

Brügelmann, Hans

Evidenzbasierung in der Pädagogik. Scheinlösung für ein schwieriges Problem

Pädagogik (Weinheim) 67 (2015) 10, S. 38-43



Quellenangabe/ Reference:

Brügelmann, Hans: Evidenzbasierung in der Pädagogik. Scheinlösung für ein schwieriges Problem - In: Pädagogik (Weinheim) 67 (2015) 10, S. 38-43 - URN: urn:nbn:de:0111-pedocs-180804 - DOI: 10.25656/01:18080

<https://nbn-resolving.org/urn:nbn:de:0111-pedocs-180804>

<https://doi.org/10.25656/01:18080>

Nutzungsbedingungen

Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen: Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

We grant a non-exclusive, non-transferable, individual and limited right to using this document.

This document is solely intended for your personal, non-commercial use. Use of this document does not include any transfer of property rights and it is conditional to the following limitations: All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

peDOCS
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation
Informationszentrum (IZ) Bildung
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Digitalisiert

Mitglied der


Leibniz-Gemeinschaft

Evidenzbasierung der Pädagogik

Scheinlösung für ein schwieriges Problem¹

HANS BRÜGELMANN

Schon früh sind die großen Testprogramme wie PISA, IGLU und VerA usw. mit dem Hinweis kritisiert worden, vom vielen Wiegen werde die Sau nicht fett, d. h. die Leistungen der Schule bzw. ihrer Schüler(innen) nicht besser. Diese Kritik trifft zu, greift aber zu kurz. Sie stellt zum Ersten nicht in Frage, ob die Waage (sprich: der jeweils eingesetzte Leistungstest) überhaupt misst, was gemessen werden soll, und sie akzeptiert zum Zweiten, dass eine messbare Gewichtszunahme (sprich: Verbesserung der Testwerte) ein angemessener Maßstab für die angestrebten Lernerfolge sei.

Mit dem Anspruch der »Evidenzbasierung« ihrer Praxis ist die Pädagogik in den letzten Jahren unter Druck geraten.

Beide Annahmen sind problematisch – sei es auf der Ebene des Bildungssystems (PISA & Co), sei es auf der Ebene der Schule oder auf der Ebene der einzelnen Schüler(innen) (VerA). Nicht, weil Leistungstests nutzlos wären. Im Gegenteil: Sie sind hilfreiche Instrumente, solange man ihren Status nicht überschätzt und den Geltungsbereich ihrer Ergebnisse beachtet. Überzogen wird ihr Kredit, wenn man ihre Befunde im Vergleich zur Alltagserfahrung und zum professionellen Urteil als höherwer-

Die Zunahme von empirischen Studien zum Unterricht hat zu der Forderung geführt, dass pädagogisches Handeln sich an Befunden zur Wirkung des Handelns orientieren sollte. Inwieweit ist dieser Anspruch angemessen? In welchem Verhältnis stehen Befunde empirischer Studien zu komplexen pädagogischen Handlungsanforderungen? Eine kritische Sichtung der Potentiale und Grenzen von Tests für die pädagogische Praxis.

tig betrachtet – und nicht als komplementär, d. h. als wichtige Ergänzung.

Mit dem Anspruch einer stärkeren »Evidenzbasierung« ihrer Praxis ist die Pädagogik in den letzten Jahren unter Druck geraten (vgl. die Beiträge in *Böttcher u. a.* 2009). Die Forderung: In der Schule dürften nur Methoden eingesetzt werden, die sich in empirischen Untersuchungen als »wirksam« erwiesen haben. Der »Goldstandard« für solche Studien verlangt nicht nur eine bestimmte Anlage der Untersuchung (Kontrollgruppen-Design mit großen Zufallsstichproben), er setzt auch auf bestimmte Verfahren der Erhebung und Auswertung von Daten. Beide Forderungen sind problematisch.

