Leseprobe aus: Brügelmann, Vermessene Schulen - standardisierte Schüler, ISBN 978-3-407-25729-1

http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-407-25729-1

Einladung

Titel und Botschaft dieses Buches sind kein Urteil, sondern eine Warnung. Als ein Instrument unter anderen können standardisierte Tests und Fragebögen auch in der Bildungsforschung hilfreich sein – wenn ihr Kredit bei der Deutung der Ergebnisse nicht überzogen wird. Sie gefährden jedoch die pädagogische Arbeit, wenn sie zur Norm dafür werden, wie der Erfolg von Lehren und Lernen zu erfassen ist.

In meinen gut 40 Jahren als Bildungsforscher und Evaluator habe ich ganz unterschiedliche Untersuchungsmethoden genutzt: normierte Tests, standardisierte Frage- und Beobachtungsbögen, aber ebenso offene Aufgaben, informelle Gespräche und beiläufige Beobachtungen. Ich habe Daten in großen Stichproben und an einzelnen Fällen erhoben, sie in narrativen Porträts ausgewertet, mit statistischen Verfahren verdichtet oder aus Kennwerten anderer Studien synthetisiert. Nach jeder Untersuchung habe ich mehr davon verstanden, wie Kinder und Jugendliche lernen, wie Lehrer/innen sie dabei unterstützen können, welche Bedingungen in der Schule und im Bildungssystem als Ganzem dabei förderlich bzw. hinderlich sind. Vor allem aber sind mir jedes Mal erneut die Grenzen eines solchen Zugangs bewusst geworden. Für die sogenannten qualitativen Methoden sind diese weithin bekannt, z. B. Verzerrung durch die begrenzte Fallauswahl und Personenabhängigkeit der Wahrnehmung. Anders für die sogenannten quantitativen Methoden: Unter dem Etikett »Evidenzbasierung« beanspruchen sie seit einigen Jahren eine besondere Autorität als »Goldstandard« oder bekommen diese von außen zugesprochen. Dieser Anspruch ist aus vielen Gründen problematisch – eine Einschätzung, die auch von vielen geteilt wird, die selbst empirische Forschung betreiben – allerdings unterschiedlich stark, wie sich auf dem Forum »Kritik empirischer Bildungsforschung« im Dezember 2014 in Hamburg gezeigt hat. Seine Ergebnisse sollen als fachinterne Diskussion 2015 in der *Zeitschrift für empirische Erziehungswissenschaft* veröffentlicht werden.

Dieses Buch ist dagegen für all jene geschrieben, die von Bildungsforschung, Evaluation und ihren Ergebnissen im pädagogischen Alltag betroffen sind. Es versteht sich als Lesehilfe für Interessierte in Politik, in den Medien, in den Schulen und Familien – auch damit sie sich gegenüber unberechtigten Geltungsansprüchen von »evidenzbasierten« Urteilen behaupten können. Ich hoffe, dass die Lektüre darüber hinaus auch für meine Kolleginnen und Kollegen in Forschungs- und Evaluationsprojekten ein Gewinn ist. Jedenfalls wenn sie die Perspektive wechseln. Denn nicht das, was sie – oft mit den besten Absichten – wollen oder meinen, sondern was sie tatsächlich bewirken und wie sie verstanden oder auch missverstanden werden, sollte bestimmen, wie sie die Ergebnisse ihrer Forschung darstellen.

In den vergangenen 40 Jahren habe ich dazu an verschiedenen Orten kritische Anmerkungen publiziert – als Versuch, eine andere Sicht auf Schule und ihre Evaluation starkzumachen. Nun also diese Zusammenfassung, um die Ansprüche und Grenzen

der geforderten »Evidenzbasierung« noch einmal systematischer zu diskutieren. Nach der Einführung in Probleme der Evaluation in Form einer inszenierten Diskussion (Kap. 1) folgt ein Überblick über grundlegende Fragen der »Evidenzbasierung« (Kap. 2). Diese Fragen werden anschließend – in jeweils für sich lesbaren Beiträgen – diskutiert im Blick auf ihren Ertrag

