pedocs
Open Access Erziehungswissenschaften

DIPF

Brügelmann, Hans

Vermessene Schulen - standardisierte Schüler. Zu Risiken und Nebenwirkungen von PISA, Hattie, VerA & Co.

Weinheim, Basel : Beltz 2015, 143 S. - (Pädagogik)

Quellenangabe/ Reference:
Brügelmann, Hans: Vermessene Schulen - standardisierte Schüler. Zu Risiken und Nebenwirkungen von PISA, Hattie, VerA & Co. Weinheim, Basel : Beltz 2015, 143 S. - (Pädagogik) - URN: urn:nbn:de:0111-pedocs-278901 - DOI: 10.25656/01:27890

https://nbn-resolving.org/urn:nbn:de:0111-pedocs-278901
https://doi.org/10.25656/01:27890

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

PÄDAGOGIK

Hans Brügelmann

Vermessene Schulen – standardisierte Schüler

Zu Risiken und Nebenwirkungen von PISA, Hattie, VerA und Co.

BELTZ

Hans Brügelmann

Vermessene Schulen – standardisierte Schüler

Zu Risiken und Nebenwirkungen von PISA, Hattie, VerA & Co.

BELTZ

Dr. Hans Brügelmann war bis 2012 Professor für Erziehungswissenschaft mit dem Schwerpunkt Grundschulpädagogik und -didaktik. Seitdem arbeitet er als Fachreferent für Qualitätsentwicklung im Grundschulverband. Er zählt zu den profiliertesten Kritikern einer einseitig testorientieren Bildungsforschung.

Dieses Buch ist auch als E-Book erhältlich
(ISBN 978-3-407-29204-9).

Das Werk und seine Teile sind urheberrechtlich geschützt.
Jede Nutzung in anderen als den gesetzlich zugelassenen Fällen
bedarf der vorherigen schriftlichen Einwilligung des Verlages.
Hinweis zu § 52a UrhG: Weder das Werk noch seine Teile dürfen
ohne eine solche Einwilligung eingescannt und in ein Netzwerk
eingestellt werden. Dies gilt auch für Intranets von Schulen
und sonstigen Bildungseinrichtungen.

© 2015 Beltz Verlag · Weinheim und Basel
Werderstr. 10 · 69469 Weinheim
www.beltz.de

Lektorat: Dr. Erik Zyber
Herstellung: Michael Matl
Satz: Sonja Frank
Druck und Bindung: Beltz Bad Langensalza GmbH, Bad Langensalza
Umschlagabbildung: © Nisian Hughes, getty images
Reihengestaltung: glas ag, Seeheim-Jugenheim
Umschlaggestaltung: Michael Matl
Printed in Germany

ISBN 978-3-407-25729-1

Inhaltsverzeichnis

Einladung 7

Vorspiel 11

1. **Wozu Evaluation? Inszenierte Kontroverse in verteilten Rollen** 17

2. **Über das Spiel mit Zahlen hinaus – Grundprobleme einer »Evidenzbasierung«** 31
   - 2.1 Sind Forschungsergebnisse »evident«? 31
   - 2.2 Von der Medizin lernen? 33
   - 2.3 Untersuchungsdesign: viele Fälle oder reiche Daten? 35
   - 2.4 Untersuchungsmethode: eindeutige Daten oder bedeutungsvolle Informationen? 51
   - 2.5 Maßstab der Bewertung: starke Wirksamkeit oder hohe Wichtigkeit? 57
   - 2.6 Wege aus den Dilemmata: Anregungen und Forderungen 58
   - 2.7 Balance von Ebenen und Rollen: ein differenziertes System von Rechenschaftsformen 62

3. **Hattie und der Zauber großer Zahlen** 65
   - 3.1 Was bringen Meta-Analysen der Unterrichtsforschung? 65
   - 3.2 Das Kind zwischen Hattie und PISA 70

4. **PISA & Co.: Nutzen und Grenzen von Leistungsvergleichen auf Systemebene** 77
   - 4.1 Warum wir PISA & Co. brauchen, die Ergebnisse aber nur mit Vorsicht nutzen sollten 77
   - 4.2 Stärken, Reichweite und Schwächen von PISA 80
   - 4.3 Eignet sich PISA als Paradigma für die Evaluation von Schule und Unterricht generell? 89
   - 4.4 Fazit 94

5. **Evaluation von Schule und Unterricht** 95
   - 5.1 Qualität und die Kunst, den Erfolg von Unterricht zu bestimmen 95
   - 5.2 Wahrheit durch VerA? 102
   - 5.3 Schulevaluation durch »kritische Freunde« 113

6. Die Not mit den Noten 117
6.1 Differenzierung der Notenkritik 117
6.2 Tests oder Texte als Alternative zu Noten 120
6.3 Dialogische Verbindung von Fremd- und Selbstbeurteilungen in einer pädagogischen Leistungskultur 125

7. Zehn Thesen zur Diskussion 129

Hinweise für eine vertiefende Lektüre 131

Literatur 133

Quellennachweise 139

Glossar 141

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.

