daher häufig gezwungen, hinter den empirischen Differenzierungsgrad methodisch ausgefeilter Primärstudien zurückzugehen, um eine Ergebnisaggregation durchführen zu können. So können im ungünstigen Fall die „besten“ Veröffentlichungen z. B. zur Wirkung von zwei- vs. mehrgliedrigen Schulsystemen auf die Leistungsentwicklung nicht in eine Metaanalyse einbezogen werden, weil jene nur adjustierte (d. h. um den Einfluss bestimmter Kontextmerkmale, wie z. B. den sozialen Hintergrund der Schülerschaft bereinigte) Zusammenhänge berichten.

Ein Beispiel dafür, wie diese statistisch „erzwungene“ Simplifizierung in Metaanalysen zu problematischen, weil verkürzten inhaltlichen Schlussfolgerungen führen kann, sind die Studien zur Wirkung von Hausaufgaben auf Schulleistung. Cooper et al. (2006) fanden in ihrer sehr aufwändigen Metaanalyse den über viele Studien konsistenten Befund, dass die Menge der Hausaufgaben und die Dauer der Hausaufgabenbeschäftigung signifikant positiv mit Leistungsindikatoren assoziiert waren, wenngleich auch schwächer im Primar- als im Sekundarbereich. Trautwein et al. (2009) greifen die Ergebnisse dieser Metaanalyse auf. Sie zeigen in einer längsschnittlichen und spezifisch auf Hausaufgabeneffekte angelegten Mehrebenenanalyse, dass die Effekte von Hausaufgaben je nach Analyseebene unterschiedlich ausfallen bzw. kausal verschieden interpretiert werden müssen. Auf der *Klassenebene* verschwand beispielsweise der signifikant positive Effekt von Hausaufgabenmenge auf Leistung, wenn als Moderatorvariable die

Schulart berücksichtigt wird. Schülerinnen und Schüler im untersten von zwei (bzw. drei) Bildungsgängen (Schweizer Schulsystem) erhielten systematisch weniger Hausaufgaben, so dass sich der Zusammenhang zur Leistung eher auf das Merkmal Schulart als auf die Hausaufgabenmenge an sich zurückführen lässt. Auf der *Schülerebene* ergab sich ein umgekehrter Zusammenhang zwischen Hausaufgabenzeit und nachfolgender Leistung, wenn das zuvor bestehende Leistungsniveau in die Analyse einbezogen wurde. Schwächere Schüler benötigten mehr Zeit für ihre Hausaufgaben und erzielten dennoch geringere Zugewinne. Grundsätzlich ist es möglich, Mehrebenenstrukturen im Rahmen von Metaanalysen zu berücksichtigen und statistisch zu modellieren. Schmid et al. (2013) geben einen Überblick über entsprechende Ansätze und Softwareoptionen.

Diese und andere Differenzierungen unterstreichen die Notwendigkeit, Befunde aus Metaanalysen mit denen aus methodisch anspruchsvollen Einzelstudien abzugleichen, um kausale Fehlschlüsse zu vermeiden.

Prinzipiell ist zwar auch die Aggregation komplexerer Variablenzusammenhänge – etwa von Kovarianzstrukturen mehrerer Strukturgleichungsmodelle –metaanalytisch zu bewerkstelligen. Dies hat jedoch fast immer den Nachteil, dass automatisch die Menge der auffindbaren Primärstudien drastisch reduziert wird, da nur Studien mit (zumindest fast) exakt identischen Variablensets aggregiert werden können. Damit aber fiele einer der Hauptvorteile des quantifizierenden Ansatzes der Metaanalyse, die Aggregierung *großer Mengen* von Einzelergebnissen, weg. Ein möglicher Ausweg besteht im Nachschalten komplexer Analyseverfahren. Zunächst werden bivariate Primärbefunde metanalytisch aggregiert und anschließend die aggregierten Statistiken (z. B. gemittelte Korrelationen) in multivariaten Modellen (z. B. Pfadmodellen) auf Studienebene weiter analysiert (für ein Anwendungsbeispiel aus dem Bildungsbereich s. Robbins et al. 2009). Spätestens bei einem solchen Vorgehen ließe sich jedoch diskutieren, ob statistische Analysen auf der Studienebene inhaltlich das Gleiche aussagen wie die Daten auf der Primärstudienebene. Sohn (1995) formuliert dieses Unbehagen wie folgt: „Meta-analysts have acted as if there is no difference, as if conclusions based on the study of the literature have the same epistemological standing as those based on the direct study of nature. Are meta-analysts correct in this regard? Is it proper to treat the research literature as a proxy for nature?“ (S. 109).

Ein weiterer Aspekt betrifft die Wahl des *statistischen Modells* der Aggregation von Effektgrößen aus Primärstudien, d. h. die Berechnung der mittleren Effektstärke entweder nach einem Modell mit festen Effekten *(Fixed Effects Model)* oder nach einem Modell mit zufälligen Effekten *(Random Effects Model)*. Bei dieser Unterscheidung geht es allgemein formuliert um die Frage, in welchem Verhältnis die Stichprobe von Studien zur Grundgesamtheit aller Primärstudien steht, die den interessierenden Effekt untersucht haben. Inhaltlich bedeutet die Verwendung eines *Fixed-Effects*-Modells, dass man davon ausgeht, bei den betrachteten Einzelstudien handelt es sich quasi um Replikationsuntersuchungen aus derselben Population. Dies mag für den Bereich medizinisch-experimenteller Forschung mit randomisierten Kontrollgruppendesigns noch plausibel erscheinen; für naturalistische Studien im Bildungsbereich dagegen kaum, da hier eine Vielzahl von z. T. unkontrollierbaren Kontext- und Moderatoreneinflüssen die Regel darstellt.