Die Durchschnittsfalle: Menschliches Verhalten ist kontextabhängig

Die Reichweite von Befunden der Bildungsforschung ist nicht nur zeitgebunden, sondern auch situationsabhängig. Für unterschiedliche nationale Kulturen mag dies unmittelbar einleuchten. Kontextabhängigkeit gilt aber auch für Subkulturen wie das soziale Milieu der Familie oder einzelner Klassenzimmer. Was Lehrerin A gelingt, mag bei Lehrer B schiefgehen – selbst bei gleicher Kompetenz. Trotz Super Nanny und ihrer »Erfolge«: Unterrichtsmethoden und Didaktik, erst recht aber pädagogische Konzepte sind keine Techniken, die situationsübergreifend funktionieren. So setzen Lehrer(innen) pädago-

gische Programme ganz unterschiedlich um. Wer in verschiedene Klassen hineinschaut, die mit demselben Schulbuch, selbst mit demselben, in kleinen Schritten ausgearbeiteten Lehrgang arbeiten, wird einen ganz unterschiedlichen Unterricht erleben. Umgekehrt nutzen Lehrer(innen) auch spezifische Methoden im Rahmen ihrer jeweiligen didaktischen Konzeption auf verschiedene Weise: Freies Schreiben mit einer Anlauttabelle hat in Reichens Konzeption »Lesen durch Schreiben« eine ganz andere Bedeutung als im Vier-Säulen-Modell des Spracherfahrungsansatzes oder als Anhängsel eines Fibellehrgangs. Diese Konzept- und Kontextabhängigkeit der Umsetzung beeinflusst nicht nur die reale Intervention, sondern auch deren Wirkungen. Anders gesagt: Die Lehrperson hat eine Schlüsselrolle bei der Übersetzung des Papier-Curriculums in ein Curriculum-in-Aktion. Es stellt sich damit die Frage nach dem Alltagswert von Studien, die versuchen, den Einfluss der Randbedingungen möglichst weitgehend auszuklammern.

Für deren Anwendung auf den Einzelfall stellt sich ein zusätzliches Problem: Befunde aus Stichproben sind immer Wahrscheinlichkeitsaussagen. Folgt man ihnen, hat man die Sicherheit, in hundert oder gar tausend Fällen weniger Fehler zu machen, als wenn man sie nicht beachtet. Ob sie aber auf den Einzelfall zutreffen, kann man vorher nicht sagen. In den Mittelwerten der Stichprobenergebnisse werden die – in der Regel breit

streuenden – Einzelwerte gemittelt. Die einzelne Lehrerin aber hat es mit einem konkreten Kind zu tun, und sie muss entscheiden, ob ihre Jule oder ihr Ben ein im Sinne der Statistik »typisches« Kind ist oder nicht. Als Vereinfachungen erleichtern solche Typisierungen das Alltagshandeln unter Zeitdruck – mit dem Risiko von Fehlentscheidungen im Einzelfall. Unter welchen Bedingungen eine Allgemeinaussage auf einen konkreten Schüler zutrifft – z. B. bei der Übergangsentscheidung am Ende der 4. Klasse – das sagen uns Studien wie IGLU oder VerA nicht.

Zu Recht wird kritisiert, wer glaubt, aus Einzelbeobachtungen generalisierbares Wissen ableiten zu können. Aber weniger beachtet wird der umgekehrte, mindestens gleich gewichtige Irrtum: aus statistischem Wissen könne man Fallwissen für praktisches Handeln im Alltag direkt ableiten. Denn für Pädagogen besteht das eigentliche Problem darin, allgemeine Aussagen – seien es empirische Befunde der Forschung oder normative Vorgaben wie Lehrpläne – auf ihre jeweilige besondere Situation hin auszulegen.

Stellen Sie sich analog zum Vergleich von zwei Unterrichtsmethoden eine Untersuchung von Medizinern bzw. Optikern vor, die die Sehfähigkeit von Grundschulkindern mit zwei Brillen vergleichen, eine mit Gläsern +2 Dioptrien und eine mit Gläsern -2. Da in dem Alter Kurzsichtigkeit häufiger ist, würde im Durchschnitt die -2-Brille besser abschneiden – und zwar statistisch signifikant, wenn die Stichproben nicht zu klein sind. Darf man auf Grund dieses Befundes allen Kindern dieselbe Brille verordnen? Und selbst wenn die Studie nach Kurzsichtigen (analog zu »Kinder ohne Migrationshintergrund«) und Weitsichtigen (analog zu »Kinder mit Migrationshintergrund«) unterschiede, würde doch niemand akzeptieren, dass aufgrund der unterschiedlichen Effekte ersterer eine -2-Brille und letzterer eine +2-Brille verschrieben würde. Denn wir wissen um die große Streuung der Fehlpassung selbst innerhalb der Gruppen Kurz- und Weitsichtigkeit, dass die Fehlsichtigkeit ja von 0,5 bis 6 und mehr variieren kann und dass es zudem noch weitere Eigenschaften des Auges gibt, die über die Passung ei-

