Rolf Schulmeister / Jörn Loviscach

Kritische Anmerkungen zur Studie „Lernen sichtbar machen“ (Visible Learning)[1] von John Hattie

John Hattie hat mit seiner Synthese „Visible Learning“ von mehreren hundert Meta-Analysen ein beeindruckendes Werk produziert, das vor allem deshalb breite Zustimmung erfahren hat, weil die inhaltlichen qualitativen Erkenntnisse, zu denen er mit Bezug auf Lernende, Lehrende, Unterrichtsmethoden, Schuleinfluss usw. gelangt, von vielen Rezipienten akzeptiert werden können (Spiewak, 2013). Daran soll im Folgenden keine Kritik geübt werden. Wir wollen allerdings eine Reihe methodischer Bedenken gegenüber Meta-Analysen an Beispielen aus Hatties Analysen illustrieren und einige handwerkliche Fehler seiner Studie aufzeigen.

Die Kritik an Meta-Analysen ist fast ebenso alt wie die Methode selbst. Dazu zählt der Einwand, man würde Äpfel mit Birnen vergleichen, indem man mehrere Studien zusammenfasst. Es zählt dazu auch die Meinung, man dürfe keine Mittelwerte von Effektstärken bilden (s. Brügelmann, o. J.). Eine weitere Kritik betont: Da es sich bei Meta-Analysen um post-factum-Analysen handelt, gehen in solche Analysen meistens ältere Studien ein, die durch neuere Entwicklungen längst überholt sein können. Tatsächlich befasst sich Hattie mit Meta-Analysen, die in einigen Effektgruppen aus den 1970/80er Jahren stammen und deren inkludierte empirische Studien noch älter sein können (s. die Kritik von Terhart, 2011). Ausgenommen von diesem Vorwurf sind Themen, die mit Technologien, Medien, Computer etc. zu tun haben – aber auch hier kann man bei einer Zeitdifferenz von 20 Jahren schon von „veraltet“ sprechen. Insofern kann man Terharts Argument zustimmen, dass möglicherweise in vielen Bereichen die Wissenschaft inzwischen weiter ist, als es der Stand der Originalstudien und deren Meta-Analysen repräsentiert. Es haben sich zwar einige mit einer allgemeinen Kritik an Hattie hervorgetan und inhaltlich mit seinen Schlussfolgerungen auseinandergesetzt (z. B. Lind, 2013); aber bisher scheint sich niemand die Mühe gemacht zu haben, nachzuvollziehen, wie Hattie gearbeitet hat.

Hattie selbst nimmt zu den bekannten Argumenten gegen Meta-Analysen in seinem Buch Stellung. Und auch wir betrachten die Argumente nicht als grundsätzliche Einwände gegen Meta-Analysen, sondern denken, dass es Meta-Analysen geben könnte, die diese Fehler vermeiden. Aber wir halten es für unbedingt geboten, die Einwände empirisch am Fall zu klären, was bedeutet, dass man

1. die zu Effektgruppen zusammengefassten Meta-Analysen auf ihre inhaltlich-qualitative Vergleichbarkeit durch geschulte Dritte überprüfen müsste, wie es in der empirischen Sozialforschung z. B. bei Codierung von Interviewtexten üblich ist,
2. den Einfluss extremer Werte einzelner oder weniger Studien auf die Mittelwerte stärker kontrollieren müsste,
3. die historische Bedingtheit älterer Meta-Analysen klären müsste, falls keine neueren Meta-Analysen zu den jeweiligen Faktoren zur Verfügung stehen.

Es soll an dieser Stelle keine generelle Kritik der Hattie-Studie (s. Lind, 2013) oder des Konzepts der Effektstärke durchgeführt werden (s. Lind, 2014). Wir werden uns auf eine

[1] Die Anmerkungen beziehen sich auf die zweite, korrigierte Ausgabe (Hattie 2014).

stichprobenartige Analyse der Meta-Meta-Analyse von John Hattie beschränken und dabei auf die ersten beiden Kritikpunkte näher eingehen, um Mängel der Studie aufzuzeigen. Wir haben willkürlich wenige Stellen herausgegriffen und wollen es bei diesem exemplarischen Vorgehen belassen, weil eine komplette Re-Analyse ein größeres Projekt erfordern würde.

Zur Statistik der Meta-Meta-Analyse

Die Zusammenfassung von Effektstärken aus mehreren Einzelstudien zu einem Mittelwert ist ein erster problematischer Schritt in Meta-Analysen, weil dadurch die Streuung der Einzelstudien nicht mehr erkennbar ist: Der Mittelwert sagt nichts über die Breite der Verteilung. Dabei sind die Verteilung, insbesondere ihre Breite und ihre Extrema sehr wichtige Informationen z. B. über die Homogenität der Stichprobe oder die in der Stichprobe eventuell enthaltenen Minderheiten. Aber die Leser solcher Meta-Analysen haben gelernt, mit diesem ersten Schritt der Mittelwertbildung auskommen zu müssen, weil sie im Überblick über die Einzelstudien die Verteilung noch erkennen können – zumindest, wenn die Meta-Analyse eine entsprechende Übersichtstabelle über die inkludierten Studien anbietet.