- für die *Bildungsforschung* selbst, z. B. durch Meta-Analysen, wie sie Hattie (2013) in seinem Band »Lernen sichtbar machen« zusammengefasst hat (Kap. 3);
- für *Bildungspolitik und -verwaltung*, etwa aus internationalen Leistungsvergleichen wie TIMSS*, PISA* & Co. (Kap. 4);
- für die *Schul- und Unterrichtsentwicklung*, z. B. aus Studien wie KESS und LAU* in Hamburg oder die bundesweiten Vergleichsarbeiten VerA* in Klasse 3 und 8 (Kap. 5);
- auf der *Schülerebene*, vor allem durch »diagnostische« Tests anstelle der Lernbeobachtung durch die Lehrperson (Kap. 6).[1]

Ursprünglich war dieses Buch als Sammelband bereits publizierter Einzelarbeiten geplant. Im Prozess des Schreibens zeigte sich jedoch, dass die angestrebte Systematik so nicht zu erreichen war und es zudem immer wieder Dopplungen gab. Die Texte wurden deshalb stellenweise gekürzt und im Kontext der Gesamtargumentation überarbeitet (die Originalquellen finden sich am Ende des Buches). Dennoch ist die Eigenständigkeit der Kapitel erhalten geblieben, um Einstiege an verschiedenen Stellen zu ermöglichen. Insofern gibt es Überlappungen und Wiederholungen. Sie helfen, zentrale Einsichten in unterschiedlichen Kontexten neu zu beleuchten. Auch die Stilform der Beiträge wechselt, um verschiedenen Lesergruppen unterschiedliche Zugänge zu ähnlichen Fragestellungen zu eröffnen.

Über die Querverweise zwischen den Kapiteln lassen sich zudem eigene Lesewege gehen. Die als Diskussion angelegten Kapitel erleichtern vor allem Laien den Zugang zu den Kontroversen über »Evaluation und Rechenschaft« (Kap. 1) sowie über das Verhältnis von Großstudien mit standardisierten Instrumenten und Fallstudien, die mit offeneren Verfahren arbeiten (Kap. 3.2). Das zweite Kapitel enthält grundlegende Überlegungen zu den Problemen einer »Evidenzbasierung« von Bildungspolitik und pädagogischer Praxis. Fachkollegen werden vermutlich am ehesten von Kapitel 4 profitieren, in dem einzelne Aspekte von PISA diskutiert werden.

Um der besseren Lesbarkeit willen habe ich die Literaturverweise auf direkt zitierte Quellen beschränkt. Für die vertiefende Lektüre finden sich am Ende einige sparsam ausgewählte Leseempfehlungen zu den Hauptkapiteln. An der fachinternen Diskussion interessierte Leser/innen können die kompletten Literaturnachweise beziehen über:

hans.bruegelmann@grundschulverband.de.

---

1 Begriffe, die mit * gekennzeichnet sind, werden im Glossar erläutert.

Zur Entstehung dieses Buches haben viele wichtige Beiträge geleistet. Mein Dank gilt den Gegenlesern der ersten Entwürfe für ihre hilfreichen Hinweise: Axel Backhaus, Erika Brinkmann, Karl-Heinz Imhäuser, Gerheid Scheerer-Neumann und ganz besonders Hans Werner Heymann, der mir mit seinen klugen und sehr detaillierten Anmerkungen geholfen hat, den vertrauten Text aus Leserperspektive gründlich zu überarbeiten. Den letzten Feinschliff hat das Manuskript durch das behutsame Lektorat von Erik Zyber im Beltz Verlag erhalten.

Einen besonderen Schub haben mir beim Nachdenken über die Inhalte dieses Buches die Diskussionen in der Ferienakademie »Wem oder was nutzt Evidenzbasierung in der Pädagogik?« der Studienstiftung im Sommer 2014 gegeben. Für die aktive Mitgestaltung dieses für mich produktivsten Seminars meines Berufslebens danke ich Pia Algermissen, Christine Blinn, Dominik Bott, Fabio Bove, Andreas Bugl, Viktoria Dreieicher, Gerlinde Eberle, Markus Stephan Forster, Muriel Frenznick, Kathrin Hildebrandt, Marlene Kowalski, Lisa Landeck, Nicola Désirée Schulte, Anna Schwenke, Anne Sprink, Stefanie Tamke, Holger Wende und Justus Zokaie.