ner Brille entscheiden. Von einer Augenärztin oder einem Optiker erwarten wir deshalb, dass sie die Gläser für die einzelnen Patient(inn)en auswählen und gemeinsam mit ihnen ausprobieren, welche am besten passen. Denn über die bloße Dioptrienzahl hinaus sind weitere Aspekte – bis hin zum subjektiven Sehgefühl – zu berücksichtigen. In der Didaktik aber wird eine »evidenzbasierte Forschung« als »Goldstandard« akzeptiert, die aufgrund von Mittelwertvergleichen Methoden generell vorschreibt oder ausschließt, obwohl deren Wirkungen je nach Kontext breit streuen. Da ist der Erfahrungstransfer über Analogie (»von Fall zu Fall«) nicht fehleranfälliger – dank seines Kontextreichtums sogar leichter auf Passung einzuschätzen als abstrakte Kennwerte.

Hinzu kommt: Die Bedeutung menschlichen Verhaltens wechselt, sie ist kultur-, ja sogar biografieabhängig. Diese Qualitäten werden in den Verallgemeinerungen der Großstudien neutralisiert – oder durch die Standardisierung der Erhebungen gar nicht erst erfasst. Damit erweist sich ihre besondere Stärke zugleich als grundlegende Schwäche.

Die Objektivitätsillusion: Menschliches Verhalten ist mehrdeutig

Auf den ersten Blick lassen die vielen empirischen Befunde zur Fehleranfälligkeit des Lehrerurteils bei der Bewertung von Schülerleistungen es attraktiv erscheinen, an seine Stelle standardisierte Tests zu setzen. Denn immer wieder haben Studien zur Vergabe von Noten bestätigt, was Alltagserfahrung von Schüler(inne)n ist:

- Verschiedene Lehrer(innen) beurteilen dieselbe Arbeit – selbst in Mathematik – auf der üblichen Sechser-Skala mit bis zu fünf unterschiedlichen Noten (mangelnde Objektivität).
- Eine Arbeit wird nach wenigen Monaten nur von einer Minderheit der Lehrer(innen) mit derselben Note bewertet wie beim ersten Mal (mangelnde Reliabilität).
- Noten weichen nicht zufällig von Testergebnissen ab, sondern sind je nach Sympathie oder Antipathie systematisch verzerrt (mangelnde Validität).

- Beim Wechsel in eine andere Klasse mit einem anderen Leistungsniveau kann sich die Bewertung derselben Schülerin um mehrere Notenstufen verbessern oder verschlechtern (mangelnde Vergleichbarkeit).

Kein Wunder also, dass man nach Alternativen sucht. Die Hoffnung heißt: standardisierte Tests. Sowohl die Erhebung als auch die Auswertung von Daten wird vorweg festgelegt (»operationalisiert«), um sie von den beteiligten Personen unabhängig zu machen. Autor(inn)en eines Leistungstests sollen Aufgaben demnach so gestalten, dass sie von allen getesteten Personen gleich verstanden und dass die Lösungen von verschiedenen Beurteilern gleich klassifiziert bzw. bewertet werden.

Menschliches Verhalten ist bedingt durch den kulturellen und situativen Kontext, seine (Be-)Deutung zudem abhängig von Erwartungen und Annahmen.

Die grundlegende Annahme des Standardisierungsparadigmas: Objektiv definiertes, beobachtbares Verhalten lässt sich eindeutig bestimmten Tiefenstrukturen (Kompetenzen bzw. Intentionen) zuordnen. Diese Annahme ist brüchig. Menschliches Verhalten ist immer mehrdeutig – bedingt durch den kulturellen und situativen Kontext und abhängig von Erwartungen und Deutungen der Beteiligten.

Dass jemand »die Hand hebt«, werden verschiedene Beobachter im Regelfall übereinstimmend feststellen. Wie sie dieses Verhalten deuten, das heißt, welche Absicht oder Aussage sie der beobachteten Person unterstellen, hängt aber vom Kontext ab:

- »Hier bin ich, komm her!« (auf dem Bahnsteig)
- »Ich stimme der Vorlage zu« (bei einem Wahlgang)
- »Bin bereit, zu dem angebotenen Preis zu kaufen« (bei einer Versteigerung)
- »Ich weiß die Antwort« (in einer Schulklasse)

Am Beispiel des Meldens im Unterricht lässt sich die Problematik erweitern. Auch wenn der Kontext klar ist, können Intention einerseits und Deutung andererseits immer noch ausein-

anderfallen. Manche Schüler(innen) melden sich, obwohl sie die Antwort nicht wissen – in der Hoffnung, dass die Lehrperson ihre Meldung im üblichen Sinne deutet und sie deshalb nicht aufruft. Es soll sogar Lehramtsanwärter(innen) geben, die ihre Schüler(innen) vor einer Lehrprobe instruieren, mit dem rechten Arm aufzuzeigen, wenn sie die Antwort wissen, und mit dem linken, wenn nicht. So können sie den uneingeweihten Prüfer(inne)n eine »breite aktive Mitarbeit« vortäuschen.