Das *Random-Effects*-Modell hebt die restriktiven Annahmen des *Fixed-Effects*-Modells insofern auf, als es nicht mehr davon ausgeht, dass alle Studien denselben Popu-

lationsparameter („wahren Effekt“) schätzen. Vielmehr wird die Studienstichprobe einer Metaanalyse buchstäblich als das Resultat einer Stichprobenziehung aus einer Grundgesamtheit von Studienrealisationen zu dem beobachteten Zusammenhang bzw. dem Interventionseffekt betrachtet. In diesem Studien-Universum existiert nicht notwendig nur eine einzige „wahre“ Effektstärke, sondern eine Verteilung von mehreren „wahren“ Effektstärken. Die Annahme einer Verteilung von „wahren“ Effektstärken reflektiert die Auffassung, dass mögliche Einflussgrößen (Moderatoren) beim Zustandekommen einer Effektstärke einerseits zu zahlreich und andererseits aus dem Datenmaterial (Veröffentlichungen) gar nicht rekonstruierbar sind, um sie explizit kontrollieren zu können. Die Bevorzugung eines *Fixed*- bzw. *Random*-Ansatzes im Rahmen von Metaanalysen ist damit sowohl eine Frage theoretischer Überlegungen, ob man metaanalytische Befunde im Sinne des oben dargestellten Verständnisses als *Wirksamkeits*- oder *Wirkungs*-Nachweis begreift, als auch eine empirisch vor dem Schritt der Datenaggregation zu klärende Angelegenheit. Denn im Falle empirisch feststellbarer substanzieller Heterogenität in den Effektstärken wird es inhaltlich immer unwahrscheinlicher, dass diese allein durch den Umstand verschiedener Personenstichproben (d. h. den Stichprobenfehler) und damit im Rahmen eines *Fixed-Effects*-Modells erklärbar sind. Methodisch differenzierte Metaanalysen kontrastieren häufig die Ergebnisse beider Aggregationsverfahren (z. B. Cooper et al. 2009; Tamim et al. 2011).

Als Zwischenfazit der Diskussion von Problemen der Studienauswahl und der Ergebnisaggregation zeichnen sich drei Empfehlungen für die Nutzung von Metaanalysen als Instrument der Evidenzgenerierung in der Schul- und Bildungsforschung ab.

1. Für das „Äpfel-und-Birnen-Problem“ gilt: Unter den von Berliner (2002) benannten Komplexitätsstrukturen von Bildungsprozessen sind Metaanalysen in erster Linie dann ein nützliches Instrument, wenn sie im Sinne eines *Wirkungs*-orientierten Verständnisses (vgl. Abschn. 1) zur Klärung der Bedingungen erfolgreicher Interventionen und Programme in realen Kontexten beitragen. Wie Seidel und Shavelson (2007, S. 485) formulieren: „Instead of estimating and searching for the true effect of teaching on learning, the role of meta-analysis primarily would be to capture context and outcome variation in reporting nuanced findings of teaching effectiveness.“ Die Sichtweise, Metaanalysen im Bildungsbereich würden definitive Wirksamkeitsnachweise von Interventionen und Programmen liefern, die sich beliebig generalisieren ließen, ist daher durch eine explizite und transparente Kommunikation gegenüber Bildungspolitik und -praxis zu relativieren.
2. Hinsichtlich des Problems der Artefaktkontrolle ist festzuhalten: Das Potenzial von Metaanalysen, den Effekt von methodischen Merkmalen der Primärstudien auf deren Ergebnisse zu identifizieren, sollte in erster Linie innerhalb der *Scientific Community* genutzt werden, um zukünftige Wirkungsstudien so zu gestalten, dass vorhandene Effekte sichtbar werden können.
3. Metanalysen greifen in der Regel auf unterkomplexe, bivariate Ergebnisdarstellungen in den Primärstudien zurück. Die metaanalytischen Befunde sollten daher mit den Ergebnissen besonders valider Einzelstudien, die aufgrund ihrer komplexen Analyseverfahren nicht in die Metanalyse eingehen konnten, abgeglichen werden (vgl. auch *Best Evidence Synthesis*, Slavin 2008).

Tab. 1: Ausgewählte Phasen und Hauptprobleme des metaanalytischen Prozesses

Phase	Entscheidungsproblem	Strategien zur Überprüfung der Sensitivität der Ergebnisse
Suche und Auswahl der Primärstudien	Konzeptionelle Heterogenität untersuchter Konstrukte in den Primärstudien (*Uniformitätsproblem* oder *„Äpfel-und-Birnen-Problem“*)	Theoriegeleitete Bildung von Studien-Subgruppen mit anschließenden Homogenitätsanalysen bzw. Regressionsanalysen
	Methodische Qualität der Primärstudien heterogen *(„Garbage-in-Garbage-out-Problem“)*	Konstruktion von Validitätsindikatoren mit anschließender Subgruppenanalyse bzw. Regressionsanalyse
Kodierung	Fehlende Daten zur Berechnung der Effektstärke	Kontrastierende Analyse unter Verwendung von Studiensets nur mit vollständigen Daten bzw. nach Ersetzung fehlender Datenpunkte durch Imputationsverfahren
Effektstärkenaggregation	Behandlung abhängiger ES	Multivariate Verrechnung
	Reliabilität der ES	Vergleich der Aggregationsergebnisse mit und ohne Gewichtung der Studien anhand des Stichprobenumfangs
	Heterogenität der ES zwischen den Studien und statistisch-theoretische Modellannahmen (*Fixed vs. Random Effects Model*)	Testung der Heterogenität der ES und Bestimmung der Varianzkomponenten ‚*innerhalb*‘ und ‚*zwischen*‘ den Studien Vergleich der Aggregationsergebnisse unter *Fixed*- bzw. *Random*-Modellannahme
Darstellung und Interpretation der Ergebnisse	Mehrdeutige Interpretierbarkeit der Befundintegration	Transparente Darstellung des Einflusses der konzeptionellen Heterogenität und der methodischen Qualität der Primärstudien auf die Ergebnisse der Metaanalyse Vergleich mit den Befunden der „besten“ Primärstudien *(Best Evidence Synthesis)*

ES Effektstärke

In Tab. 1 werden ausgewählte Phasen des metaanalytischen Prozesses mit zentralen Entscheidungsproblemen sowie Strategien zur Überprüfung der Sensitivität der Befunde in der Übersicht dargestellt. Sie dienen im folgenden Abschnitt auch als Kriterienraster, um die einflussreichen Metaanalysen von John Hattie einzuordnen.