Hans Brügelmann *Findorff/Laubach im Winter 2014*

Vorspiel

Evaluation ist ein Alltagsphänomen. Sie dient der Vorbereitung von Entscheidungen durch die Sammlung, Beschreibung, Analyse und Bewertung von Informationen – etwas, das jeder von uns tagtäglich und in vielfältigen Situationen tut. In der Küche probieren wir neue Rezepte aus, und aufgrund unserer Erfahrungen verfeinern wir zunehmend unsere Kochkünste; wenn wir umziehen, sammeln wir Informationen über die Handwerker und Ärzte vor Ort und bauen uns aufgrund von Empfehlungen eine neue Infrastruktur auf; vor dem Kauf eines neuen Autos reden wir mit Bekannten, werten Berichte im Internet aus, machen Probefahrten bei verschiedenen Händlern; wenn wir nur zwei Bücher auf eine Wanderung mitnehmen können, wählen wir unsere Ferienlektüre gezielt nach den Erfahrungen der Vorjahre und nach Empfehlungen aus Zeitschriften oder von Freunden aus. Jeder von uns evaluiert also ständig sein eigenes und fremdes Handeln. Das kann aber auf sehr verschiedene Weise geschehen. Es gibt drei Archetypen des Evaluators (→ »Scharfe Brillen, wache Augen und ein einfühlsamer Blick« 2007):[2]

(1) Der Warentester im Bereich der Produktionskontrolle und des Konsumentenschutzes arbeitet mit technisch definierten Maßstäben (z. B. einer zulässigen Fehlertoleranz), anhand derer er die Güte von gleichen oder zumindest gleichartigen Produkten feststellt. Ich bezeichne dies als die *technisch-methodische* Lösung des Urteilsproblems.

(2) Der Kritiker in Kunst, Musik, Theater und Literatur zeichnet sich demgegenüber durch seine individuelle Fähigkeit zur originellen Wahrnehmung und Vermittlung ästhetischer Qualitäten aus. Kann er als Repräsentant bestimmter Denktraditionen über die Zeit hinweg eine eigene Klientel gewinnen oder zum Meinungsführer einer bestehenden Subkultur aufsteigen, wird er zur Autorität – eine *personengebundene* Lösung des Urteilsproblems.

(3) Der Richter ist im juristischen Prozess zwar an inhaltliche Normen gebunden, trägt aber für ihre Auslegung auf den Einzelfall und für die Feststellung des Sachverhalts, auf den die Norm anzuwenden ist, die Verantwortung. Das Verfahren ist gleichwohl durch soziale Regeln derart formalisiert, dass man diesen Typ als *institutionelle* Lösung des Urteilsproblems bezeichnen kann.

Im pädagogischen Bereich finden wir alle drei Typen wieder: den technisch Messenden in den internationalen Leistungsvergleichen, den Kritiker als Gutachter bei der Lehrmittelzulassung und den Richter als Prüfer in Examina.

Dieses Buches enthält dazu zwei zentrale Botschaften: Was für das Gelingen von Lernen wesentlich ist, kann nicht auf dieselbe Art und Weise erfasst werden, wie man

2 Das Zeichen → verweist auf Beiträge, die im Internet unter der Adresse www.beltz.de (Buchdetailseite) verfügbar sind.

das Funktionieren technischer Geräte oder ökonomischen Erfolg misst. Erst recht kann man es nicht planen und mithilfe von Evaluationsverfahren steuern, die aus diesen Subkulturen entlehnt worden sind. Das heißt nicht, Schulen könnten ohne Evaluation auskommen. Aber die Pädagogik braucht andere Formen der Untersuchung und Bewertung von Leistung, als sie in Technik und Wirtschaft üblich sind. Dafür muss sie nicht am Punkt Null anfangen, denn in der pädagogischen Praxis sind bereits viele brauchbare Beispiele entwickelt worden.