Leider fehlt meist schon in solchen Tabellen eine direkte Angabe über die Breite der jeweils gemessenen Verteilungen. Oft wird ein Standardfehler angegeben; aber dieser sagt zunächst nur etwas darüber aus, wie sicher man sich beim geschätzten Erwartungswert ist. Wirft man einen Würfel 1000mal, mag man dessen Erwartungswert von 3,5 mit einem Standardfehler von 0,05 bestimmen. Nichtsdestotrotz steuern die einzelnen „Würfelexperimente“ von 1 bis 6.

Problematisch wird es erst, wenn wie bei Hattie eine Mittelwertbildung auf einer Ebene oberhalb von Meta-Analysen erfolgt, als Aggregation der Mittelwerte der Effektstärken mehrerer Meta-Analysen (s. Brügelmann). Warum? Die Varianz der empirischen Studien, die in den Originalstudien noch enthalten war, geht bereits in den Meta-Analysen durch die Mittelwertbildung bei der Kalkulation von Effektstärken verloren und wird nun bei Hattie durch einen zweiten Vorgang der Mittelwertbildung mehrerer Effektstärken völlig nivelliert.

Im Text diskutiert Hattie nicht mehr die Streuungen der Einzelfälle, gibt aber immerhin öfters Gründe für die variierenden gemittelten Effektstärken der Metastudien an. „Instead of considering only the size of an effect, we should be looking for patterns and implications across effect sizes.“ schreibt er in einem aktuellen Buchkapitel (Hattie et al., 2014, S. 200). Allerdings werden auch diese noch verbliebenen Unterschiede durch die Mittelwertbildung der Meta-Meta-Meta-Analysen eingeebnet. Noch fragwürdiger ist die Meta-Meta-Meta-Analyse der Faktoren „Lernende“, „Elternhaus“, „Schule“, „Lehrperson“, „Unterrichten“ in der Tabelle 01 auf Seite 22.

Das statistische Interesse gilt in der wissenschaftlichen empirischen Forschung stets der Varianz, nämlich der Frage, inwiefern sich der Effekt von Versuchsperson zu Versuchsperson unterscheidet. Der Vorgang der Mittelwertbildung von Korrelationen und Effektstärken, der in Meta-Analysen gewählt wird, ist diesem Erkenntnisinteresse bereits abträglich. Das von Hattie gewählte Verfahren verstärkt den Verlust an Varianz nochmals, indem er die streuenden Mittelwerte der Effektstärken aus mehreren Meta-Analysen ein weiteres Mal zu einem Mittelwert aggregiert.

In den kleinen Tabellen neben dem „Tachometer“ jeder Effektgruppe findet sich auch eine Angabe zu einem Standardfehler. Die Beschreibung auf S. 23 f. lässt darauf schließen, dass

dies der Standardfehler der geschätzten Effektstärke der Gruppe sein soll. Aber dort steht bloß der ungewichtete Mittelwert der Standardfehler (soweit verfügbar!) der Meta-Analysen dieser Gruppe. So werden etwa bei „Integrierte Curricula“ zwei Meta-Analysen mit d = 0,48 bzw. 0,31 und Standardfehler 0,086 bzw. 0,015 im Mittel zu d = 0,39 und Standardfehler 0,050. Die beiden Studien sollen damit also mehr als drei Standardfehler auseinander liegen. Das wäre sehr unwahrscheichlich. Der Standardfehler muss deutlich größer als der angegebene Wert von 0,05 sein.

Diskussionswürdig ist schon, dass Hattie die mittleren Effektstärken schlicht mit dem ungewichteten arithmetischen Mittel bildet. Weil für viele der benutzten Meta-Analysen Daten über die Streuung und/oder die Zahl an Personen fehlen, ist dies ein naheliegender Ansatz. Man könnte allerdings solche unvollständig dokumentierten Meta-Analysen auch als mangelhaft aus der Meta-Meta-Analyse ausschließen. Ein Beispiel für solche fehlenden Angaben ist die im Komplex „Feedback“ zitierte Metastudie von Standley. Die Autorin erwähnt selbst, dass die zwei der vor ihr verwendeten Studien mit den gigantischen Werten d=35 und d=33 Studien „with single-subject“ sind, will sagen: N=1. Um die Fallzahlen der anderen Studien herauszufinden, die sie benutzt hat, müsste man sich durch knapp 100 Artikel wühlen, teilweise aus den 1960er Jahren.

Dass Hattie das ungewichtete arithmetische Mittel bildet, könnte teilweise drastische Auswirkungen haben. Wir haben uns die Effektgruppe „Geburtsgewicht“ angesehen, denn dafür werden nur zwei Meta-Analysen verwendet, für die obendrein die Zahlen der erfassten Personen vorliegen. Die Analyse von Bhutta et al. umfasst 3276 Kinder, die von Corbett & Drewett 1213. Als Effektstärken listet Hattie d=0,73 bzw. 0,34. (Ersteres ist überraschend, weil Bhutta et al. $r^2=0,51$ angeben, man also d=2,0 erwarten würde. Aber ignorieren wir dies.) Hattie mittelt die beiden Effektstärken zu (0,74+0,34)/2=0,54. Wenn man auf simpelste Weise die Zahlen der Kinder berücksichtigt (was in diesem Fall einfach geht, weil die Zahlen von Fällen und Kontrollen praktisch gleich sind), erhält man dagegen (3276·0,73+1213·0,34)/(3276+1213)=0,62. Damit würde dieser Komplex in Hatties Rangliste einen Sprung nach oben machen. Falls nicht nur die Fallzahlen, sondern sogar auch die Varianzen der d-Werte der Meta-Analysen verfügbar wären, ließe sich die Rechnung noch verfeinern.

