

## Metaanalysen: Nutzen und Grenzen

Die empirische Bildungsforschung - vor allem in den angelsächsischen Ländern - hat in den vergangenen 20, 30 Jahren eine Unzahl von Befunden aus Einzelprojekten zu verzeichnen, die kaum jemand mehr überblicken kann.

Angesichts dieser zunehmenden Fülle von Detailstudien hat (in der Sozialforschung generell) die Bedeutung sog. „Metaanalysen“ zugenommen. Um zu sehen, ob es eine Beziehung zwischen Faktoren wie z. B. Fernsehkonsum von Kindern und ihrer Lesefähigkeit gibt und wie hoch sie statistisch ausgeprägt ist, verrechnet man Daten, z.B. gefundene statistische Effekte. Man tut dies, obwohl die Daten unter verschiedenen Bedingungen, in verschiedenen Kontexten (Bildungssystemen, Sprachen, Kulturen...) und mit unterschiedlichen Instrumenten gemessen worden sind.

Auf diese Weise hofft man robustere Werte zu gewinnen und den Einfluss von Ausreißer-Befunden aus Einzelstudien zu minimieren. Auf den ersten Blick ein überzeugender Ansatz - und auf alle Fälle eine Hilfe, um einen groben Überblick zu gewinnen.

Vor kurzem nun ist ein Buch erschienen, das die *Times Educational Supplement* als „... heiligen Gral der Didaktik“ gewürdigt hat, weil es in einem Band zusammenfasse, was man heute aus empirischen Studien über Bedingungen für einen *erfolgreichen Unterricht* wisse:

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

Auch in Deutschland wird diese Zusammenschau immer häufiger zitiert, um bestimmte „Methoden“ oder „Ansätze“ als mehr oder weniger sinnvoll zu bewerten<sup>1</sup>.

Denn Hattie ist noch einen Schritt weitergegangen als die Kolleg/inn/en vor ihm: Er hat die Werte aus über 800 Metaanalysen zum Schulerfolg unter verschiedenen Bedingungen in einer Meta-Meta-Analyse erneut verdichtet und findet in der Tat einige beeindruckende

---

<sup>1</sup> Vgl. etwa Helmke in der ZEIT ( <http://www.zeit.de/2011/51/Interview-Helmke/seite-2> ) oder Köller unter: [http://www.lv-schleswig-holstein.mnu.de/index.php?option=com\\_rubberdoc&view=doc&id=31&format=raw](http://www.lv-schleswig-holstein.mnu.de/index.php?option=com_rubberdoc&view=doc&id=31&format=raw) ); s. für eine deutschsprachige Einführung:

[http://www.afl.hessen.de/inj/servlet/prt/portal/prtroot/slimp.CMReader/HKM\\_15/AfL\\_Internet/med/6e9/6e940673-173f-1131-f012-f31e2389e481,22222222-2222-2222-2222-222222222222](http://www.afl.hessen.de/inj/servlet/prt/portal/prtroot/slimp.CMReader/HKM_15/AfL_Internet/med/6e9/6e940673-173f-1131-f012-f31e2389e481,22222222-2222-2222-2222-222222222222) S. 25-27

Zusammenhänge. Hier die Plätze 1-10 der angeblich einflussreichsten Faktoren<sup>2</sup> (s. für einen breiteren Überblick unten die Auswahl-Übersicht auf Deutsch):

Rank	Influence	Studies	Effects	ES
1	Self-reported grades	209	305	1.44
2	Piagetian programs	51	65	1.28
3	Providing formative evaluation	30	78	.90
4	Micro teaching	402	439	.88
5	Acceleration	37	24	.88
6	Classroom behavioral	160	942	.80
7	Comprehensive interventions for learning disabled students	343	2654	.77
8	Teacher clarity	na	na	.75
9	Reciprocal teaching	38	53	.74
10	Feedback	1287	2050	.73

Die formal aufgelisteten Merkmale zeigen allerdings Spannungen auf, die den Sinn von Ranglisten, wie Hattie sie publiziert, in Frage stellen. Zum Beispiel wirke positiv, wenn Lehrer/innen klare Anweisungen bei der Aufgabenstellung geben, aber auch wenn Schüler/innen anderen Schüler/innen etwas beibringen oder wenn sie sich gar selbst beurteilen dürfen. Auch die folgende Auswahl über das ganze Spektrum der Effektstärken zeigt, dass es darauf ankommt, was man jeweils unter den einzelnen Etiketten konkret versteht - sowohl in den verschiedenen Studien, die als „gleich“ miteinander verrechnet werden, als auch bei der Übersetzung in andere Sprachen (s. die Originalbegriffe in Klammern) als auch in jeder Auslegung der stark verknüpften „Etiketten“ durch einzelne Leser/innen<sup>3</sup>:

- 1.4 Selbstbeurteilung
- 1.3 Piaget-orientierte Programme
- .9 förderorientierte Beurteilungen
- .9 Einüben von methodischen Fertigkeiten („micro teaching“)