1. Wozu Evaluation? Inszenierte Kontroverse in verteilten Rollen

Moderatorin: Kürzlich habe ich über die Medien folgenden Vorschlag der Krankenkassen mitbekommen: Ärzte sollten – ähnlich wie im alten China – danach bezahlt werden, wie gut und wie rasch sie Krankheiten heilen. Nur so könne die Qualität des Gesundheitswesens verbessert und der finanzielle Aufwand Erfolg versprechend konzentriert werden. Dieses Beispiel wirft die Frage auf, wie der Lehrerfolg des Schulsystems und einzelner Lehrer/innen erfasst werden kann und welche Rolle dabei landesweite Testprogramme wie VerA* oder LAU* in Hamburg, aber auch internationale Leistungsvergleiche wie die IEA*-Lesestudie, wie TIMSS*, PISA* und IGLU* spielen können. Dazu haben wir, wie ich hoffe, eher nachdenkliche Vertreter verschiedener Verbände eingeladen. Bitte sehr!

Arbeitgeber: Wir begrüßen sehr, dass endlich überprüft wird, was unsere Schulen leisten. Als Steuerzahler haben wir ein Recht darauf zu wissen, was mit unserem Geld passiert und ob die Schulen ihren gesellschaftlichen Auftrag erfüllen.

Lehrer: Und das soll mit diesen Tests gelingen? Ich nehme ihre zunehmende Verbreitung im Schulalltag als sehr problematisch wahr. Die Leistungen, vor allem die fachlichen Leistungen der Schüler/innen, machen doch nur einen Teil des Auftrags der Schule aus. Qualität erweist sich nicht in kurzfristigen Ergebnissen, sondern in den langfristigen Wirkungen des persönlichen Umgangs der Lehrer/innen mit den Schüler/inne/n, im inhaltlichen Anspruch von Aufgaben und Aktivitäten und im methodischen Niveau ihrer Bearbeitung. Prozesse haben einen Eigenwert. Es ist wichtig, wie junge Menschen in unserer Gesellschaft aufwachsen sollen, und es geht damit um langfristige Entwicklungen, die sich messtechnisch nicht zureichend erfassen lassen. Auch den Wert der Demokratie würde doch niemand nur an ökonomischen Indikatoren festmachen wollen, oder?

Elternvertreterin: Ich verstehe gar nicht, warum in den Schulen so ein Wirbel um Tests gemacht wird. Leistungen werden doch überall bewertet. Wenn ich ein neues Auto oder auch nur einen neuen Toaster kaufen will, besorge ich mir einen aktuellen Warentest. Und auch Personen bewertet jeder von uns tagtäglich, z. B. wenn wir den Friseur wechseln, weil wir unzufrieden sind, oder wenn wir Bekannten unsere Hausärztin empfehlen. Warum tun sich Pädagogen so schwer mit der Forderung nach Leistungsmessung?

Lehrer: Wir tun uns nicht schwer, wir hinterfragen lediglich die Art der Leistungsmessung. Sie gehört doch zu unserem Alltag. Wir Lehrer/innen erfassen und bewerten Leistungen von Schüler/inne/n, und unsere Vorgesetzten entscheiden nach einer Un-

terrichtsbeurteilung über die Verbeamtung junger Lehrer/innen und über die Besetzung von Funktionsstellen.

Arbeitgeber: Aber diese Vorgänge müssen transparenter werden! Warum reagieren Sie so empfindlich auf den Einsatz von standardisierten Instrumenten? Sie werden in Zukunft nicht mehr damit rechnen können, dass öffentliche Mittel einfach so fließen, ohne dass sich der Aufwand durch einen entsprechenden Ertrag rechtfertigen lässt. Die knappen Mittel müssen sparsam und effektiv eingesetzt werden – wie in der Industrie und anderen Bereichen der öffentlichen Verwaltung auch. Schulen können sich zukünftig nicht mehr diffus auf »guten Unterricht« berufen, wenn ihre Schüler/innen nicht anständig lesen, schreiben und rechnen lernen.

Schulaufsicht: Sie sehen das zu technisch. Die Bewertung von Lern- und Lehrerfolgen ist kein Warentest, und Unterricht ist auch kein Markt, auf dem Kunden zwischen Dienstleistungen wählen – wie bei Friseur und Ärztin. Schule ist ebenso wenig ein Betrieb zur »Produktion« von Qualifikationen, die auf dem Arbeitsmarkt verwertbar sind, sondern ein Ort der Bildung. Die Qualität eines Theaters oder eines Krankenhauses können Sie ja auch nicht nur nach dem Verhältnis von finanziellem Input und Output bewerten. Bildung ist mehr als eine Addition von Kenntnissen und Fertigkeiten, die man isoliert abprüfen könnte. Tests erfassen doch nur ein Oberflächenverhalten und nicht das, was eine Person denkt.