Interessant ist dabei zum einen, dass wir in der Praxis häufig Mischformen dieser Evaluationstypen finden, z. B. die Lehrerin, die testet, aber auch informell bewertet und am Ende »von oben« Noten vergibt. Auch ist die diagnostische Funktion – wie beim praktischen Arzt – in der Regel mit der »therapeutischen« Verantwortung verknüpft: Die Lehrerkonferenz vergleicht nicht nur Lehrmittel, sie wählt auch aus; Lehrer/innen bewerten nicht nur Leistungen, sondern fördern die Schüler/innen auch.

Zum anderen ist in den letzten 30 bis 40 Jahren eine Zunft von Evaluationsspezialisten entstanden, die sich vor allem aus der pädagogischen Psychologie rekrutiert. Diese Entwicklung hat zu einer Differenzierung der Aufgaben, Rollenmuster und Urteilsformen geführt und zu einer »Expertokratie«, die die Praxis zunehmend entmündigt. Denn die Fülle an Optionen, die Vielfalt relevanter Kriterien und die wachsende Verfügbarkeit von Informationen erschweren diese Aufgabe zunehmend – nicht nur im Privatleben, sondern auch im Beruf. Angesichts dieser wachsenden Komplexität haben wir zwei Optionen, im pädagogischen Feld wie auch anderswo:

(1) Wir können Evaluationen an Spezialisten delegieren: an die Stiftung Warentest, an den TÜV, an Computer-Bild – und im Bereich der Schule an ein Institut für Bildungsforschung oder eine professionelle Inspektion. An dieser Stelle kommt mit dem Zauberwort »Evidenzbasierung« das Versprechen wissenschaftlicher Absicherung ins Spiel. Der Preis: Abhängigkeit von externen Experten und Übersetzungsprobleme bei den Ergebnissen.
(2) Die Alternative heißt Entwicklung der Evaluationskompetenz im Handlungsfeld, also Unterstützung der Entscheidungsträger selbst, um die Diagnosefähigkeit von Lehrer/inne/n und die Fähigkeit zur Selbsteinschätzung von Schüler/innen/n zu fördern und um Schulen zu helfen, den Erfolg ihrer Arbeit selbst zu untersuchen und Rechenschaft über ihn abzulegen (→ »Evaluation als Dienstleistung für die Unterrichtspraxis« 1976). Der Preis: Vermittlungsprobleme bei den Verfahren und Mehrarbeit vor Ort.

Bei der Organisation von Evaluation geht es immer auch um Macht. In einem viel zitierten und mehrfach nachgedruckten Artikel hat Barry Macdonald (1976) schon vor vierzig Jahren drei Typen der Machtverteilung in Evaluationsvorhaben unterschieden:

(1) Als *bürokratische* Evaluation bezeichnet er Bildungsforschung im Dienste eines Auftraggebers (meist der Bildungsverwaltung), dessen Ziele unhinterfragt als Kriterium für die Beurteilung eines Programms übernommen werden. Wirksame und sparsame Verwendung der vorhandenen Mittel ist wesentlicher Maßstab für den Erfolg eines Programms.

(2) Eine *autokratische* Evaluation bezieht ihre Kriterien aus der akademischen Subkultur. Sie bietet den Entscheidungsträgern eine externe Rechtfertigung ihrer Politik, sofern diese die Empfehlungen des Evaluators befolgen. Der Evaluator verlangt dafür Unabhängigkeit in der Bestimmung der Urteilsmaßstäbe und in der Durchführung der Untersuchung. Die objektive Prüfung des Programms an übergreifenden Maßstäben ist der Anspruch dieses Evaluationstyps.
(3) Eine *demokratische* Evaluation akzeptiert Wertkonflikte als Ausgangspunkt der Untersuchung. Sie handelt die Interessen der Auftraggeber mit den Betroffenen aus und bezieht unterschiedliche Deutungen in die Beurteilung des Programms ein, um die Auseinandersetzung auf einer besseren Informationsgrundlage (aller Beteiligten) neu anzuregen. Ein allgemeiner Zugang zu handlungsbedeutsamer Information ist der zentrale Wert dieses Modells.

Mit diesen Alternativen stehen wir im Bildungsbereich vor einem grundsätzlichen Dilemma der Organisation moderner Gesellschaften: Vertrauen auf Spezialistenwissen vs. demokratische Kontrolle von Ungewissheit.