Nun lässt sich einwenden, dass diese Beispiele allenfalls für Befragungen relevant sind, etwa weil Personen dazu neigen, sozial erwünschte (genauer: von ihnen individuell als sozial erwünscht eingeschätzte) Antworten anzukreuzen. Aber bei der Bewertung von Leistungen stellt sich das Problem der mehrdeutigen Beziehung zwischen Oberflächen- und Tiefenstruktur analog. Insofern war es kühn von einer deutschen Kultusministerin, Eltern zu versprechen:

Bei Aufgabenlösungen geht es nicht um »falsch« vs. »richtig«, sondern um das kognitive Niveau der zugrunde liegenden Kompetenzen.

»Jetzt wissen Sie wirklich, wo Ihr Kind steht!«, als die Vergleichsarbeiten VerA in Klasse 3 flächendeckend eingeführt wurden. Denn die richtige Antwort zu einer Aufgabe kann aus verschiedenen Gründen gewählt worden sein, z. B.,

- weil Schüler(innen) wie erwartet das von den Testautoren avisierte Wissen und Können zur aktuellen Problemlösung genutzt haben, aber auch;
 - weil sie diese oder eine ähnliche Aufgabe vor Kurzem bearbeitet und die verfügbare Lösung nur abgerufen oder analog entwickelt haben;
 - weil sie dank ihrer allgemeinen Intelligenz durch eigenes Nachdenken die Lösung gefunden haben;
 - weil sie unter Zeitdruck rasch noch ein Kreuz (zufällig richtig) gesetzt haben;
 - weil sie die Lösung aus Hinweisen im Text und den verfügbaren Alternativen als plausibel erraten haben.
- Umgekehrt gilt auch für Fehler, dass sie unterschiedliche Gründe haben

können: Unaufmerksamkeit, anderer Lösungsweg, ... Vor allem aber gibt es bessere und schlechtere Fehler – und falsche Ergebnisse, die auf ein höheres Kompetenzniveau verweisen als eine richtige Lösung aus falschem Grund.

Ein Beispiel aus dem Anfangsunterricht

Fünf Kinder haben geschrieben:

KINO		
	KINNO	
KINO		KIENO
	KINO	

Bei der Auswertung eines standardisierten Tests würde herauskommen: Zwei Kinder haben das Wort falsch geschrieben und drei richtig – eine objektive und reliable, das heißt von beliebigen Testern reproduzierbare Aus- und Bewertung. Aber ist sie auch valide?

Unabhängig von Person und Situation ist nur eine Oberflächendeutung möglich. Um die Kompetenz der Kinder beurteilen zu können, muss man aber wissen, wie sie sonst schreiben und wie sie vorher geschrieben haben, das heißt auf welcher Stufe der Schreibentwicklung sie sich befinden. Aufgrund einer solchen – theoriegeleitet begleitenden – Lernbeobachtung könnten sich ganz andere Einschätzungen ergeben:

Tom, das erste Kind, schreibt KINO (richtig), weil er in einer Straße wohnt, in der er beim täglichen Gang zum Kindergarten an einem Kino vorbeikommt, auf dessen Schild ihn seine Mutter mehrfach hingewiesen hat. Auf ähnliche Weise hat er ein Dutzend Wörter gelernt, die er als Buchstabenfolge reproduzieren kann, ohne zu wissen, warum man z. B. KINO gerade mit diesen Buchstaben schreibt.

Nora schreibt mithilfe einer Anlauttabelle (HUNT, RATFARA usw.). Sie hat das Wort nach ihrer Aussprache verschriftet – und Glück gehabt, dass diese Strategie im konkreten Fall zum richtigen Ergebnis führte. Sie schreibt also aus falschem Grunde richtig.

Carl ist schon einen Schritt weiter als Nora, weil er weiß, dass es nicht reicht, Laut für Laut zu verschriften.