Lehrer: Genau. Denken Sie nur an die Evaluation der Programme kompensatorischer Vorschulerziehung in den 70er Jahren. Damals kamen unmittelbar nach Programmende von »Head Start« und anderen Projekten die großen Erfolgsmeldungen aus den Begleitstudien: höherer IQ, bessere Sprachfähigkeit und bessere Leistungen in den anderen trainierten Bereichen. Wenige Jahre später hieß es, die Überlegenheit der Versuchsgruppen gehe schon während der Grundschulzeit verloren, der Aufwand lohne nicht. Wieder einige Jahre später wurden sogenannte Sleeper-Effekte festgestellt, das heißt diejenigen, die als Kinder an den Programmen teilgenommen hatten, erreichten bessere Schulabschlüsse, waren im Beruf erfolgreicher und hatten weniger Probleme im privaten und sozialen Umfeld.

Meine Sorge: Kurzfristige Leistungsmessungen erfassen wie in einer Momentaufnahme nur einen kleinen Ausschnitt der Wirkungen, der sich ohne Kenntnis der Erfahrungen im Unterricht nicht sinnvoll interpretieren, wohl aber gut als Munition im politischen Tagesgeschäft verwenden lässt.

Bildungsforscherin: Ihre Skepsis gegenüber Tests als Instrument schulischer Leistungsmessung hat eine lange Geschichte. Seit jeher wehrt sich die Schule gegen Urteile von außen und verweist auf ihre ständigen eigenen Leistungsüberprüfungen. Dabei hätten gerade Lehrer/innen allen Grund, selbstkritisch gegenüber ihrer eigenen Urteilsfähigkeit zu sein. In den 60er und 70er Jahren wurden z. B. reihenweise Untersuchungen zur

Notengebung in Klassenarbeiten publiziert. Dabei kam heraus: Verschiedene Lehrer/innen beurteilen dieselbe Arbeit sehr unterschiedlich, und zwar nicht nur Aufsätze, sondern auch vermeintlich eindeutige Mathearbeiten und Diktate; sogar dieselben Lehrer/innen beurteilen dieselbe Arbeit zu verschiedenen Terminen unterschiedlich. Und: Urteile über (gleiche) Schülerleistungen hängen mit sozialen Faktoren wie Status der Eltern und persönlicher Beliebtheit der Schüler/innen zusammen. Wenn man denkt, was alles von Noten abhängt, muss Leistungsmessung stärker objektiviert werden.

Lehrer: Theoretisch klingt das plausibel, aber auch standardisierte Tests haben ihre Schwächen: Dieselbe Leistung kann Ergebnis angelernten Wissens oder selbstständigen Denkens sein. Und politisch ist es naiv zu glauben, in der Öffentlichkeit würden Daten aus Leistungsstudien so differenziert wahrgenommen, wie die Befunde es erfordern.

Schulaufsicht: Vor allem: Auch quantitative Daten beschreiben nur einen Teil der Wirklichkeit, und sie sprechen nicht für sich selbst. Sie müssen interpretiert werden. Man müsste die Ergebnisse gleich von verschiedenen Positionen aus kommentieren, um eine einseitige Verwertung zu erschweren. So einfach ist das nicht mit der »Objektivität«.

Elternvertreterin: Die ist aber für uns Eltern ganz wichtig, um die Schullaufbahn unserer Kinder verlässlich planen, und auch, um die Arbeit ihrer Lehrer/innen einschätzen zu können. Insofern interessieren mich ganz konkret die Klagen der Betriebe über unzureichende, ja sinkende Leistungen der Schule. Was ist dran am sogenannten Leistungsverfall?

Arbeitgeber: Wir stützen uns dabei auf Ergebnisse von Tests bei den neu eingestellten Lehrlingen. Gewerkschaften, Erziehungswissenschaftler/innen, aber auch die jeweils betroffenen Regierungsparteien stellen unsere Klagen immer gleich infrage. Die Pädagogik und die Bildungspolitik verstehen sich wohl als unfehlbar. Aber ohne Bereitschaft, sich einer Außenkritik zu stellen, sind Fortschritte nicht denkbar.

Bildungsforscherin: Das sehe ich auch so. Andererseits leiden die Einstellungsuntersuchungen unter einer Reihe von Mängeln: Inhaltlich erfassen die Tests nur kleine und in der Praxis oft nicht sonderlich relevante Ausschnitte schulischer Ziele (meist nur die Rechtschreibung, und die nur im Diktat, und die Grundrechenarten). Methodisch sind die Aufgaben oft problematisch (z. B. exotische Wörter oder das alltagsferne und irritierende Angebot, unter mehreren falschen Schreibungen die richtige auszuwählen). Für Verallgemeinerungen sind die Stichproben* nicht repräsentativ, und für historische Vergleiche fehlen zudem vergleichbare Bezugsgruppen aus früheren Zeiten.

Lehrer: Und Zeitungsberichte wie »Leistung zählt nicht mehr« oder gar »Neue Max-Planck-Studie widerlegt die Reformpädagogik« haben die Aussagekraft der Daten oft weit überzogen. Ihre methodischen Grenzen werden nicht beachtet, und zum Teil werden auch die Grundsätze einer fairen Interpretation empirischer Daten verletzt.