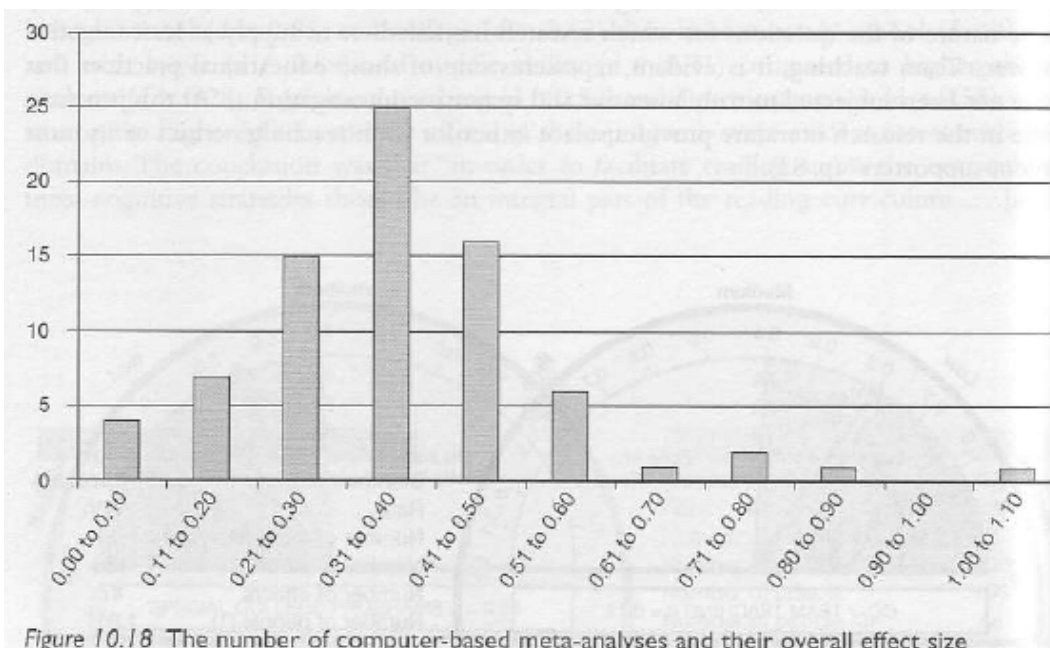
<sup>2</sup> ES = Effektstärke; ab 0,5 gilt ein Effekt gemeinhin als beachtenswert - der Anschaulichkeit halber oft gleichgesetzt mit dem Vorsprung von etwa einem Schuljahr):

<sup>3</sup> Vgl. zu der sehr unterschiedlichen Definition von „offenem Unterricht“ in entsprechenden Studien: Brügelmann, H. (1998): Öffnung des Unterrichts - Befunde und Probleme der empirischen Forschung. In: Brügelmann, H./ Fölling Albers, M./ Richter, S. (Hrsg.) (1998): Jahrbuch Grundschule. Fragen der Praxis Befunde der Forschung. Erhard Friedrich Verlag: Seelze, 8-42.

- .7 klare Anweisungen des Lehrers („teacher clarity“)
- .7 Schüler als Lehrer („reciprocal teaching“)
- .6 Fortbildung von Lehrer/inne/n
- .5 Schüler als Tutor\_inn\_en („peer tutoring“)
- .0 Schülerkontrolle über Lernen
- .0 offen vs. traditionell
- .0 Altersmischung
- .0 Fachwissen Lehrer
- .0 Förderung von Wahrnehmung und Motorik

Wie kompliziert das Bild tatsächlich ist, zeigen beispielhaft die Werte zum Faktor „Computereinsatz“ im Unterricht. Mit einer durchschnittlichen Effektstärke von .39 ist sein Einfluss als mäßig positiv einzuschätzen. Wenn die Lernschritte durch das Programm vorgegeben werden, verschwindet dieser sogar ganz (-.02). Andererseits steigt der Effekt, wenn die Schüler das Vorgehen im Programm selbst steuern konnten, auf .49, und wenn sie mit einem Partner arbeiteten, sogar auf .96.

Das heißt aber, wenn man sich diesen Überblick über die verschiedenen Studien anschaut,



dass nicht der Mittelwert das eigentlich interessante Datum ist (noch einmal: wir sehen hier eine Verteilung der Häufigkeiten mittlerer Effektstärken aus verschiedenen Metaanalysen), sondern eher die Extremwerte und damit die Streuung. Diese verweisen darauf, dass es auf die *Art* des Computereinsatzes und auf die *Bedingungen* seiner Nutzung ankommt, ob er didaktisch produktiv wird oder nicht. Was aber nützt dann der

Lehrperson das Wissen um einen Mittelwert von 0.39 wenn sie eine Form des Computereinsatzes wählt, die einen Wert von 0.0 erzielt?