Man muss beispielsweise aufpassen, ob ein Selbstlaut lang oder kurz gesprochen wird, und dass dann »da etwas Besonderes ist«. Aber ob man einen Buchstaben verdoppelt – und gegebenenfalls welchen –, das hat er noch nicht verstanden. Carl schreibt also aus halbrichtigem Grund falsch.

Jule weiß dagegen schon: Wenn man ein langes /i:/ hört und das Wort nicht kennt, dann ist es fast immer richtig und deshalb klug, <ie> zu schreiben. Diese statistisch begründete Faustregel hilft ihr Fehler zu vermeiden. Sie schreibt also aus richtigem Grund falsch.

Leonie kennt diese Faustregel auch. Aber sie hat darüber hinaus gelernt, dass es Ausnahmen gibt wie MASCHINE, APFELSINE und eben KINO, die man sich als Einzelwörter merken muss. Nur sie schreibt aus richtigem Grund richtig.

Bei anspruchsvolleren Aufgaben nehmen die Interpretationsspielräume zu. Kinder können schon die Aufgabe unterschiedlich verstehen – und die Auswertung kann leicht verketten, was sich hinter der individuellen Lösung verbirgt. Es geht nicht um »falsch« vs. »richtig« auf der Leistungsebene, sondern um das Niveau der zugrunde liegenden Kompetenzen. Diese Probleme verschärfen sich im musisch-ästhetischen Bereich, bei sozialen und politischen Kompetenzen – generell: je höher die Anforderungen an das individuelle Denk- und Urteilsvermögen steigen, also je mehr es um Bildung und nicht nur um Leistung geht. Leider werden diese beiden Dimensionen nicht immer sauber getrennt – z. B. wenn die Kultusministerkonferenz ihre Kompetenzvorgaben »Bildungsstandards« nennt, die eigentlich »Leistungsstandards« heißen müssten, da sie sich nur auf Fachleistungen beziehen. Die Qualität grundlegender schriftsprachlicher und mathematischer Kompetenzen hängt zudem davon ab, wie, das heißt über welche konkreten Erfahrungen sie erworben wurden, in welche persönliche Gesamtkompetenz sie eingebunden sind.

Leistung ist abhängig von der persönlichen Bedeutung der Aufgabe

Was hier für die Überprüfung des Lernerfolgs beschrieben wurde, spielt

auch eine Rolle für das Lernen selbst. Sogar für eine scheinbar nur technische Leistung wie die Rechtschreibung lässt sich die Bedeutung der persönlichen Bedeutung (hier: von Wörtern) zeigen. In entsprechenden Untersuchungen schneiden Jungen erwartungsgemäß im Durchschnitt schlechter ab als Mädchen. In der Regel nimmt der Rückstand der Jungen jedoch ab, wenn die Wörter aus ihrem Erfahrungs- und Interessensbereich stammten, z. B. bei »Schiedsrichter«, »Computer« oder »Fahrrad«. Das letzte Wort ist besonders interessant, weil in einer Schweizer Teilstichprobe kein Aufholeffekt der Jungen festzustellen war. Diese scheinbare Irritation ist aber gerade ein Beleg für das angenommene Muster, ist in der Deutschschweiz doch das Wort »Velo« üblich. Was hier als ein geschlechtsgruppenspezifischer Effekt sichtbar geworden ist, zeigt wie durch eine Lupe ein grundsätzliches Problem auf: die unterschiedliche Bedeutung derselben Aufgabe für das Lernen wie auch für dessen Überprüfung.

Befunde sind interpretationsbedürftig

Diese Mehrdeutigkeit von Aufgaben und ihrer Lösung bezieht sich zunächst auf die unterste Ebene einer Begriffspyramide, die im Forschungsprozess dann immer abstrakter und damit noch interpretationsbedürftiger wird. Die unterschiedlichen Auslegungen »derselben« Unterrichtsmethode durch verschiedene Lehrer(innen) habe ich bereits angesprochen. In Untersuchungen werden sie dennoch ein- und derselben Intervention zugeordnet. In einer Meta-Analyse wiederum werden verschiedene Studien zusammengefasst, die dieselbe Methode, z. B. »entdeckendes Lernen«, untersucht haben. Unberücksichtigt bleibt dabei, dass auch Forscher(innen) unterschiedliche Vorstellungen von diesem Konzept haben. Noch diffuser wird ein solcher Begriff in einer Meta-Meta-Analyse, wie Hattie (2013) sie vorgelegt hat.² So entstehen Sammelbegriffe wie »computerunterstützter

Unterricht«, die in sich ähnlich heterogen sind wie auf der anderen Seite »Gruppenarbeit« oder »offener Unterricht«. Damit verschärft sich das Problem der Interpretationsbedürftigkeit von Ergebnissen – und die Frage, wer die Deutungsmacht hat bzw. welche Möglichkeiten für eine unabhängige Urteilsbildung eröffnet werden.