Schulaufsicht: Das ist aber doch Vergangenheit. Mit TIMSS* oder LAU* sind forschungsmethodisch versierte Institute in das Geschäft eingestiegen. Meine Sorge ist eher die Eigendynamik, die solche Testprogramme entfalten. Ob die Forscher/innen der Politik nicht doch zu viel versprechen? Um auf das Ärzte-Beispiel zurückzukommen: Wer kann eigentlich sagen, wann ein Patient »gesund« ist? Analog ist unser Problem: Kann man Bildung (und nicht nur Kenntnisse und Fertigkeiten) verlässlich messen, z. B. mit Kompetenztests die Einhaltung von Bildungsstandards? Und dann auch noch belegen, wo die Gründe für unterschiedliche Lernerfolge liegen?

Arbeitgeber: Wir haben doch dasselbe Problem bei der Beurteilung unserer Mitarbeiter/innen in den Betrieben. Warum wollen Sie sich vor dieser Anforderung drücken? Die zentrale Frage ist doch: Ist unser Bildungssystem international konkurrenzfähig? Und eine zweite: Investieren wir in den Bildungshaushalten an den richtigen Stellen? Ein Kollege von Ihnen hat es mal treffend mit dem schönen Satz gesagt, notwendig sei eine Haltung, »die die pädagogische Version der protestantischen Rechtfertigungslehre zu überwinden trachtet, wonach es für die Erlösung allein auf den Glauben ankommt, während die Werke keine Bedeutung haben« (Lange in VBE 2000, S. 48).

Lehrer: Das klingt so einfach. Als ob man nur ein Metermaß oder ein Thermometer nehmen müsste und dann daran den Zustand des Bildungswesens ablesen könnte. Wir haben doch schon oft die Erfahrung gemacht, dass wissenschaftliche Untersuchungen nicht zu einem klaren Ergebnis führen. Ich erinnere nur an den Methodenstreit in der Lesedidaktik, der in den 60er Jahren in einem Patt endete. Oder an die Versuche mit Gesamtschulen in den 70er Jahren, die von Verfechtern und von Gegnern bis heute unterschiedlich ausgeschlachtet werden. In den letzten Jahren ist die Integration behinderter Schüler/innen umstritten, weil verschiedene Untersuchungen unterschiedliche Ergebnisse erbringen. Mit der »Präzision« und »Objektivität« scheint es da nicht so weit her zu sein. Selbst Meta-Meta-Analysen sind in ihrer Aussagekraft zu Recht umstritten. Ich fürchte, mit Tests erwischt man nur ein paar kleine Zipfel von dem, worum es im Unterricht wirklich geht. Schlimmer noch: In den Händen der Messtheoretiker wird aus pädagogischen Konzepten wie »Selbstständigkeit« ein mechanisches System von Fähigkeitsmodulen, in dem ich meine Idee eines partnerschaftlichen Umgangs mit Kindern und Jugendlichen nicht mehr wiedererkenne.

Bildungsforscherin: Zugegeben: Wir stecken da in einer Falle. Versuchen wir, komplexe Fähigkeiten zu erfassen, wirft man uns vor, ihr »Wesen« zu verfehlen; verzichten wir darauf, diese Fähigkeiten einzubeziehen, wird uns vorgehalten, wesentliche Ziele der

Schule nicht zu berücksichtigen. Jeder von uns muss pragmatische Lösungen finden. Und wenn ich mir anschaue, mit was für primitiven Aufgaben immer noch viele Lehrer/innen Leistungen überprüfen, dann wünschte ich mir, dass das von uns in der Testentwicklung erreichte Niveau auch nur halbwegs Standard im Unterrichtsalltag wäre.

Elternvertreterin: Ich stimme Ihnen zu. Vom grünen Tisch aus kann man leicht immer neue Forderungen erheben, was ein »vernünftiges« Messinstrument alles zu leisten habe. Alles, was besser ist als der bisherige Schlendrian, ist einen Versuch wert. Heute können Lehrer/innen doch machen, was sie wollen. Und die faulen und unfähigen Kollegen bringen den ganzen Berufsstand in Verruf. Ich verstehe nicht, dass kompetente und engagierte Lehrer/innen diese Missstände decken. Sie sollten doch froh sein, wenn ihre Leistung anerkannt wird.

Schulaufsicht: So einfach ist das nicht. Vor allem hinkt der Vergleich mit den Prüfaufgaben der Lehrer/innen. Landesweite Testprogramme haben eine ganz andere Funktion. Und die Nebenwirkungen falsch gewählter oder falsch verstandener Aufgaben können fatal sein. Testergebnisse haben eine hohe Suggestivkraft. Durch ihre scheinbar technische Präzision gewinnen sie das Image einer Unparteilichkeit, die in der öffentlichen Diskussion leicht überschätzt wird. Laien, und dazu zählen auch Bildungspolitiker, Verbandsvertreter und Journalisten, können im Einzelfall die Annahmen, unter denen Aussagen von Tests gelten, und ihre konkrete Leistungsfähigkeit kaum einschätzen. In der Diskussion wird eine zu simple Alternative konstruiert: »objektive« Tests hier gegen »subjektive« (Vor-)Urteile da.