3 Die Meta-Metaanalysen von John Hattie

Während Evidenzbasierung und insbesondere Metaanalysen im bio-medizinischen Anwendungsbereich inzwischen als unverzichtbar gelten (Volmink et al. 2004) – sowohl gesundheitspolitisch für die Populationsebene also auch für individuelle ärztliche Entscheidungsprozesse – hatten sie diesen Status in pädagogischen und bildungspolitischen Handlungsfeldern bis vor kurzem auch nicht annähernd. Die Meta-Metanalysen von John Hattie (2009; 2012) zur Frage, welche Bedingungen und Merkmale auf Seiten von

Schülerinnen und Schülern, Lehrkräften, der Unterrichtsgestaltung sowie der schulischen Rahmenbedingungen mit schulischer Leistung in Zusammenhang stehen, haben diese Situation verändert. Sein Buch *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement* (Hattie 2009) wurde der wissenschaftlichen Suchmaschine *Google Scholar* zufolge nach Erscheinen fast zweitausend Mal zitiert; daneben wurden seine Hauptbefunde in zahlreichen bildungspolitischen Debatten aufgenommen und z. T. äußerst kontrovers diskutiert (Hattie 2010; Terhart 2011). Vor dem Hintergrund der hier interessierenden Frage, welchen Stellenwert quantitative Forschungsbefundsynthesen für die Evidenzerzeugung im Bildungsbereich haben, erscheint es daher sinnvoll, den Aspekt des *strategischen* Stellenwerts von Hatties Analysen vom *methodischen* zu trennen. Der strategische Nutzen erscheint auf den ersten Blick unbestreitbar: *Visible Learning* hat im internationalen Maßstab als „Weckruf" für eine Evidenzdebatte im Bildungsbereich funktioniert und damit eines von Hatties Hauptzielen erreicht, die Fragen nach Wirksamkeit, Wirkung und Effizienz von Bildungsmaßnahmen zu Schlüsselfragen zu machen: „[W]e spend millions, if not trillions, of dollars investing in innovations, changes, and policies in education without a lot of evidence that this investment is making a difference to student outcomes" (Hattie 2009, S. 255). Im Folgenden werden zunächst in sehr knapper Form Hatties (2009) Studienanlage und Kernbefunde beschrieben, um dann auf einige methodische Aspekte unter Bezugnahme auf die oben dargestellten Entscheidungen im metaanalytischen Prozess einzugehen.

3.1 Studienanlage und Hauptbefunde von Visible Learning

Hatties Forschungsbefundsynthese betrachtet die Effekte von Schüler-, Unterrichts- und Schulvariablen auf Schulleistungen. In seine Analyse gehen ausschließlich bereits durchgeführte Metaanalysen ein, d. h. in *Visible Learning* werden keine Primärstudien betrachtet (Hattie 2009, S. 255). In diese Meta-Metaanalyse (oder „Megaanalyse", Terhart 2011) von mehr als 800 publizierten Metaanalysen anderer Forscherinnen und Forscher fließen indirekt die Daten und Ergebnisse von über 52 000 Primärstudien mit insgesamt mehr als 200 Mio. untersuchten Schülerinnen und Schülern ein. Aus diesen Studien extrahiert Hattie 138 (potenzielle) Einflussfaktoren der Schulleistung, die er in sechs Faktorengruppen zusammenfasst: Merkmale der *Lernenden* (z. B. Besuch vorschulischer Angebote, Vorleistung, Interesse am Fach), Merkmale des *Elternhauses* (z. B. sozioökonomischer Status der Eltern, elterliches Interesse an Schule), Merkmale der *Schule* (z. B. durchschnittliche Klassengröße, jahrgangsgemischtes Lernen), Merkmale der *Lehrkraft* (z. B. Ausbildung, Fachwissen), Merkmale des *Curriculums* (z. B. bilinguale Klassen, schulische Zusatzangebote) und die größte Faktorengruppe, Merkmale des *Unterrichts* selbst (z. B. Feedback-Kultur, Förderung meta-kognitiver Strategien).

Tabelle 2 zeigt, dass die durchschnittlich stärksten Effekte im Bereich der Lehrermerkmale sowie bei Faktoren der Lehr-Lernsituation (*Teaching, Curricula*) zu verzeichnen sind und Merkmale der Schülerinnen und Schüler erst danach folgen.

Die schnelle und breite Rezeption von Hatties Befundsynthese ist sicherlich zum einen dem Respekt vor der schieren Masse an wissenschaftlich fundierter Information geschuldet, die in seiner Studie verdichtet wird; zum anderen dürften dazu aber auch folgende Faktoren beigetragen haben: 1) eine explizite theoretische Konzeption von „gutem",

Tab. 2: Aggregierte Effektstärken d für den Zusammenhang von sechs untersuchten Faktorengruppen mit Schulleistung (gekürzt und übersetzt aus Hattie 2009, S. 18)

Charakteristika	Anzahl der Meta-analysen	Anzahl der Primärstudien	Summe des Primärstudien-*N*	Anzahl der Effektstärken	Aggregierte Effektstärke *d*	Durchschnittlicher *SE*
Schüler/-innen	139	11.101	7.513.406	38.287	0,40	0,044
Familiärer Hintergrund	36	2211	11.672.658	5182	0,31	0,058
Schule	101	4150	4.416.898	13.348	0,23	0,072
Lehrkraft	31	2225	402.325	5559	0,49	0,049
Curriculum	144	7102	6.899.428	29.220	0,45	0,076
Unterrichts-/Lehrmethode	365	25.860	52.128.719	55.143	0,42	0,071
Gesamt	816	52.649	83.033.433	146.626	0,40	0,062

SE Standard Error (Standardfehler)

leistungswirksamem Unterricht, 2) eine eingängige, für alle Teilanalysen einheitliche grafische Darstellungsform und Bewertung der Befunde mithilfe eines von ihm so bezeichneten *„Einflussbarometers"* (s. u., Abb. 1) und 3) die Präsentation aller 138 potenziellen Wirkfaktoren in Form von Rankings gemäß der durchschnittlichen Effektstärke.