Das genaue Vorgehen bei der Wichtung wird davon abhängen, ob man davon ausgeht, dass die Studien exakt denselben Effekt messen („fixed effect“, sozusagen nur Exemplare einer einzigen Apfelsorte) oder nicht („random effect“, sozusagen alle Obstarten). Im Extrem kann man behaupten, dass jede der (Meta-)Studien einem anderen, quasi-zufälligen Punkt im Universum der Effekte entspricht. Dann wäre in der Tat das ungewichtete arithmetische Mittel denkbar, um eine mittlere Effektstärke des Komplexes zu bilden. Aber dies müsste man jeweils pro Komplex diskutieren.

Es werden die Effektstärken von Studien zusammengefasst, die unterschiedlicher kaum sein können. Die Mittelwertbildung bei Themengruppen, zu denen es sehr viele Studien gibt, könnte im Sinne eines „random effect“ noch als halbwegs sinnvoll erscheinen, sofern sich eine kompakte Verteilung ergibt, so dass der Mittelwert der Masse zumindest eine statistische Aussage macht, selbst wenn zwischen den Studien deutliche semantische Unterschiede existieren. So vereint die Gruppe „Computerunterstützung“ (S. 260 ff.) 81 Meta-Analysen, deren Mittelwert zumindest auch ins Intervall der häufigsten Effektwerte fällt. Zum Beispiel bei der Themengruppe „Feedback“ führt dagegen der Ausreißer d=2,87 (nämlich die in diesem Beitrag mehrfach angesprochene Metastudie von Standley) zu einer ausgesprochen schiefen Verteilung (Schiefe von knapp 3).

Besonders problematisch erscheint es uns, wenn extreme Unterschiede zwischen zwei[2] oder drei Studien, die eine Gruppe bilden, durch den Mittelwert nivelliert werden. So berichtet Hattie beispielsweise für die Effektgruppe „Induktives Lernen“, in der es nur zwei Meta-Analysen gibt, die Effektstärken d=0,06 und d=0,59 und zieht sie zur Effektstärke d=0,33 zusammen. Die beiden Meta-Analysen liegen 15 Standardfehler der zweiten Studie auseinander. Schon das macht es unwahrscheinlich, dass sie denselben Effekt messen. Hatties Beschreibungen (S. 246) legen das auch inhaltlich nahe. Dies ist also ein Fall für eine „random effects“-Analyse: Man versteht die beiden Meta-Analysen als zwei zufällige Punkte im Universum der Effekte, die zu dieser Gruppe gehören. Auf Basis von zwei Stichproben eine Aussage über eine Gruppe zu treffen, ist allerdings gewagt. Es lässt sich aus diesen beiden Werten nicht einmal plausibel schließen, ob der Mittelwert der Gesamtheit *über* oder aber *unter* 0 liegt.

Probleme dieser Art ergeben sich immer, wenn nur wenige Studien zusammgenfasst werden, deren Effektstärken (und deshalb höchstwahrscheinlich auch deren Effektarten) sich gravierend unterscheiden. Nehmen wir an, es seien drei Studien, die betrachtet werden, dann kann man zwei Fälle unterscheiden: Im ersten Fall sind die Effektstärken aller drei deutlich verschieden, dann erhalten wir einen nichtssagenden Mittelwert einer breiten Verteilung. Etwa bei der Effektgruppe „Klassenzusammenhalt“ werden drei Meta-Analysen mit den jeweiligen Effektstärken 0,17, 0,92 und 0,51 zu einer gemeinsamen Effektstärke von 0,53 verrechnet. Der angegebene Standardfehler von 0,016 spiegelt diese Spanne nicht wider; vielmehr hat Hattie hier den einzigen in den drei Meta-Analysen verfügbaren Standardfehler benutzt.

Im zweiten Fall sind zwei der Effekte ähnlich hoch bzw. niedrig. Dann bestimmen die beiden hohen bzw. niedrigen Werte den Mittelwert (Beispiel: „Jahrgangsübergreifende Klassen“ mit den Werten -0,03, -0,01 und 0,17 ergibt einen Mittelwert von 0,04; „Berufswahlunterricht“ mit Werten von 0,50, 0,48 und 0,17 ergibt im Mittel 0,38). In solchen Fällen wäre es sinnvoller, nach qualitativen inhaltlichen Gründen für den vermeintlichen Ausreißer zu suchen. Für solche Fälle sollte man Grenzen bestimmen, wie groß die Spannen zwischen den Werten sein dürfen, um sie mitteln zu können, zum Beispiel ein Abstand zwischen Mittelwert und Median von weniger als einer halben (geschätzten) Standardabweichung. Jenseits dieser Grenzen sollte man gezwungen sein, nach inhaltlichen Gründen für die Verschiedenheit zu suchen oder zumindest die Breite der Verteilung anzugeben.