Angemessen sind weder eine Expertokratie noch eine Bürokratie; notwendig sind – im Sinne einer demokratiédienlichen Forschung – Transparenz, Machtverteilung und Dialog. Wo die technische Präzisierung von Methoden an ihre Grenzen stößt, sind Mechanismen sozialer Kontrolle erforderlich – sowohl bei der Unterrichts- und Schulentwicklung als auch bei der Leistungsrückmeldung auf Individualebene.³ Dialogische Verfahren bieten keine Sicherheit für die Richtigkeit von Deutungen, aber Teilsichten können infrage gestellt, Autoritätsansprüche gebrochen werden. Ein Grundproblem der Sozialwissenschaften: Keine Methode kann gleichzeitig technische Genauigkeit/Objektivität und kulturelle Bedeutung/Gültigkeit von Beobachtungen sichern.

Diese Relativierung des Geltungsanspruchs von standardisierten Instrumenten ist keine Aufforderung, Tests wegen der genannten Schwächen wegzuzwerfen. Tests können ein hilfreiches Instrument sein, wenn man sie als eine zusätzliche (nicht höhere) Sicht auf Schülerleistungen nutzt. Schon aus ökonomischen Gründen macht es Sinn, von Zeit zu Zeit die ganze Klasse mit einem Kurzverfahren zu testen, um bei Abweichungen von der eigenen Einschätzung genauer und differenzierter zu beobachten. Aber Lehrer(innen) und Eltern sollten sich ihren Ergebnissen auch nicht unterwerfen. Tests können reflektierte Erfahrung und Urteilskraft nicht ersetzen.

Fazit

Befunde in den Humanwissenschaften sind in hohem Maß situationsabhängig. Die Bedeutung eines Faktors

verändert sich von Fall zu Fall mit dem Einfluss anderer Faktoren.

Auch Untersuchungen mit standardisierten Verfahren sind angewiesen auf die Interpretationsfähigkeit ihrer Autoren. Zahlen sprechen nicht für sich.

Die verallgemeinernde Absicherung von Befunden sichert keine Übertragbarkeit auf neue Situationen oder Einzelfälle. Nicht das arithmetische Mittel, sondern die Streuung der Einzelwerte um den Durchschnitt ist häufig das praxisrelevantere Datum.

Lehrer(innen) können Befunde aus Großstudien als Hypothesen nutzen, um ihre Aufmerksamkeit auf möglicherweise bedeutsame Bedingungen zu richten. Ob eher das Potenzial oder die Risiken dieser Faktoren zum Tragen kommen, muss in jedem Einzelfall neu überprüft werden.

Anmerkungen

¹ Dieser Beitrag ist die überarbeitete Fassung von Kap. 2.3 und 2.4 meines gerade erschienenen Buches »Vermessene Schulen – standardisierte Schüler« (Beltz: Weinheim 2015), in dem auch die zugrundeliegende Literatur detaillierter nachgewiesen ist.

Tests können ein hilfreiches Instrument sein, wenn man sie als eine zusätzliche (aber nicht höhere) Sicht auf Schülerleistungen nutzt.

² Siehe zu der sehr eingeschränkten Aussagekraft solcher (Meta-)Meta-Analysen Kap. 3.1 in »Vermessene Schulen«.

³ Siehe dazu Kap. 5.3 und 6.3 in »Vermessene Schulen«.

Literatur

Böttcher, W., u. a. (Hg.) (2009): Evidenzbasierte Bildung: Wirkungsevaluation in Bildungspolitik und pädagogischer Praxis. Münster

Brügelmann, H. (2015): Vermessene Schulen – standardisierte Schüler. Zu Risiken und Nebenwirkungen von PISA, Hattie, VerA & Co. Weinheim/Basel

Hattie, J. A. C. (2013): Lernen sichtbar machen. Baltmannsweiler (engl. 2009)

Dr. Hans Brügelmann (Prof. i. R.) ist Fachreferent im Grundschulverband.

Adresse: Admiralstraße 14, 28215 Bremen

E-Mail: hans.bruegelmann@grundschulverband.de