Hatties theoretischer Ansatz des *„Visible teaching – Visible learning"* geht – stark verkürzt – davon aus, dass Lehr-Lernprozesse dann erfolgreich verlaufen, wenn zwischen Lehrern und Schülern eine *reziproke Perspektivübernahme* gelingt: Lehrkräfte müssen die jeweils anstehenden Lernprozesse mit den Augen ihrer Schülerinnen und Schüler sehen können und Schülerinnen und Schüler müssen die Verantwortung für das eigene Lernen erkennen und dadurch temporär zu ihren eigenen Lehrern werden (Hattie 2009, S. 238). Für beide Prozesse sei Feedback ein entscheidender Faktor. Diese Grundfigur

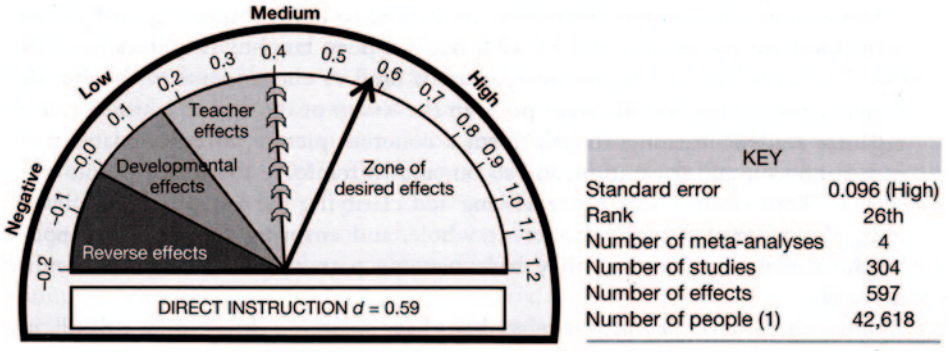

Abb. 1: Typisches Einflussbarometer zur Darstellung der aus Metaanalysen aggregierten mittleren Effektgröße für den Zusammenhang eines Faktors mit Schulleistung. (Copyright: Hattie (2009), S. 205)

Tab. 3: Unterschiedliche Lehrerverhaltensweisen (aktivierend vs. lernbegleitend) in ihrer Wirksamkeit auf Lernerfolge (übersetzt aus Hattie 2009, S. 243)

Lehrkraft als „Aktivator“	*d*	Lehrkraft als „Lernbegleiter“	*d*
Reziprokes Unterrichten	0,75	Lernen mit Planspielen	0,32
Feedback	0,72	Entdeckendes Lernen	0,31
Training von Selbstverbalisationsstrategien	0,67	Reduzierung der Klassengröße	0,21
Metakognitive Strategien	0,67	Individualisiertes Lernen	0,20
Direkte Instruktion	0,59	Problembasiertes Lernen	0,15
Zielerreichendes Lernen *(Mastery Learning)*	0,58	Gender-spezifische Angebote	0,12
Herausfordernde Ziele setzen	0,56	Internetgestütztes Lernen	0,12
Häufiges Testen mit Rückmeldung	0,46	Ganzheitlicher Erstleseunterricht	0,06
Strategien zur Verknüpfung von Lernzielen und Vorwissen	0,41	Induktives Unterrichten	0,06
Mittlerer Effekt „Aktivator“	0,60	Mittlerer Effekt „Lernbegleiter“	0,17

dient Hattie als Folie, um die Masse an Meta-Metanalysebefunden zu ordnen und zu interpretieren. Am Ende stehen zwei Kontrastierungen von Wirkfaktorengruppen, zum einen zweier unterschiedlicher Lehrerverhaltensweisen (nicht notwendig Lehrertypen!) und zum anderen von unterrichtsnahen Merkmalen vs. äußeren Bedingungen des Lehr-Lernkontexts. Demzufolge sind Lehrkräfte empirisch deutlich lernwirksamer, wenn sie eine aktivierende, direkt involvierende und involvierte Rolle einnehmen (*Teacher as Activator*) anstatt sich primär als Lernbegleiter bzw. „Anbieter“ von Lerngelegenheiten *(Teacher as Facilitator)* zu verhalten (vgl. Tab. 3)[3]. Die zweite Kontrastierung belegt nach Hatties (2009, S. 244) Interpretation, dass Merkmale „guten“ Unterrichts wesentlich entscheidender für den Lernerfolg sind als äußere Struktur- und Organisationsmerkmale, wie z. B. reduzierte Klassengröße, jahrgangsübergreifendes Lernen oder externe Differenzierung (vgl. Tab. 4).

Diese beiden Gegenüberstellungen wirksamer und weniger wirksamer Schul- und Unterrichtsfaktoren stehen ganz am Schluss von Hatties (2009) Buch. Zuvor wird für jeden einzelnen der betrachteten Faktoren ein „Einflussbarometer“ präsentiert (vgl.

Tab. 4: Unterrichtsfaktoren und Rahmenbedingungen schulischen Arbeitens in ihrer Wirksamkeit auf Lernerfolge (aus Hattie 2009, S. 244; übersetzt nach Köller 2012)

Unterrichtsmerkmale	*d*	Rahmenbedingungen	*d*
Unterrichtsqualität	0,77	Interne Differenzierung	0,28
Reziprokes Lernen	0,74	Steigerung der Finanzen	0,23
Lehrkraft-Schüler-Verhältnis	0,72	Reduzierung der Klassengröße	0,21
Feedback	0,72	Differenziertes Schulsystem (externe Differenzierung	0,12
Training von Selbstverbalisationsstrategien	0,67	Jahrgangsübergreifender Unterricht	0,04
Metakognitive Strategien	0,67	Offener Unterricht	0,01
Direkte Instruktion	0,59	Lernangebote in den Sommerferien	−0,09
Herausfordernde Ziele setzen	0,57	Sitzenbleiben	−0,16
Mittlerer Effekt	0,68	Mittlerer Effekt	0,08

Abb. 1). In dieser Darstellung mittelt Hattie jeweils die durchschnittlichen Effektstärken *d* aus mehreren publizierten Metaanalysen, bildet also den Durchschnitt von Durchschnittsangaben. Dieser zweifach aggregierte *d*-Wert wird dann im Barometer visuell in einer von vier inhaltlich beschriebenen „Zonen" lokalisiert. Die für Hattie wichtigste „Zonengrenze" liegt bei $d = 0.4$, dem Durchschnittswert der Effektgrößen aller mehr als 800 Metaanalysen seiner Synthese. Oberhalb dieser *Benchmark* liegen, so Hattie (2009, S. 18), die Effekte von pädagogischen Maßnahmen und Programmen, die nicht nur irgendeinen trivialen, von Null verschiedenen Lernerfolg bei den Schülerinnen und Schülern bewirken *(„What works?")*, sondern den anzustrebenden Effekt *(„What works best?")*. In der „Zone" $d < 0$ befinden sich solche Faktoren, die lernhinderlich wirken (z. B. viel Zeit mit Fernsehkonsum, Klassenwiederholung), zwischen $d = 0$ und $d = 0{,}15$ liegen Faktoren, die nicht über das hinaus wirksam sind, was normale kognitive und soziale Reifungsprozesse auch ohne Beschulung hervorbringen würden, die also unter Lernerfolgsgesichtspunkten bereits als potenziell schädlich gelten müssten (z. B. offene vs. traditionelle Lernformen, jahrgangsübergreifendes Lernen). Im Bereich $0{,}15 < d < 0{,}4$ schließlich ordnet Hattie (2009, S. 20) den Lernfortschritt ein, den Lehrkräfte typischerweise innerhalb eines Schuljahres bewirkten.