Zum Faktor „Klarheit der Lehrperson“ wird nur eine einzige Meta-Analyse von Fendick mit einer Effektstärke d = 0,75 herangezogen. Zudem handelt es sich bei der Arbeit von Fendick um eine „unpublished dissertation“ (s. u.). Unbekannt ist die Anzahl der Studien, die in diese Analyse eingegangen sind; unbekannt ist auch die Zahl der Versuchspersonen sowie die Zahl der ermittelten Effekte. Thematisch vergleichbare Meta-Analysen, die vielleicht niedrigere Werte eingebracht und damit den Mittelwert gedrückt hätten, sind in der Hattie-Studie nicht enthalten. Und so landet dieser Effekt auf Rangplatz 8. Die meisten anderen Effektgruppen beinhalten mehrere Meta-Analysen, sodass nicht eine einzelne

[2] Es gibt mehrere Effektgruppen, die nur mit jeweils zwei Meta-Analysen gestützt werden. Sofern deren Werte praktisch gleich sind, erscheint eine Mittelwertbildung als einfach (z. B. „Akzeleration“, „Reziprokes Lehren“, „Konfessionsschulen“, „Interne Differenzierung“, „Lernen in Kleingruppen“, „Wiederholendes Lesen“, „Werte- und Moralerziehung“, „Metakognitive Strategien“, „College-Förderkurse“). Unterschiede in der Effektstärke bestehen bei „Motorische Aktivität und Gender“, „Hausbesuche durch Lehrende“, „Sätze kombinieren“, „Leseförderung“, „Spielförderung“, „Freiarbeit“, „Zuschnitt von Methoden auf Schülermerkmale“.

Meta-Analyse über den Rangplatz entscheidet. Weitere Effekte, die nur mit einer einzigen Meta-Analyse belegt wurden, sind „Kreativität“, „Kognitive Entwicklungsstufe nach Piaget“, „Spezielle Ernährung“, „Positive Sicht auf die eigene Ethnizität“, „Bezug staatlicher Transferleistungen“, „Charter Schools“, „Internatsunterbringung“, „Schulgröße“, „Dauer der Sommerferien“, „Klassenführung“, „Peer-Einflüsse“, „Lehrpersoneneffekte“, „Lehrer-Schüler-Beziehung“, „Bewegungserziehung“, „Taktile Stimulation“, „Unmittelbarkeit der Rückmeldung“, „Nichtetikettieren von Lernenden“, „Lernzielhierarchisierung“, „Fallbeispiele“, „Technologiegestütztes Lernen zu Hause“.

Sobald mehr als zwei Meta-Analysen zu einer Effektgruppe erfasst wurden, die alle ähnliche Effektstärken aufweisen (z. B. die drei Studien in den Effektgruppen „Fernsehen“ und „Sommerschulen“), scheint die Zusammenfassung der Effektstärken unproblematisch zu sein, denn die Mittelwertbildung verändert dann ja nichts oder nicht viel. So streuen zum Beispiel die Effektstärken der Gruppe „Selbstkonzept“ mit sechs Meta-Analysen zwar von d=0,32 bis d=0,76, aber selbst das ist vergleichsweise wenig im Vergleich mit anderen Effektgruppen. Der Mittelwert von 0,43 ist noch nah an fünf der sechs Meta-Analysen, wenn auch deutlich wird, dass bereits ein einziger Wert den Mittelwert von den fünf anderen Werten wegzieht und die Effektstärken von fünf Meta-Analysen kleiner sind als der Mittelwert und nur der von einer größer ist als der Mittelwert.

Einige Effektgruppen im Detail

Im Folgenden haben wir willkürlich Effektgruppen herausgegriffen und genauer betrachtet, deren Literatur mit überschaubarem Aufwand beschafft werden konnte.

Selbsteinschätzung des eigenen Leistungsniveaus (S. 53 und S. 380)

Auf Rang 1 der Hattie-Liste landet der Themenkomplex „Selbsteinschätzung des eigenen Leistungsniveaus“ („self-reported grades“). Sechs Meta-Analysen sind zusammengefasst worden, als Effektstärke wird von Hattie d=1,44 angegeben. In dieser Gruppe sind Meta-Analysen von:

	d
Mabe & West	0,93
Falchikov & Boud	0,47
Ross	1,63
Falchikov & Goldfinch	1,91
Kuncel, Crede & Thomas	3,10
Kuncel, Crede & Thomas	0,60

Mabe & West berichten aus 55 Studien Korrelationen mit einer großen Streuung von r=-0,26 bis r=0,80, die sie zum Mittelwert von r=0,29 zusammenfassen bei einer Standardabweichung von r=0,25. Das scheint nicht unberechtigt zu sein, weil auch der Modus in der Gegend des Mittelwerts liegt, aber die hohe Streuung stimmt nachdenklich. 43 der 55 Studien geben Korrelationen an, die in die Analyse einbezogen werden. Es werden manche Subgruppen erwähnt, die das eine oder andere Kriterium nicht erfüllen und zur Reduktion der Stichprobe führen könnten. Die Angabe von n=35 Studien mit 35 Effekten und 13.565 Versuchspersonen, die in der Tabelle bei Hattie steht, kommt so bei Mabe & West aber nirgends vor. Mabe & West reduzieren die Zahl der Studien auf 43 und gewichten die mittlere Korrelation aufgrund von Stichprobengrößen und Reliabilität neu. So bereinigen sie die mittlere Korrelation erst zu r=0,31, dann zu 0,36 und schließlich zu 0,42. Mabe & West

warnen, dass es sich bei den rechnerisch korrigierten Werten nur um grobe Approximationen an die „wahren“ Werte handelt, „given the large number of missing reliability coefficients that had to be estimated.“ Hattie transformiert die Korrelation von r=0,42 in die Effektstärke d=0,93.