Diese *Konzept-* und *Kontext*abhängigkeit von Wirkungen einer „Methode“ zeigt sich auch bei anderen Interventionen. So hat nach Hattie ein auf Problem-Aufgaben fokussierter Unterricht mit einer Effektstärke von .15 insgesamt nur einen geringen positiven Effekt - mit .57 aber eine hohe Wirksamkeit, wenn vorher Beispielaufgaben durchgearbeitet worden sind. Dabei kommt es zusätzlich auf die angezielte Kompetenzebene an: für die Sicherung von Prinzipien steigt der Effekt bis zu auf .75, für den Erwerb von Basiswissen ist die Beziehung dagegen sogar negativ, d. h. ein traditioneller Unterricht ist in dieser Hinsicht erfolgreicher.

Für die Einschätzung des Erfolgs einer Methode kommt es also erstens darauf an, an welchem *Kriterium* man diesen misst. Dann ist zu prüfen, in welcher *Form* unter welchen *Bedingungen* sie umgesetzt wird. Die für einen Überblick über die Forschung und für bildungspolitische Entscheidungen verdienstvolle Verdichtung der Daten in Hatties Meta-Meta-Analyse bedeutet also, dass diese genau durch die Ablösung von den konkreten Kontexten für Praxis weitgehend ihre Bedeutung verlieren. Denn die Zahlen sind mehrdeutig. Was bedeutet zum Beispiel eine mittlere Korrelation oder Effektstärke im konkreten Fall<sup>4</sup>: leicht positiver Einfluss in allen Fällen ODER starker in einigen und gar keiner in anderen ODER hoher im Zusammenwirken mit Faktor X, aber niedriger im Zusammenwirken mit Variable Y?

Das heißt aber: Selbst wenn 99 von 100 Studien eine Überlegenheit der Methode B zeigen, dürfte das kein Grund sein, einer Lehrerin, die mit Methode A erfolgreich arbeitet, deren Nutzung zu untersagen (vgl. wieder das Beispiel der Computernutzung: zwei Ausreißerstudien weisen besonders hohe Effekte auf). Es wäre auch kein Grund, weiteren Lehrer/inne/n, die Methode A ausprobieren wollen, dies zu verbieten. Allenfalls könnte man eine besondere Begründung verlangen, warum jemand von einer weithin bewährten Praxis abweichen will. Juristisch gesprochen, erzeugen statistische Befunde aus großen

---

<sup>4</sup> Ganz abgesehen von dem häufig übersehenen Problem, dass ein gleichzeitiges Auftreten von zwei Faktoren nicht gleichbedeutend ist mit einer inhaltlichen Beziehung: *Korrelation* bedeutet nicht *Kausalität*. Ein hoher TV-Konsum geht häufig einher mit schwacher Leseleistung - aber ist sie die Ursache? Oder kann nicht auch eine schwache Lesekompetenz verhindern, dass Kinder zum Buch greifen (und damit ihre Lesefähigkeit verbessern) und stattdessen lieber fern sehen? Oder ist gar ein Drittfaktor (z. B. Vernachlässigung durch die Eltern) möglicherweise Ursache für beide Verhaltensweisen?

Stichproben dann eine „Beweislast“, d. h. eine Begründungspflicht für Abweichungen - sie können diese aber nicht abwerten oder gar ausschließen.

Kontexte verändern Beziehungen - das gilt selbst in den Naturwissenschaften, wenn wir an die Gesetze der Mechanik im Mikrobereich der Atome denken. In den Sozialwissenschaften wiederum lassen sich Befunde aus kontrollierten Experimenten nicht einfach auf Alltagssituationen übertragen, Befunde aus einem (schul)kulturellen Kontext nicht unbesehen auf einen anderen.

Insofern gerät die vertraute Grundfigur der Forschung und ihres Verhältnisses zur Praxis ins Wanken: zunächst Ableitung allgemeiner Regeln aus der Häufigkeit von Beziehungen bei einer großen Zahl von Fällen und dann deduktive Ableitung einer Handlungsanweisung für den Einzelfall aus diesen Regeln. Für Lehrer/innen bedeutet das: Jede Situation ist ein neuer Fall und bedarf einer eigenen Einschätzung. Die Durchschnittsbefunde aus (Meta-)Metanalysen können dafür hilfreiche Hypothesen liefern - aber keine Vorschriften<sup>5</sup>.

---

<sup>5</sup> Vgl. ausführlicher zum Status statistischer Befunde und zu den Anforderungen an eine für die Praxis nützliche Forschung: Brügelmann, H. (2011): Miss Marple neben PISA & Co. - Plädoyer für eine Bildungsforschung, die der Praxis nützt. In: Moser, H. (Hrsg. (2011): Aus der Empirie lernen? Forschung in der Lehrerbildung. Professionswissen für Lehrerinnen und Lehrer, Bd. 10. Schneider Verlag Hohengehren: Baltmannsweiler, 221-234.