Die sechs Wirkfaktorengruppen (vgl. Tab. 2 oben) werden nach inhaltlichen Aspekten nochmals zweifach untergliedert: im Bereich *Schule* u. a. in Effekte des *Schultyps,* Effekte von *Klassenkompositionsmerkmalen* und diese weiter in Effekte von *Charter Schools*, *konfessionellen Schulen* etc. bzw. von *Klassengröße*, *Leistungsdifferenzierung*, *Jahrgangsmischung* etc. Für die entstehenden Untergruppen von thematisch relativ homogenen Metaanalysen führt Hattie jeweils seine Meta-Metaanalyse durch, deren Ergebnisse in den beschriebenen Barometer-Darstellungen verdichtet werden. Mit dieser Untergliederung definiert Hattie *a priori* inhaltliche Moderatoren der Wirkung auf Schulleistung.

Alle 138 Einzelfaktoren werden schließlich im Anhang des Bandes nach ihrer Effektstärke in ein Ranking gebracht (vgl. Tab. 5).

3.2 Methodische und methodologische Vorbehalte gegen Hatties Vorgehen

In Abschn. 2.1 wurden zwei Stationen im metaanalytischen Prozess als Weichenstellungen hinsichtlich der Frage charakterisiert, ob das Instrument Metaanalyse eher im Sinne eines an *Wirksamkeits*- oder an *Wirkungsnachweisen* orientierten Verständnisses von Evidenz begriffen wird: die Auswahlstrategie der Primärstudien und die Methode der Effektstärkenaggregation. In Bezug auf die beiden Kardinalprobleme bei der Auswahl, das *Äpfel-und-Birnen*- und das *Garbage-in-Garbage-out*-Problem, votiert Hattie eindeutig für ein inklusives Vorgehen (Hattie 2009, S. 10 f.). Er plädiert dafür, *a priori* möglichst keine Studie (in seinem Fall: Metaanalyse) wegen theoretischer oder methodischer Qualitätsbedenken auszuschließen. Sein Standpunkt ist, dass eine mögliche Heterogenität der Ergebnisse empirisch auf Moderatoreneinflüsse hin untersucht werden müsse.

Zwischen dieser Grundhaltung und seinem eigenen methodischen Vorgehen bei der Aggregation der Effektstärken tut sich allerdings eine Kluft auf. Zwar benennt Hattie in den einzelnen Kapiteln gelegentlich Befunde zu Moderatorenanalysen, die die Autoren der ursprünglichen Metaanalysen berichtet haben; er versäumt es jedoch, die von ihm auf der *Meta*-Metaebene aggregierten Befunde hinsichtlich deren eigener Homo- bzw.

Tab. 5: Ranking ausgewählter effektstarker bzw. -schwacher Faktoren in Hatties (2009) Meta-Metaanalyse (Faktorenbezeichnungen teilweise übersetzt nach Köller 2012)

Faktoren mit lernhinderlichen oder schwachen Effekten	*d*	Faktoren mit sehr starken Effekten	*d*
Mobilität (Umzüge der Eltern)	−0,34	Formatives Feedback für Lehrkräfte	0,90
Erkrankungen des Schülers	−0,23	Interventionen für Lernbehinderte	0,77
Fernsehkonsum	−0,18	Klarheit der Instruktion	0,75
Alleinerziehende Eltern	−0,17	Reziprokes Unterrichten	0,74
Klassenwiederholung	−0,16	Feedback	0,73
Offener Unterricht	0,01	Verteiltes vs. massiertes Lernen	0,71
Jahrgangsübergreifender Unterricht	0,04	Lehrkraft-Schüler-Verhältnis	0,72
Leistungsgruppierung (differenziertes Schulsystem)	0,12	Metakognitive Strategien	0,69
Problembasiertes Lehren	0,15	Wortschatztraining	0,67
Interne Differenzierung	0,16	Förderung der Leseflüssigkeit	0,67

Heterogenität kenntlich zu machen und ggf. Moderatoren zu identifizieren. So wird ein Vergleich der Effekte von Studien aus den 1980er Jahren mit aktuelleren Untersuchungen an keiner Stelle vorgenommen, obwohl methodische Standards sich zwischenzeitlich erheblich verändert haben. Insgesamt unterscheiden sich die einbezogenen Metaanalysen mitunter ganz erheblich in ihren Befunden zu dem jeweils betrachteten Wirkfaktor. Hierdurch blendet Hattie *de facto* eine Ebene seiner mehrfach geschachtelten Datenstruktur aus. Indem er als finale Effektgröße im Einflussbarometer jeweils nur den gemittelten Mittelwert anderer Metaanalysen und eine nicht näher erklärte Angabe zum Standardfehler[4] (*Standard Error*, vgl. oben, Abb. 1) darstellt, wird suggeriert, dass diese „finale" Effektgröße den „wahren" Effekt des betrachteten Schul- oder Unterrichtsfaktors wiedergibt.

Dies wird auch durch Hatties Entscheidung für das *Fixed-Effects*-Modell als statistisches Modell der Datenaggregation unterstrichen. Er begründet diese Entscheidung rein technisch damit, dass fast alle der von ihm aggregierten Metaanalysen ihrerseits ein *Fixed-Effects*-Modell zugrunde gelegt hätten (Hattie 2009, S. 12). Seit ungefähr einem Jahrzehnt werden jedoch *Random-Effects*-Modellen bevorzugt (Cafri et al. 2010; Schmidt et al. 2009b), gerade weil sich in komplexen sozialwissenschaftlichen Handlungsfeldern die Annahme einer „wahren" Effektgröße angesichts vielfältiger Interaktionszusammenhänge (Berliner 2002) nicht plausibel vertreten lässt. Zudem existieren inzwischen statistische Anwendungsvarianten, die eine hierarchische Modellierung von Meta-Metaanalysedaten ermöglichen (z. B. Sterne et al. 2002). Hierbei wird dem Umstand Rechnung getragen, dass sowohl Varianz innerhalb als auch zwischen Metaanalysen modelliert werden kann.