Falchikov et al. geben die Effektstärken der einbezogenen Studien an. Sie variieren von d=-0,62 bis d=1,42 mit einem Mittelwert von d=0,47. Kann man bei derart großer Streuung dem Mittelwert noch trauen? Vorsichtshalber rechnen die Autoren Korrelationen nicht in d-Werte um, sondern berechnen sie getrennt: „Correlation coefficients (r) relating to the relationship between teacher and student marks vary from -0.05 to 0.82, with the mean value of r being 0.39.“ Falchikov et al. stellen sich dieselbe Frage wie wir. Die Skepsis ist berechtigt, denn bei den Studien, die Korrelationen berichten, würde die Cohensche Effektstärke d=0,85 ergeben.

Kuncel, Crede & Thomas kommen in dieser Gruppe mit Teilstichproben zweimal vor und bieten Daten zu mehreren unterschiedlichen Effekten. Für beide Effekte gibt Hattie die Studienanzahl von 29 und die sample size von 56.265 an. Diese Zahlen beziehen sich jedoch auf die gesamte Studie und nicht auf die beiden selektierten Teilstichproben mit 12 Studien und 12.089 Probanden, die den College-GPA betreffen. Hattie rechnet in eine Effektstärke von d=3,10 um, s.u. Der zweite Aspekt, die Differenz zwischen Selbsteinschätzung und GPA misst bei ihnen d=1,38 (College), bzw. d=0,32 (High School). Woher der hier von Hattie verwendete Wert von d=0,60 stammt, bleibt unklar. Kuncel, Crede & Thomas geben die Effektstärken der einbezogenen Studien nicht an: die angegebenen Mittelwerte können daher nicht nachvollzogen werden.

Unter allen 724 (s.u.) Meta-Analysen, die Hattie einbezogen hat, gibt es nur zwei mit Effektstärken größer als d=2,0. Der Spitzenwert von d=3,1 kommt in dem hier beleuchteten Themenkomplex vor: In der Studie von Kuncel, Crede & Thomas ist eine Korrelation von r=0,84 mit einem 90%-Glaubwürdigkeitsintervall von 0,70 bis 0,94 angegeben. Hattie hat daraus wohl mit der üblichen Umrechnung eine entsprechende Effektstärke von d=3,1 bestimmt. Was dabei unter den Tisch fällt, ist, dass sich das 90%-Glaubwürdigkeitsintervall für dieses d über den extrem großen Bereich von 2,0 bis 5,5 erstreckt: Die Umrechnung von fehlerbehafteten r-Werten zu d-Werten ist für hohe Korrelationen r heikel, denn ein r von 1,0 muss theoretisch zu einer unendlich hohen Effektstärke d führen.

Konzentration, Ausdauer und Engagement (S. 59 und S. 382)

Es stellt sich die Frage, ob die fünf in dieser Gruppe zusammengefassten Studien überhaupt in eine Gruppe gehören: Konzentration, Ausdauer und Engagement sind Variablen, denen man am besten mit Tests zu Leibe rückt. Die Operationalisierung müsste im Grunde Methoden umfassen, die genau diese Konstrukte messen. Das ist bei keiner der in diesem Komplex enthaltenen Meta-Analysen der Fall.

Die Meta-Analyse von Feltz & Landers hat den Effekt mentaler Übungen auf das motorische Lernen zum Gegenstand; die Analyse von Kumar befasst sich mit dem Effekt von Lehrstrategien auf die Zeit, die Schüler mit einer Aufgabe verbringen; die Analyse von Cooper & Dorr vergleicht die Leistungsmotivation von ethnischen Gruppen, und die Studie von Mikolashek untersucht die Belastbarkeit von Schülern. Auch hier scheinen die Studien wenig gemein zu haben. Aber das soll in diesem Fall nicht das Hauptargument sein.

Die Meta-Analyse von Kumar befasst sich mit „student on-task ‘engagement’“ im naturwissenschaftlichen Unterricht, was aber weder Konzentration noch „Engagement für

Naturwissenschaften“ bedeutet, wie es bei Hattie heißt, sondern mit Kumars eigenen Worten: „An appropriate definition of on-task engagement is the effective time within the allocated class time a student actively participates in learning“ (S. 50). Die Studien, die Kumar betrachtet, befassen sich mit instruktionalen Lehrerstrategien. Die unabhängige Variable ist das Lehrerhandeln, die abhängige die Zeit, die Pre-College-Studierende mit der Aufgabe verbringen. Diese Meta-Analyse ließe sich eher im dem Teil des Buches von Hattie ansiedeln, der sich mit Unterrichtsstrategien befasst als in diesem Komplex, der den Lernenden zugeordnet ist. Mit anderen Worten: Die Effektstärke gilt hier den Lehrenden und ihren instruktionalen Strategien, aber nicht den Lernenden.