Für mich ist das auch berufsbiografisch eine zentrale Frage. So ist Evaluation Thema eines Buches, das mir seit seinem Erscheinen vor 40 Jahren bis heute viel bedeutet »Zen und die Kunst, ein Motorrad zu warten« von Robert Pirsig (1978, engl. 1974). Dieses Buch bündelt – wie durch ein Brennglas – Erfahrungen, die mein berufliches Leben zur Zeit der ersten Welle von Evaluation und *Accountability* (Rechenschaft) nachhaltig geprägt haben.

Ich hatte nach meinem ersten juristischen Staatsexamen das Glück, ein Stipendium der Stiftung Volkswagenwerk zu bekommen. Dadurch konnte ich zunächst zwei Jahre beim deutschen Bildungsrat in der Kommission »Strategien der Curriculum-Entwicklung« an Konzepten der Qualitätsentwicklung mitarbeiten und danach für einige Zeit in England, in den USA und in Kanada in Instituten für pädagogische Evaluation konkrete Verfahren, sozusagen »on the job«, kennenlernen. In England waren es Lawrence Stenhouse, Barry MacDonald und John Elliott am *Centre for Applied Research in Education*, die mich mit ihren Ideen zur Demokratisierung und Rechenschaft von Unterricht und Schule geprägt haben, und in Urbana-Champaign/Illinois war es neben Ernest House vor allem Bob Stake mit seinem Konzept »naturalistischer Evaluation«.

Als viel gefragter Experte hatte Bob Stake wenig Zeit, sich um den jungen Gast aus Deutschland zu kümmern. Aber am letzten Tag meines Aufenthaltes lud er mich zu einem Abschlussgespräch in sein Zimmer. Ich stand schon in der Tür, um mich zu verabschieden, als er mich fragte: »Kennst du ›Zen and the art of motorcycle maintenance‹?« Als ich verneinte, meinte er: »Das wird dir gefallen. Das ist besser als jedes Fachbuch über Evaluation. Ich habe es in meinem letzten Seminar als Basistext benutzt.«

Schon wegen des exotischen Titels war ich sehr skeptisch, hielt die Empfehlung mehr für die Grille eines etwas versponnenen Gelehrten. Aber nach meinem Flug von Urbana nach Chicago hatte ich bis zum Abflug nach Deutschland noch ein paar Stun-

den Zeit und fuhr dann doch mit dem Bus vom Flughafen in die Stadt, um eine Buchhandlung aufzusuchen. Und ich hatte Glück, noch ein Exemplar des Buches zu erwischen, das schon wenige Monate nach seinem Erscheinen ein Kultbuch geworden war.

In mehrfacher Hinsicht ist es eine Reisebeschreibung: eine mentale Reise durch die Welt im Kopf eines Philosophie-Dozenten, der mit seinem Denken an die Grenzen menschlicher Erkenntnis stößt und darüber verrückt (oder für verrückt erklärt) wird; die reale Motorradtour dieses Mannes mit seinem Sohn und zwei Freunden, um nach »erfolgreicher« Elektroschock-Behandlung in der Psychiatrie sein Gedächtnis und damit seine Identität durch das Aufsuchen biografisch bedeutsamer Orte wiederzugewinnen; schließlich eine Reise durch die abendländische und zum Teil auch durch die östliche Geistesgeschichte.

Pirsigs zentrales Thema ist »Qualität« und ihre Bestimmung in einem technisch geprägten, aber ohne subjektive Bedeutung wertlosen Alltag:

> »Er sah zwei Welten, gleichzeitig. Auf der intellektuellen Seite, der ›squareness‹-Seite, erkannte er jetzt, dass Qualität ein Spaltbegriff war. (...) Man nimmt sein analytisches Messer, setzt die Spitze genau auf den Begriff Qualität, klopft einmal darauf, gar nicht fest, nur ganz leicht, und die ganze Welt teilt sich, zerfällt glatt in zwei Hälften – hip und square, romantisch und klassisch, humanistisch und technologisch – und der Bruch ist sauber. (...)
> Die romantische Anschauungsweise ist vorwiegend durch Inspiration und Phantasie bedingt, kreativ und intuitiv. Gefühle sind wichtiger als Fakten. ›Kunst‹ als Gegensatz zu ›Wissenschaft‹ ist oft romantisch. (...) Die klassische Anschauungsweise beruht hingegen auf der Vernunft und auf Gesetzen. (...) Motorradfahren ist romantisch, Motorradwartung hingegen rein klassisch. (...) Der klassische Stil ist direkt, schmucklos, gefühlsfrei, ökonomisch und von ausgewogenen Proportionen. Er will nicht das Gefühl ansprechen, sondern Ordnung ins Chaos bringen und das Unbekannte bekannt machen. (...) Alles ist unter Kontrolle. Sein Wert bemisst sich nach der Geschicklichkeit, mit der diese Kontrolle aufrechterhalten wird. Einem Romantiker kommt die klassische Betrachtungsweise oft stupide und hässlich vor, genau wie die Wartung technischer Gegenstände selbst. Alles dreht sich nur um Teile und Einzelteile und Bestandteile und Beziehungen. Nichts lässt sich restlos klären, bevor es nicht zehnmal durch einen Computer gelaufen ist. Alles muss gemessen und bewiesen werden. (...)
> ...es ist wichtig, dass ich jetzt eine Beziehung zwischen der Liebe zur Sache und der Qualität herstelle, indem ich zu zeigen versuche, dass Liebe zur Sache und Qualität der innere und der äußere Aspekt ein und derselben Sache sind. Wer Qualität sieht und sie bei der Arbeit spürt, dem liegt etwas an den Dingen. (...)
> Die klassische und die romantische Auffassung von Qualität müssen kombiniert werden.« (Pirsig 1978, S. 74 f., 224 f., 284 f., 302)

Die Spannung zwischen klassischer und romantischer Weltsicht ist konstitutiv für die Sozialwissenschaften. Sie hat mich mein ganzes Berufsleben begleitet – bis in die Neuropsychologie hinein. Was sie konkret für die Bildungsforschung, für Politik und Unterricht bedeutet, will ich in den folgenden Kapiteln an dem Anspruch diskutieren, dass Entscheidungen in diesen Bereichen nur noch »evidenzbasiert« zu fällen seien. Und dafür ist entgegen der verbreiteten Publikationshektik nicht nur Literatur relevant, die erst nach dem Jahr 2000 veröffentlicht wurde.

Es ist immerhin auch schon 35 Jahre her, dass die OECD ihr »Basic Education Policies Project« durchgeführt hat. In meinem Beitrag → »Experimental decision making and responsive accountability« (1980) habe ich damals Verfahren der Rechenschaftslegung in den verschiedenen Mitgliedsstaaten begutachtet. Dabei hatte ich den Eindruck gewonnen, dass die Diskussion eine solche Reife erreicht hat, dass man sich der Chancen und Risiken verschiedener Methoden allseits bewusst ist. Meine Erfahrungen mit Evaluation in den vergangenen 15 Jahren – ob internationale Systemvergleiche oder Bewertung individueller Leistungen – lassen mich inzwischen wieder daran zweifeln. Technisch-methodisch wird heute auf einem sehr hohen Niveau gearbeitet und diskutiert, aber die sozialen Kontexte und die politischen Implikationen werden nicht gleichermaßen reflektiert. Dies ist kein deutsches Phänomen, wie die Verhandlungen auf den (1972 bis 2004) wiederkehrenden *Cambridge Conferences on Evaluation* zeigen (vgl. Elliott/Kushner 2007). Die Kritik an einer technologisch verkürzten »Qualitätswende« wird allerdings auch in den angelsächsischen Ländern kaum zur Kenntnis genommen. Dort dominiert wie bei uns die Forderung nach einer »Evidenzbasierung« mithilfe methodisch perfektionierter Tests.

Qualität von Evaluation hat mit der Qualität von Beziehungen zu tun, es geht dabei um Offenheit, Vertrauen, Glaubwürdigkeit. Wir brauchen eine faire Verteilung der Deutungsmacht zwischen den Beteiligten – gerade in einem so heiklen Feld wie der Pädagogik. Menschen und ihr Lernen lassen sich nicht vermessen wie Maschinen und technische Prozesse. Und noch vermessener wäre es zu glauben, sie ließen sich in vergleichbarer Weise steuern und planen.