Darüber hinaus aggregiert Hattie (2009) die mittleren Effektstärken der ursprünglichen Metaanalysen, ohne sie durch die Zahl der jeweils eingegangenen Studien zu gewichten. Metaanalysen, die auf vielen hundert Einzelstudien beruhen, gehen dabei mit dem gleichen Gewicht in das *d*- Barometer ein, wie Metaanalysen mit nur fünf Primärstudien. Welche Folgen dieses Vorgehen für die inhaltlichen Schlussfolgerungen haben kann, soll kurz an einem Zahlenbeispiel aus Hatties (2009) Daten demonstriert werden. Der aus vier Metaanalysen ermittelte Effekt der Unterrichtsmethode der Direkten Instruktion *(Direct Instruction)* beträgt nach Hattie (2009, S. 205; vgl. oben Abb. 1) $d = 0{,}59$ und fällt damit

in die „erwünschte Zone“ ($d > 0{,}4$). *Direkte Instruktion* stellt eine keinesfalls unumstrittene, hochstrukturierte und lehrerzentrierte Unterrichtsmethode dar. Schaut man sich die verarbeiteten Metaanalysen einzeln an, so fällt auf, dass die mit 232 Primärstudien bei weitem größte Analyse (Borman et al. 2003) gleichzeitig diejenige mit der geringsten Effektstärke ($d = 0{,}21$) ist. Würde man die drei Metaanalysen, für die Informationen zum Standardfehler vorlagen, nach ihrer Primärstudienanzahl gewichtet mitteln (Hill et al 2007; Shadish und Haddock 2009), so läge die resultierende Effektstärke bei $d = 0{,}39$ und damit nicht mehr in der von Hattie definierten „erwünschten“ Wirkungszone.

Dieses Beispiel soll nicht dazu dienen, Hatties Befunde kategorisch in Frage zu stellen. Dazu bedürfte es einer detaillierten Re-Analyse aller seiner Effektstärkenaggregierungen. Allerdings wäre zu diskutieren, ob die *Sensitivität* solcher datenreichen Analysen ausreichend transparent gemacht wird, um die Ergebnisse unter Evidenzgesichtspunkten angemessen bewerten zu können. Hattie nämlich verwendet das positive Resultat für die Methode der *Direkten Instruktion* später weiter, um sein theoretisches Erklärungsmodell der erfolgreichen *Lehrkraft als „Aktivator“* zu stützen (vgl. oben, Tab. 3). Insbesondere aber die Präsentation möglicherweise sehr sensitiver Effektstärkenschätzungen in Form von Rankings sollte daher nicht unkommentiert an potenzielle Nutzer in Bildungspolitik und pädagogischer Praxis vermittelt werden. Wie Snook et al. (2009, S. 104 f.) resümieren: „We are concerned, however, that: i) Despite his own frequent warnings, politicians may use his work to justify policies which he does not endorse and his research does not sanction; ii) Teachers and teacher educators might try to use the findings in a simplistic way and not, as Hattie wants, as a source for ‚hypotheses for intelligent problem solving‘.“

4 Schlussbemerkung

In diesem Beitrag wurde versucht zu zeigen, wie das Verständnis von Evidenz hinsichtlich der Wirksamkeit pädagogischer und bildungspolitischer Maßnahmen sich in der Art und Weise niederschlägt, wie Metaanalysen gesehen, durchgeführt und genutzt werden. Erfolgen Suche und Auswahl der Primärstudien für eine Metaanalyse nach strengen methodischen Standards, z. B. dem Einschluss nur randomisierter Kontrollstudien, dann liegt der Fokus auf dem Nachweis der Wirksamkeit *(Efficacy)* von Maßnahmen und Programmen. Die Kontext- und Realweltbedingungen werden als grundsätzlich kontrollierbar angenommen. Eine solche Konzeption von empirischen Wirksamkeitsnachweisen mag bestenfalls für sehr genau beschreibbare, eng definierte Maßnahmen im pädagogischen Bereich angemessen sein, wie z. B. Studien zur Wirksamkeit eines klar definierten Leseförderprogramms für dyslektische Kinder, das auch „in der Fläche“ unter immer gleich standardisierten Bedingungen administriert wird. Der Regelfall interessierender Maßnahmen und Programme ist nach Berliner (2002) jedoch gekennzeichnet durch die Wirkung multipler Kontexteinflüsse in realen Lehr-Lernsituationen, die Allgegenwart von Wechselwirkungen zwischen eben solchen Kontextbedingungen und dem interessierenden Wirkfaktor und die geringe „Halbwertszeit“ von Evidenz empirischer Bildungsforschung.

In dieser Perspektive sind es vor allem die deskriptiven und explorativen Potenziale der Metaanalyse, die wertvolle Hinweise für eine Verbesserung von Studiendesigns und

Konstruktoperationalisierungen sowie Erkenntnisse über moderierende Kontextfaktoren ermöglichen und damit zur Steigerung der Validität von Untersuchungen mit notwendigerweise nicht-experimentellem Design geben können. Im Einzelfall können durch metaanalytische Befunde auch politisch motivierte Mythen entzaubert werden, wie z. B. die Vorstellung, dass die Reduzierung der Klassengröße an sich leistungsförderlich wirken müsse.

Weiterhin wurde versucht zu argumentieren, dass die bisher umfangreichste Forschungsbefundsynthese im Bildungsbereich, John Hatties Meta-Metaanalyse *Visible Learning* (2009), zwar an vielen Stellen ein Bekenntnis zur Wichtigkeit von differenzierenden Betrachtungen der Befunde von Metaanalysen abgibt, dass seine eigene methodische Vorgehensweise dies allerdings eher wie ein Lippenbekenntnis erscheinen lässt. Hatties Beispiel zeigt, dass es wichtig und an der Zeit wäre, sich über Standards der quantitativen Befundintegration zu verständigen, auf Seiten der Abnehmer metaanalytischer Befunde ein Verständnis über Möglichkeiten und Grenzen des Verfahrens zu fördern und mögliche Hegemonialansprüche hinsichtlich der Generierung von Evidenz in pädagogischen und bildungspolitischen Handlungsfeldern zu relativieren. Andernfalls droht eine neue Runde in der Dialektik von Versachlichung und Verdinglichung, dieses Mal auf der Meta-Metaebene.