Lehrer: Zusätzlich sehe ich Probleme in der Passung von Test und Unterricht. Vor allem die Beteiligung an internationalen Vergleichen erzwingt oft problematische Kompromisse. Da wird schon mal der Lehrplan nur eines von 16 Bundesländern und der Aufgabenpool eines Schulbuchs zum Maßstab für »das Curriculum der BRD«. Und der Zeitdruck, unter dem die Testaufgaben entwickelt worden sind, erschwert es, solchen Schwächen auf die Spur zu kommen oder gar bessere Alternativen zu entwickeln. Wir haben ja schon bei TIMSS* gesehen, dass eine Reihe von Aufgaben inhaltlich oder sprachlich problematisch war, auch wegen der Übersetzungsprobleme.

Bildungsforscherin: Es ist sicher richtig, dass Papier-und-Bleistift-Aufgaben bei der Untersuchung von Lesen und Schreiben näher an den untersuchten Fähigkeiten sind als etwa in den Naturwissenschaften. Aber in Zusatzstudien haben sich die Antworten in den Auswahlfragen der Tests als gute Annäherung an die Leistungen in realen Experimenten erwiesen. Die Ergebnisse der Tests dienen ja nur als möglichst gute Indikatoren für die gemeinten Fähigkeiten, sie beanspruchen nicht, diese selbst zu erfassen.
Lehrer: In der Theorie mag das stimmen. Aber sie müssen wieder an die Wahrnehmung durch die Betroffenen denken. Schüler/innen oder Lehrer/innen, die in Tests gut abschneiden wollen, werden sich an den Indikatoren orientieren. Was bringt es z. B.

noch, im Unterricht zeitlich aufwendige Experimente durchzuführen, wenn Schüler/innen Tests auch bestehen können, indem man einfach solche Testaufgaben mit ihnen übt? Wir kennen das ja vom privaten Repetitor, den Studierenden vor dem juristischen Staatsexamen besuchen, um Prüfungsaufgaben zu pauken, statt zu studieren. »Teaching to the Test« lässt sich kaum vermeiden, wenn von den Ergebnissen der Tests persönliche Vor- und Nachteile für die Betroffenen abhängen. Denken Sie nur an die Nebenwirkungen von erfolgsabhängigen Verträgen für Vertreter in der Wirtschaft, z. B. im Zusammenhang mit der Bankenkrise. Da wurde auch ohne Rücksicht auf langfristige Folgen für die Firma oder gar für die Kunden auf rasche Erfolge gesetzt, um die Provision zu kassieren.

Bildungsforscherin: Ich finde, hier gehen verschiedene Dinge durcheinander. Wir müssen uns über den konkreten Zweck der Leistungsmessungen verständigen. Ich sehe da drei ganz verschiedene Funktionen: Landesweite Testprogramme können einmal dazu dienen, die Vergabe von Berechtigungen auf einen einheitlichen Maßstab zu beziehen. Vereinheitlichung von Prüfungen wie das Zentralabitur haben diese Funktion. Ein anderes Ziel ist es, die Arbeit einzelner Schulen oder gar Lehrer/innen zu bewerten. Vor allem in den angelsächsischen Ländern werden standardisierte Tests zu diesem Zweck eingesetzt. Bei Veröffentlichung der Daten kommt es dann zu dem viel diskutierten Ranking, einer Rangliste von Schulen. In einigen Kantonen der Schweiz wird auch überlegt, die Bezahlung der Lehrer/innen an Ergebnisse in solchen Tests zu knüpfen. Eine ganz andere Funktion haben schließlich die aktuell diskutierten länderübergreifenden Stichprobenuntersuchungen*, die lediglich auf einen Vergleich von Schulsystemen zielen.

Elternvertreterin: Wieso das? Wir Eltern haben doch ein berechtigtes Interesse an Informationen über die Qualität des Unterrichts in der Klasse unseres Kindes.

Bildungsforscherin: Wir müssen aktuell zwei Typen von Untersuchungen auseinanderhalten. Die Kultusministerkonferenz hat in ihren Konstanzer Beschlüssen vom Oktober 1997 einerseits länderübergreifende Untersuchungen wie TIMSS*, PISA* und für die Grundschulen IGLU* gefordert. Da geht es gar nicht um einzelne Lehrer/innen und Schulen, sondern um Vergleiche von Schulsystemen. Der zweite Ansatz zielt auf »qualitätsverbessernde Maßnahmen« innerhalb der einzelnen Bundesländer, und da gibt es unterschiedliche Ansätze. Landesweite Testprogramme wie VerA* und LAU* in Hamburg sind da die Ausnahme. Und auch dort werden keine Ranglisten veröffentlicht. Schulbezogene Daten erhalten nur die Schulen selbst, als Anstoß, intern über ihre Arbeit im Vergleich zu anderen Schulen nachzudenken. Wo liegt da das Problem?

Lehrer: Diese Unterscheidung kann man doch nur auf dem Papier so klar durchhalten. Wenn bei einem Systemvergleich herauskommt, dass bestimmte Schulformen – sagen wir mal Gymnasien mit Gruppierung nach Leistungsniveaus in den Hauptfächern –

nach bestimmten Kriterien schlechter abschneiden, dann hat das Auswirkungen auf die Wahrnehmung der Eltern in Schulen, die nach diesem Ansatz arbeiten, und damit indirekt auch auf die Arbeit in diesen Schulen.