Die Meta-Analyse von Cooper & Dorr ist eigentlich eine Rezension einer interpretierenden (narrativen) Studie von Graham, ein Methodenvergleich von normativen und empirischen Methoden. Die Studie vergleicht Unterschiede der Leistungsmotivation („need-for-achievement“) zwischen ethnischen Gruppen und sammelt zu dem Zweck „race-comparative“-Studien. Die Studie dient der Widerlegung einer früheren narrativen Arbeit von Graham und sucht vor dem Hintergrund ethnischer Unterschiede nach Differenzen in Alter, Schulniveau, soziökonomischem Status etc. Der Mittelwert der Effektstärken hingegen ist in dieser Studie weniger relevant.

Die Meta-Analyse von Mikolashek fasst 28 Studien zur Belastbarkeit („resilience“) bei Risikoschülern zusammen, wobei Belastbarkeit operationalisiert wurde durch Testergebnisse und Noten, also nicht durch einen zielgerichteten Test auf Belastung. Diese Studie ist nicht mit psychologischen Studien gleich zu setzen, die Belastung als Konstrukt messen. Belastbarkeit ist für Mikolashek das Bestehen von Leistungsanforderungen, und moderierende Variablen wie Familienbeziehungen sollen die Risikoschüler beim Bestehen unterstützen.

Gravierender ist jedoch, dass mit Datta & Narayanan in dieser Gruppe eine Studie auftaucht, in der mit dem Begriff „concentration“ die industrielle wirtschaftliche Konzentration bezeichnet wird. Diese Studie hat offenbar weder Hattie selbst noch einer vom Hattie-Team wirklich gelesen, man hat nur auf die Effektstärke geachtet. Diese ökonomische Meta-Analyse gehört nicht nur nicht in diese Gruppe, sondern überhaupt nicht in die Studie von Hattie. So weit wollen wir den Begriff Konzentration beim Lernen nicht dehnen.

Aber schon die Überschrift der Effektgruppe ist problematisch. Sie besagt, dass Konzentration, Ausdauer und Engagement semantisch in eine Gruppe passen. Das ist schwer nachzuvollziehen. Konzentration und Aufmerksamkeit könnten zusammengehören; Ausdauer hingegen würde besser zu Fleiß und Leistung passen; Engagement wäre etwas Drittes. Dass alle im Hintergrund etwas mit Motivation zu tun haben, bedeutet nicht, dass man derart unterschiedliche Studien zusammenfassen sollte.

Nach Studium der fünf Meta-Analysen in dieser Gruppe kann man zu der Überzeugung gelangen (*nicht* aus *prinzipiellen* Gründen, sondern durch die Einzelbewertung der herangezogenen Analysen), dass man diese Gruppe gar nicht hätte bilden dürfen. Hattie berechnet eine Effektstärke für den Faktorenkomplex von d=0,48. Je nachdem, welche der Studien man in dieser Gruppe meint beibehalten zu können, ändert sich die berechnete Effektstärke. Für die Studien von Kumar nimmt Hattie die Effektstärke von d=1,09 an (umgerechnete Korrelation). Nimmt man diese Studie sowie die „falsch gelesene“ von Datta & Narayanan aus der Gruppe heraus, dann ergäbe sich eine Effektstärke von d=0,24, und die Variablengruppe würde nicht auf Rang 49, sondern unterhalb von Rang 94 platziert werden müssen.

Extracurriculare Aktivitäten (S. 188 und S. 409)

Hattie fasst in dieser Gruppe fünf Meta-Analysen zusammen. Man mag darüber streiten, ob es in Wirklichkeit nur drei sind, weil eine Meta-Analyse für drei gezählt wird (Nr. 444 bis 446) wegen mehrerer registrierter Effekte in derselben Meta-Analyse. Es handelt sich um eine der üblichen amerikanischen „unpublished dissertations“ (s. u.).

Die Dissertation stammt von einer Studentin namens Charla Patrice Lewis, die im Text aber mehrfach als männliches Wesen adressiert wird (S. 188). Die von Charla Lewis ermittelten Effekte betreffen allgemeine extracurriculare Aktivitäten, Sport und Arbeiten mit den sehr unterschiedlichen Effektstärken von 0,47, 0,10 und -0,01, die von Hattie gemittelt werden, obwohl dies die Autorin selbst nicht macht. Die beiden anderen in diesem Komplex zitierten Studien haben es eigentlich mit Betreuungs- und Förderprogrammen zu tun, gehören damit eigentlich zu einem anderen Thema, werden aber mit vereinnahmt. Man fragt sich, ob nicht die wichtigere Information darin steckt, dass Erwerbstätigkeit nicht zum Erfolg beiträgt, als dass alle drei (oder fünf) Meta-Analysen zusammen eine numerische Effektstärke von 0,17 aufweisen.