Anmerkungen

1 Dem Validitätskonzept von Cook und Campbell (1979) folgend, wird mit interner Validität die eindeutig kausale Interpretierbarkeit eines Zusammenhangs (Schlüssigkeit) bezeichnet und mit externer Validität die Verallgemeinerbarkeit auf angebbare Zielpopulationen.

2 Prinzipiell abweichende Auffassungen werden auch in der Medizin, z. B. von Goldenberg (2006), Djulbegovic et al. (2009) oder Maier und Shibles (2011), vertreten.

3 Interessanterweise gibt Hattie an dieser Stelle (2009, S. 243) d = 0,06 als mittlere Effektstärke für „Induktives Unterrichten" an. Im Text weiter vorne (S. 208) weist das entsprechende Einfluss-Barometer hingegen d = 0,33 aus; aus dem Anhang seines Buches wird ersichtlich, dass er sich in dieser Tabelle (S. 243) nur auf die kleinere und ältere Metaanalyse von zweien bezieht und diejenige von Klauer und Phye (2008) mit d = 0,59 ohne Angabe von Gründen unberücksichtigt lässt.

4 Aus den Angaben im Text und im Appendix A lässt sich rekonstruieren, dass Hattie (2009) den Standardfehler für seine Meta-Metaeffektstärken berechnet, indem er die Standardfehler der ursprünglichen Metaanalysen ungewichtet mittelt, wenn diese in der Publikation angegeben wurden. Metaanalysen, die keinen Standardfehler berichten, werden in dieser Berechnung ignoriert, was zu nicht unerheblichen Verzerrungen führen kann.

Literatur

Becker, B. J. (2009). Model-based meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Hrsg.), *The handbook of research synthesis and meta-analysis* (2. Aufl., S. 377–395). New York: Russell Sage Foundation.

Beelmann, A., & Bliesener, T. (1994). Aktuelle Probleme und Strategien der Metaanalyse. *Psychologische Rundschau, 45,* 211–233.

Bellmann, J., & Müller, T. (2011). Evidenzbasierte Pädagogik – ein Déjà-vu? In J. Bellmann & T. Müller (Hrsg.), *Wissen, was wirkt. Kritik evidenzbasierter Pädagogik* (S. 9–32). Wiesbaden: VS Verlag für Sozialwissenschaften.

Berliner, D. C. (2002). Comment: Educational research: the hardest science of all. *Educational Researcher, 31*(8), 18–20. doi:10.3102/0013189X031008018.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73,* 125–230. doi:10.3102/00346543073002125.

Brüggemann, A., & Bromme, R. (2006). Anwendungsorientierte Grundlagenforschung in der Psychologie. *Psychologische Rundschau, 57,* 112–116. doi:10.1026/0033-3042.57.2.112.

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*(2), 239–270. doi:10.1080/00273171003680187.

Campbell, C., & Levin, B. (2009). Using data to support educational improvement. *Educational Assessment, Evaluation and Accountability, 21,* 47–65. doi:10.1007/s11092-008-9063-x.

Chatterji, M. (2007). Grades of evidence. *American Journal of Evaluation, 28,* 239–255. doi:10.1177/1098214007304884.

Coe, R. (2009). School improvement: Reality and illusion. *British Journal of Educational Studies, 57*(4), 363–379. doi:10.1111/j.1467-8527.2009.00444.x.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: Lawrence Erlbaum.

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

Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987–2003. *Review of Educational Research, 76*(1), 1–62. doi:10.3102/00346543076001001.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2. Aufl.). 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. doi:10.1177/1094428103257358.

Cortina, K. S., & Pant, H. A. (2009). Hierarchische Linearere Modelle. In H. Holling (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie IV: Evaluation, Bd. 1 Evaluationsforschung: Grundlagen und statistische Methoden der Evaluationsforschung* (S. 335–362). Göttingen: Hogrefe.

Davies, R. S., Williams, D. D., & Yanchar, S. (2008). The use of randomisation in educational research and evaluation: A critical analysis of underlying assumptions. *Evaluation & Research in Education, 21*(4), 303–317. doi:10.1080/09500790802307837.

Djulbegovic, B., Guyatt, G. H., & Ashcroft, R. E. (2009). Epistemologic inquiries in evidence-based medicine. *Cancer Control, 16*(2), 158–168.

Eysenck, H. J. (1984). Meta-analysis: an abuse of research integration. *The Journal of Special Education, 18*(1), 41–59. doi:10.1177/002246698401800106.

Eysenck, H. J. (1995). Meta-analysis squared – does it make sense? *American Psychologist, 50*(2), 110–111. doi:10.1037/0003-066X.50.2.110.

Fischer, F., Waibel, M., & Wecker, C. (2005). Nutzenorientierte Grundlagenforschung im Bildungsbereich. *Zeitschrift für Erziehungswissenschaft, 8*(3), 427–442. doi:10.1007/s11618-005-0149-7.

Goldenberg, M. J. (2006). On evidence and evidence-based medicine: Lessons from the philosophy of science. *Social Science & Medicine, 62*(11), 2621–2632. doi:10.1016/j.socscimed.2005.11.031.

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

Hattie, J. (2010). On being a ‚critic and conscience of society': The role of the education academic in public debates. *New Zealand Journal of Educational Studies, 45,* 85–96.

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Oxford: Routledge.

Hill, C. J. H. S., Bloom, A., Black, R., & Lipsey, M. W. (2007). *Empirical benchmarks for interpreting effect sizes in research*, MDRC Working Papers on Research Methodology. New York: MDRC.

Klauer, K. J., & Phye, G. D. (2008). Inductive reasoning: A training approach. *Review of Educational Research, 78*(1), 85–123. doi:10.3102/0034654307313402.

Köller, O. (2012). What works best in school? Hatties Befunde zu Effekten von Schul- und Unterrichtsvariablen auf Schulleistungen. *Psychologie in Erziehung und Unterricht, 59,* 72–78. doi:10.2378/peu2012.art06d.

Maier, B., & Shibles, W. A. (2011). A critique of evidence-based medicine (EBM): Evidence-based medicine and philosophy-based medicine. In D. N. Weisstub (Hrsg.), *The philosophy and practice of medicine and bioethics* (Bd. 47, S. 453–486). Dordrecht: Springer Netherlands.