Elternvertreterin: Aber das ist doch gut! Wo sollen die denn sonst die Information herbekommen, was dieses System taugt?

Schulaufsicht: Na, so einfach ist das auch nicht. Wenn bei einer Studie wie TIMSS im internationalen Teil herauskommt, dass Schulsysteme mit integrierten Sekundarschulen vergleichsweise gut abschneiden, im deutschen Bundesländervergleich aber, dass Gesamtschulen eher schlechter sind als dreigliedrige Systeme – was folgt daraus? Sollen wir die Gesamtschulen flächendeckend einführen, weil sie ihre Qualitäten nur voll entfalten können, wenn die Gymnasien ihnen nicht die besten Schüler/innen abziehen? Müssen wir die deutschen Gesamtschulen besser ausstatten, ihre Lehrer/innen besser aus- und fortbilden, damit sie das Potenzial der integrierten Schulform besser ausnutzen können? Oder sollten wir die Gesamtschulen ganz abschaffen, weil sie in unser gesellschaftliches Umfeld nicht so gut passen wie anderswo?

Bildungsforscherin: Möglicherweise werden hier die falschen Fragen gestellt. Die Organisationsform von Schule ist nach den vorliegenden Studien nicht der entscheidende Faktor für die Qualität des Unterrichts. Das hat auch die viel zitierte Meta-Meta-Analyse tausender Studien durch den neuseeländischen Bildungsforscher John Hattie gezeigt. Umso wichtiger ist es, dass wir die Wirkungen des Unterrichts erfassen und unterschiedliche Leistungen auf konkrete Bedingungen beziehen können, um diese dann zu verbessern.

Lehrer: Ich habe Probleme mit der Diskussion. Mal wird behauptet, die Programme zielten nur auf die Ebene des Gesamtsystems. Da frage ich mich, ob die Bildungspolitiker/innen überhaupt willens und in der Lage sind, Konsequenzen aus den Befunden zu ziehen. Was machen die z.B., wenn sich ein Ergebnis aus neueren US-amerikanischen Studien bestätigt, nämlich dass kleinere Klassen am Schulanfang förderlich sind und vor allem leistungsschwachen Schüler/inne/n zugutekommen? Werden die dann wirklich dort mehr Lehrer/innen einstellen?

Andererseits wird gesagt, die Testprogramme sollten den Schulen helfen, ihre Arbeit vor Ort zu verbessern. Aber dann gibt es Konflikte mit dem Datenschutz, und wegen der Anonymität und der fehlenden Dichte der Daten sind diese dafür nicht wirklich hilfreich.

Elternvertreterin: Dazu habe auch ich noch eine Frage. Landesweite Testprogramme werden ja immer wieder mit dem Argument begründet, man müsse die Arbeit der Schulen besser erfassen, um ihre Qualität steigern zu können. Warum schneidet dann aber ein Land, in dem solche Programme schon lange an der Tagesordnung sind, näm-

lich die USA, in internationalen Studien nicht besser ab als z. B. Deutschland? Dort mehren sich meines Wissens sogar die Stimmen, die eher eine schulnahe Aufsicht und Beratung der Schulen fordern, wie wir sie schon lange haben. Drehen wir uns da nicht im Kreis?

Lehrer: Da möchte ich mich gleich anhängen: Bei TIMSS* hat sich doch gezeigt, dass Länder mit zentralen Prüfungen nicht besser abschneiden als Bildungssysteme ohne. Woher rührt die Sicherheit, Tests seien besser als persönliche Beurteilungen und externe Leistungsmessungen führten zu mehr Qualität als schulinterne Prüfungen?

Schulaufsicht: Man muss doch sehen, dass kein System ohne Kontrolle auskommt. Manche Länder kontrollieren vorrangig den »Input«, z. B. die Qualifikation von Lehrer/inne/n oder die Vorgaben der Lehrpläne (so bisher die meisten kontinentaleuropäischen Staaten). Andere geben in diesem Bereich mehr Freiräume und kontrollieren dafür den »Output« stärker, z. B. durch zentrale Prüfungen (so traditionell die angelsächsischen Staaten). Im deutschen System zeigt sich in den letzten Jahren eine Tendenz zu mehr Schulautonomie, das heißt aber auch, die Aufsicht über »Input« und Arbeitsabläufe wurde verringert. Also müssen wir Rechenschaft und Kontrolle an anderer Stelle verstärken.

Lehrer: Und dazu greifen Sie auf Instrumente zurück, deren Wert und Auswirkungen äußerst umstritten sind, ganz zu schweigen von dem Problem, dass nicht einmal die Kriterien für eine grundlegende Bildung klar sind.

Arbeitgeber: Über Ziele wie Lesen, Schreiben und Rechnen dürfte es aber doch keinen Streit geben.