Einstellungen und Dispositionen (S. 54 und S. 381)

In der Gruppe „Einstellungen und Dispositionen“ zieht Hattie vier Meta-Analysen heran. Deren Effektstärken berichtet er als 0,07, 0,10, 0,06 und 0,54. Nach Hattie ergibt das eine mittlere Effektstärke von 0,19; damit landet dieser Komplex auf Rang 109. Die ersten drei Meta-Analysen, die ähnlich niedrige Effekte von d = 0,07 und 0,10 und 0,06 aufweisen, beziehen sich auf das Persönlichkeitsinventar Big Five (Neurotizismus, Extraversion, Offenheit für Erfahrungen, Verträglichkeit, Gewissenhaftigkeit); die vierte, die eine beträchtlich höhere Effektstärke von d = 0,54 berichtet, fasst jedoch Studien zum Glücklichsein zusammen. Hätte man diese Studie berechtigterweise aus dieser Gruppe herausgenommen, betrüge die Effektstärke zu dem Komplex Big Five 0,076 und würde statt auf Rang 109 etliche Ränge weiter unten landen, auf Rang 128.

Feedback (S. 206 – 211 und S. 412 f.)

Diese Gruppe umfasst 23 Studien. Wir beschränken uns auf die Zweifelsfälle. Alle 23 Meta-Analysen waren nicht zu beschaffen. Folgende Ungereimtheiten sind aufgefallen:

Während sich die Studien in diesem Komplex mit der Wirkung von Feedback auf die Lernenden befassen, behandelt die Analyse von Menges & Brinko die Beurteilung von Lehrenden durch Schüler und die Auswirkung auf Wiederholungen. Zwar ist die Evaluation von Lehrenden auch eine Rückmeldung, aber die Zielgruppe unterscheidet sich doch gravierend von den anderen Studien in der Gruppe. Es sind zwar in der Meta-Analyse von Menges & Brinko drei Studien enthalten, die einen Effekt auch auf den Lernerfolg der Schüler messen, aber diese hat Hattie nicht gemeint, denn er gibt alle 27 Studien als Bezug an: die gelistete Effektgröße ist die Größe aller enthaltenen Studien. Hätte er sich nur auf diesen Teil der Stichprobe bezogen, hätte er einen Effekt von Null vermelden müssen. Dieselbe Problematik ergibt sich mit der Meta-Analyse von L’Hommedieu, Menges & Brinko, die „feedback to teacher“ und Ratings von „instructor quality“ untersucht. Lehrerbeurteilung hätte man von Rückmeldung an Schüler trennen müssen. Die angegebene Effektstärke von d = 0,81 (r = 0,38) ist nirgends zu finden.

Eine zweite Meta-Analyse bereitet Kopfschmerzen: Standley (schon oben erwähnt) sammelt Studien, welche die Auswirkung von Musik auf verhaltenstherapeutische Interventio-

nen messen, sozusagen als Reinforcement von bestimmten Haltungen. Konkret geht es um soziales Verhalten, Körperbewegung, medizinische Studien (Kopfhaltung, unkontrollierte Tics, Saugmuster, Schreien von Kleinkindern bei Koliken), Sport, Kindesentwicklung, Transport (z. B. Unruhe in Schulbussen), aber auch Aufmerksamkeit im Mathematik-Unterricht, Lesen, Zahlen merken etc. Nur in einem sehr weiten Sinne hat diese Studie etwas mit Feedback zu tun; es handelt sich um behavioristisches Reinforcement, das Hattie nicht aus seinem Feedback-Begriff streichen will. Für Verhaltensmodifikation gibt es aber eine eigene Effektgruppe im Buch von Hattie. Die Studie hat auch nur in einem sehr weiten Sinne etwas mit Lernen zu tun, z. B. in drei Teilstudien zum Lernen in der Schule. Hattie hätte sich auf die Substichprobe der zwanzig Studien beschränken können, die es mit Lernen in der Schule zu tun haben. Dann hätte er in seinem Buch auf die zweithöchste Effektstärke von d = 2,87 verzichten müssen.

Eine weitere Ungenauigkeit in dieser Gruppe ist Hattie mit der Meta-Analyse von Azevedo & Bernard unterlaufen. Die Meta-Analyse umfasst 22 Studien mit einer Effektstärke von d = 0,80, wie bei Hattie angegeben. Aber Azevedo & Bernard teilen die Studien in eine Teilstichprobe mit 14 „immediate post-test studies“, deren Effektstärken zwischen 0,03 und 2,12 variieren, und eine zweite Teilstichprobe mit 8 „delayed post-test studies“, deren Effektstärken zwischen 0,15 und 0,62 liegen. Diese enorme Streubreite zeigt nochmals nachdrücklich, dass bereits in den einzelnen Meta-Analysen die Streuung durch Mittelwertbildung vernichtet wurde. Die von Hattie für die Azevedo-Studie angegebene Effektstärke 0,80 bezieht sich aber nur auf die erste Teilstichprobe der 14 „immediate post-test studies“. Die anderen acht Studien haben nur eine Effektstärke von 0,35. Aber Hattie gibt als Grunddaten an: 22 Studien, 22 Effekte und eine Effektstärke von d = 0,80.