Matt, G. E., & Cook, T. D. (2009). Threats to the validity of generalized inferences. In H. Cooper, L. V. Hedges, & J. C. Valentine (Hrsg.), *The handbook of research synthesis and metaanalysis* (2. Aufl., S. 537–560). New York: Russell Sage.

McGee, L. M., & Lomax, R. G. (1990). On combining apples and oranges: A response to Stahl and Miller. *Review of Educational Research, 60,* 133–140. doi:10.3102/00346543060001133.

Ong-Dean, C., Huie Hofstetter, C., & Strick, B. R. (2011). Challenges and dilemmas in implementing random assignment in educational research. *American Journal of Evaluation, 32*(1), 29–49. doi:10.1177/1098214010376532.

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics, 30,* 85–110. doi:10.1017/S 0267190510000115.

Robbins, S. B., Oh, I.-S., Le, H., & Button, C. (2009). Intervention effects on college performance and retention as mediated by motivational, emotional, and social control factors: Integrated meta-analytic path analyses. *Journal of Applied Psychology, 94*(5), 1163–1184. doi:10.1037/a0015738.

Ruthven, K. (2011). Using international study series and meta-analytic research syntheses to scope pedagogical development aimed at improving student attitude and achievement in school mathematics and science. *International Journal of Science and Mathematics Education, 9,* 419–458. doi:10.1007/s10763-010-9243-2.

Schildkamp, K., Ehren, M., & Lai, M. K. (2012). Editorial article for the special issue on data-based decision making around the world: from policy to practice to results. *School Effectiveness and School Improvement, 23*(2), 123–131. doi:10.1080/09243453.2011.652122.

Schmid, C. H., Stewart, G. B., Rothstein, H. R., Lajeunesse, M. J., & Gurevitch, J. (2013). Software for statistical meta-analysis. In J. Koricheva, J. Gurevitch, & K. Mengersen (Hrsg.), *Handbook of meta-analysis in ecology and evolution* (S. 174–193). Princeton: Princeton University Press.

Schmidt, F. L., Le, H., & Oh, I.-S. (2009a). Correcting for distorting effects of study artifacts in meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Hrsg.), *The handbook of research synthesis and metaanalysis* (2. Aufl., S. 317–335). New York: Russell Sage.

Schmidt, F. L., Oh, I.-S., & Hayes, T. L. (2009b). 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. doi:10.1348/000711007X255327.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499. doi:10.3102/0034654307310317.

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

Shadish, W. R., & Haddock, C. K. (2009). Combining estimates of effect size. In H. Cooper, L. V. Hedges, & J. C. Valentine (Hrsg.), *The handbook of research synthesis and metaanalysis* (2. Aufl., S. 257–278). New York: Russell Sage.

Slavin, R. E. (2008). Perspectives on evidence-based research in education, what works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*(1), 5–14. doi:10.3102/0013189X08314117.

Snook, I., O'Neill, J., Clark, J., O'Neill, A.-M., & Openshaw, R. (2009). Invisible learnings? A commentary on John Hattie's book – ‚Visible Learning: A synthesis of over 800 meta-analyses relating to achievement'. *New Zealand Journal of Educational Studies, 44,* 93–106.

Sohn, D. (1995). Meta-analysis as a means of discovery. *American Psychologist, 50*(2), 108–110. doi:10.1037/0003-066X.50.2.108.

Song, M., & Herman, R. (2010). Critical issues and common pitfalls in designing and conducting impact studies in education. *Educational Evaluation and Policy Analysis, 32*(3), 351–371. doi:10.3102/0162373710373389.

Stahl, S. A., & Miller, P. D. (1989). Whole language and language experience approaches for beginning reading: A quantitative research synthesis. *Review of Educational Research, 59*(1), 87–116. doi:10.3102/00346543059001087.

Sterne, J. A. C., Jüni, P., Schulz, K. F., Altman, D. G., Bartlett, C., & Egger, M. (2002). Statistical methods for assessing the influence of study characteristics on treatment effects in ‚meta-epidemiological' research. *Statistics in Medicine, 21*(11), 1513–1524. doi:10.1002/sim.1184.

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning. *Review of Educational Research, 81*(1), 4–28. doi:10.3102/0034654310393361.

Terhart, E. (2011). Has John Hattie really found the holy grail of research on teaching? An extended review of Visible Learning. *Journal of Curriculum Studies, 43*(3), 425–438. doi:10.1080/00220272.2011.576774.

Trautwein, U., Schnyder, I., Niggli, A., Neumann, M., & Lüdtke, O. (2009). Chameleon effects in homework research: The homework-achievement association depends on the measures used and the level of analysis chosen. *Contemporary Educational Psychology, 34*(1), 77–88.

U.S. Department of Education, Office of Elementary and Secondary Education. (2002). *No Child Left Behind: A desktop reference, Washington, D.C*. http://www2.ed.gov/admins/lead/account/nclbreference/reference.pdf. Zugegriffen: 16. Nov. 2013.

U.S. Department of Education. (2005). *Scientifically based evaluation methods: Notice of final priority*. Federal Register FR Doc. 05–1317. http://www2.ed.gov/legislation/FedRegister/finrule/2005-1/012505a.pdf. Zugegriffen: 16. Nov. 2013.

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

Volmink, J., Siegfried, N., Robertson, K., & Gülmezoglu, A. M. (2004). Research synthesis and dissemination as a bridge to knowledge management: the Cochrane Collaboration. *Bulletin of the World Health Organization, 82*(10), 778–783.

What Works Clearinghouse. (2011). *WWC procedures and standards handbook* (Version 3.0). http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19. Zugegriffen: 16. Nov. 2013.

Winne, P. H., & Nesbit, J. C. (2010). The psychology of academic achievement. *Annual Review of Psychology, 61,* 653–678. doi:10.1146/annurev.psych.093008.100348.

Wiseman, A. W. (2010). The uses of evidence for educational policymaking: Global contexts and international trends. *Review of Research in Education, 34*(1), 1–24. doi:10.3102/0091732X09350472.

Wortman, P. M. (1983). Evaluation research: A methodological perspective. *Annual Review of Psychology, 34,* 223–260. doi:10.1146/annurev.ps.34.020183.001255.

Zimmer, R., Gill, B., Razquin, P., Booker, K., & Lockwood III, J. R. (2007). *State and local implementation of the No Child Left Behind Act: Volume I – Title I School Choice, Supplemental Educational Services, and Student Achievement*. Washington D.C.: U.S. Department of Education.