Elternvertreterin: Na ja, die Frage ist doch, welches Niveau in diesen Bereichen angesetzt wird: Was soll als Ziel für das Ende der Pflichtschule oder der Grundschulzeit gelten? Und auch inhaltlich gibt es Unterschiede: Geht es z. B. um das sinnbetonte Vorlesen einer Geschichte oder um die Beantwortung von Auswahlfragen zu einem Sachtext?

Arbeitgeber: Zu den Anforderungen kann man z. B. unsere Betriebe befragen, welche Grundfähigkeiten und welches Niveau im Berufsalltag erforderlich sind.

Elternvertreterin: Meines Wissens gibt es in dieser Hinsicht erhebliche Unterschiede zwischen verschiedenen Branchen, aber auch zwischen Betrieben innerhalb einer Branche.

Lehrer: Zudem setzen die staatlichen Vorgaben keine klaren Prioritäten. Während jeder Lehrplan seine fachlichen Ansprüche favorisiert, stehen in den allgemeinen Richtlinien fachunabhängige Schlüsselqualifikationen wie Kooperationsfähigkeit oder

Selbstständigkeit im Vordergrund. Für Lehrer/innen bedeutet das, sie müssen ihre begrenzten Ressourcen auf konkurrierende Ziele verteilen. Da sind unterschiedliche Kompromisse denkbar und auch legitim. Die Testprogramme setzen ihrerseits eigene Prioritäten. Ich befürchte als Folge eine Uniformierung des Unterrichts, die nicht mehr auf die Besonderheiten vor Ort Rücksicht nehmen kann.

Bildungsforscherin: Ich muss noch einmal darauf hinweisen, dass es bei IGLU* oder PISA* nicht um die Arbeit der einzelnen Schulen geht. Eine solche übergreifende Bestandsaufnahme kann aber erhellen, wo ein System insgesamt seine Stärken und seine Schwächen hat. Da geht es nicht um Lehrerin A oder Lehrer B.

Schulaufsicht: Indirekt doch. Und ich denke, das ist auch so gewollt. Durch eine solche Bestandsaufnahme wird ja ein öffentlicher Druck erzeugt. Die Begriffe »Stärken« und »Schwächen« machen das deutlich. Mit den Aufgaben werden bestimmte Maßstäbe gesetzt. Die geprüften Leistungen markieren, welche inhaltlichen Bereiche besonders wichtig sind, also auch, welches Niveau in diesen Bereichen erwartet wird. Es muss deutlich werden, dass dies eine politische Frage ist. Solche Entscheidungen kann man nicht Wissenschaftlern überlassen.

Lehrer: Meine Sorge ist, dass grundlegende Anforderungen nicht nur für den Schulabschluss, sondern auch für Zwischenschritte festgelegt werden. In England gibt es ja einen nach Jahrgängen gestuften nationalen Lehrplan. Nehmen Sie den aktuellen Fall der auch in Deutschland bekannten Reformschule »Summerhill«: Die Schüler/innen haben die Freiheit zu entscheiden, wann und wie sie was lernen, sie würden also je nach Lernweg bei Zwischentests eher schwach abschneiden. Am Ende der Schulzeit erreichen sie aber überdurchschnittliche Ergebnisse. Pädagogische Konzepte, die unterschiedliche Lernwege fördern und das individuelle Lerntempo respektieren, werden durch ein so gestuftes System gleichgeschaltet. Und wenn Sie dann noch an die unterschiedlichen Lernvoraussetzungen der Schüler/innen denken: schon am Schulanfang mehr als drei Jahre Entwicklungsdifferenz …

Bildungsforscherin: Wenn Sie sich Studien wie VerA* oder LAU* ansehen, gilt dieser Einwand nicht mehr. Dort werden auch sozialstrukturelle Daten erhoben, um den Einfluss unterschiedlicher Milieus auf die Arbeit einer Schule einzuschätzen.

Elternvertreterin: Aber eine Schule hat oft unterschiedliche Einzugsgebiete. So können zwei Parallelklassen in derselben Schule ganz unterschiedliche Voraussetzungen haben. Auch aus demselben Einzugsgebiet berichten Lehrer/innen für aufeinanderfolgende Jahrgänge ganz unterschiedliche Zusammensetzungen.

Bildungsforscherin: Das Problem löst sich auf, wenn wir Längsschnitte* anlegen, also dieselben Schüler/innen über mehrere Jahre begleiten, wie bei der LAU* in Hamburg

in der 5., 7. und 9. Klasse. Da vergleichen wir nicht nur die Leistung zu einem Punkt X, sondern messen Lernzuwächse.

Elternvertreterin: Aber ein Test kann doch nie umfassend die Fähigkeiten einer Person messen, sondern nur das unter bestimmten Bedingungen gezeigte Verhalten, und das kann je nach Situation wechseln.

Bildungsforscherin: Ihr Einwand ist grundsätzlich richtig. Darum werden ja auch Vertrauensintervalle für die Genauigkeit von Testwerten angegeben, um mögliche Schwankungen bei der Interpretation zu beachten. Man muss insofern zwischen Einzelfalldiagnose und Aussagen über Stichproben* unterscheiden. Der Einfluss von Tagesform oder Missverständnissen bei einzelnen Aufgaben gleicht sich in größeren Gruppen aus. Es ist also etwas anderes, ob wir die Leistung einzelner Schüler/innen oder die Leistung von Schülergruppen, z. B. eines Bundeslandes oder einer Schulform, miteinander vergleichen.