Anzahl der Studien

Die Zählung der Studien, die Hattie im Anhang A vornimmt, hat zu der wiederholten Aussage geführt, er habe 815 Meta-Analysen verarbeitet. Es soll kein Argument gegen die umfangreiche Quellenarbeit von Hattie sein, aber es ist vielleicht dennoch erwähnenswert, dass 91 Studien mehrfach unter verschiedenen Faktoren aufgeführt werden, so dass die verarbeitete Zahl von Meta-Analysen 724 beträgt. Das kann man so oder so sehen, man hätte es vielleicht klären sollen. Neun Titel aus der Liste in Anhang A fehlen im Schriftenverzeichnis. Auch das ist kein Beinbruch.

Gravierender ist, dass viele der einbezogenen Meta-Analysen theoretisch anspruchslose „unpublished dissertations“ sind, deren Autoren als Doktoranden beweisen mussten, dass sie eine statistische Methode beherrschen, und die dafür eine Meta-Analyse gewählt haben. Das erinnert an die zahllosen Diplomarbeiten zur Sinnlose-Silben-Forschung in der Nachkriegszeit. Gerade in diesen Fällen wäre eine qualitative Analyse jeder einzelnen Meta-Analyse und Teilstudie wichtig gewesen. An dieser Stelle möchten wir anmerken, dass wir die Wertschätzung, die amerikanische PhD-Arbeiten in Europa genießen, nicht teilen. Gerade empirische PhD-Arbeiten sind meistens nur Routinestücke; die Gründlichkeit der Darstellung sowie der Theoriebezug lassen Einiges vermissen (wie früher schon bemerkt, s. Schulmeister, 2006, S. 58 ff.).

Die Problematik der Rangliste

Viel Aufmerksamkeit in der wissenschaftlichen und vor allem sonstigen Öffentlichkeit hat die von Hattie errechnete Rangliste der Effekte erfahren (s. Spiewak, 2013). Einige unserer kritischen Anmerkungen lassen aber bereits erkennen, dass die Rangliste ein reiner Show-

effekt ist, erst recht in der Illusion von Exaktheit, die sie vermittelt: Sobald man die Komplexe neu berechnet und dabei einzelne Studien herausnimmt, weil sie semantisch fehl am Platze oder statistisch problematisch sind, verschieben sich die Rangplätze. Ein Verzicht auf die Rangliste wäre der Wissenschaftlichkeit sicherlich besser bekommen. Allerdings hätte die Öffentlichkeit die Arbeit dann kaum zur Kenntnis genommen.

Literaturangaben

Brügelmann, H. (o. J.): Metaanalysen: Nutzen und Grenzen. (http://www2.agprim.uni-siegen.de/printbrue/hattie.09.metaanalyse.pdf).

Hattie, J. (2014): Lernen sichtbar machen. 2. korr. Aufl. Baltmannsweiler: Schneider Verlag Hohengehren.

Hattie, J. / Rogers, H. J. / Swaminathan, H. (2014): The role of meta-analysis in educational research. In: A. D. Reid et al. (Hrsg.) A Companion to Research in Education. Dordrecht: Springer, S. 197–207.

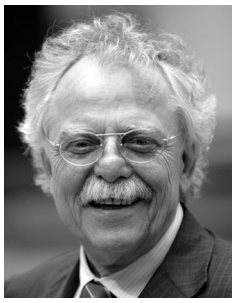
Lind, G. (2013). Meta-Analysen als Wegweiser? Zur Rezeption der Studie von Hattie in der Politik. (http://www.uni-konstanz.de/ag-moral/pdf/Lind-2013_meta-analysen-als-wegweiser.pdf).

Lind, G. (2014): Effektstärken: Statistische, praktische und theoretische Bedeutsamkeit empirischer Studien. (http://www.uni-konstanz.de/ag-moral/pdf/Lind-2014_Effektstaerke-Vortrag.pdf).

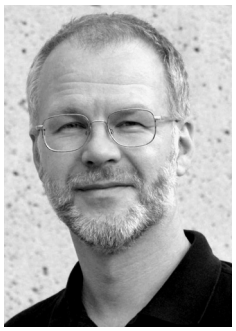
Schulmeister, R. (2006): eLearning: Einsichten und Aussichten. Oldenbourg: München.

Spiewak, M. (2013): Ich bin superwichtig! Kleine Klassen bringen nichts, offener Unterricht auch nicht. Entscheidend ist: Der Lehrer, die Lehrerin. Das sagt John Hattie. ZEIT Online 03.01.2013.

Terhart, E. (2011): Hat John Hattie tatsächlich den Heiligen Gral der Schul- und Unterrichtsforschung gefunden? Eine Auseinandersetzung mit Visible Learning. In: Keiner, E. et al. (Hrsg.). Metamorphosen der Bildung. Historie – Empirie – Theorie. Festschrift für Heinz-Elmar Tenorth. Bad Heilbrunn: Klinkhardt, S. 277–292.



Prof. em. Dr. Rolf Schulmeister
Universität Hamburg, Zentrum für Hochschul- und Weiterbildung
schulmeister@uni-hamburg.de



Prof. Dr. Jörn Loviscach
Fachhochschule Bielefeld, Lehrstuhl für Ingenieurmathematik und technische Informatik
joern.loviscach@fh-bielefeld.de