Schulaufsicht: Hier haben die Verantwortlichen ja auch dazugelernt. Wie schon gesagt: Die Daten werden nicht genutzt, um Urteile über einzelne Schulen, Lehrer/innen oder gar Schüler/innen zu fällen.

Lehrer: Aber bei VerA, den Vergleichsarbeiten in den 3. und 8. Klassen, werden die Ergebnisse schon nach einzelnen Schüler/inne/n ausgewertet. Und da erwarten die Eltern verlässliche Aussagen. Das ist für mich als Lehrer ganz schwierig, wenn es da Widersprüche zwischen dem Testergebnis und meinen Eindrücken über das Schuljahr hinweg gibt. Da kann man doch nicht einfach sagen, so ein einstündiger Test sei aussagekräftiger!

Moderatorin: Da tut sich in der Tat ein grundsätzlicher Konflikt auf. Lassen Sie uns trotzdem eine Zusammenfassung versuchen – zumindest der strittigen Punkte. Denn unsere Diskussion hat gezeigt, dass es sich bei Leistungsvergleichen um ein komplexes Problem handelt. Wir müssen meines Erachtens drei Ebenen unterscheiden:

- die *politische Ebene*: Wer bestimmt die Maßstäbe für Leistung und Schulerfolg, und wer kontrolliert deren Umsetzung? Zentrale oder dezentrale Entscheidungsträger?
- die *forschungsmethodische Ebene*: Welches Instrumentarium ist das aussagekräftigste, verlässlichste und vom Aufwand her ertragreichste? Die standardisierte Messung oder das personengebundene Urteil?
- die *reformstrategische Ebene*: Auf welchen Wegen ist Schulentwicklung am ehesten in Gang zu bringen? Durch eine interne oder durch eine externe Evaluation?

Außerdem gibt es je nach Standpunkt unterschiedliche Einschätzungen zu Funktion und Wirkungen des Instruments »Leistungstests«. Können Sie dazu die Ihnen wichtigsten Punkte noch einmal kurz auf den Nenner bringen?

Arbeitgeber: Aus meiner Sicht sollten Leistungsmessungen in allen drei Funktionen eingesetzt werden: zur *Beurteilung einzelner Schüler/innen*, um Fortschritte, aber auch Schwierigkeiten genauer erkennen und entsprechend gezielt fördern zu können und um die formelle Berechtigung für den Besuch weiterführender Schulen vergleichbar und nachprüfbar zu erfassen und auszuweisen; zur *Bewertung des Lehrerfolgs von Lehrer/inne/n*, indem die Wirkungen verschiedener methodischer Ansätze unter vergleichbaren Bedingungen überprüft werden; zur *Einschätzung von Stärken und Schwächen des Systems* insgesamt bzw. einzelner Teilsysteme, z. B. der verschiedenen Schularten.

Elternvertreterin: Ich wünsche mir, dass nicht mehr hintenherum über angeblich »gute« Schulen und »schlechte« Lehrer/innen geredet wird. Wir Eltern brauchen verlässlichere Informationen. Aber dazu müssten zunächst einmal die Ziele der Schule in Abstimmung aller beteiligten Gruppen geklärt und klarer umrissen werden: Wo liegen die Prioritäten für Allgemeinbildung heute? Wenn die Kriterien geklärt sind, könnten auch Zeugnisse aus verschiedenen Schulen und Bundesländern durch zentrale Tests vergleichbarer werden. Meine Sorge: Ich glaube nicht daran, dass eine genauere Information über den Stand des deutschen Bildungssystems politisch etwas bewirkt. Ein Beispiel: Seit Jahren berichtet die OECD, dass Deutschland für die Grundschule, in der doch die Grundlagen für alles Weitere gelegt werden sollen, viel weniger ausgibt als andere Länder, für die Oberstufe dagegen mehr. Geändert hat sich bis heute nichts.

Lehrer: Ich glaube auch nicht an die Objektivität von Tests: Der eine löst die Aufgabe, weil er kurzfristig viel auswendig gelernt hat, der andere ist durch eigenes Denken zu der Lösung gekommen. Das sind verschiedene Qualitäten, die ich an den Kreuzchen im Test nicht ablesen kann.

Aber auch für die Schulen liegt mir daran, dass die Ministerien jetzt nicht nur auf das Pferd »zentrale Leistungsmessung« setzen. Was nutzt es, wenn wir wissen, dass wir schlecht sind – oder jedenfalls schlechter als andere? Vom wiederholten Wiegen wird die Sau auch nicht fetter …

Und eine schärfere Kontrolle mag helfen, einzelne Extremfälle zu verhindern, besser wird der Unterricht in der Breite dadurch jedoch nicht. Um die Qualität von Unterricht zu verbessern, wäre es viel wichtiger, Mittel und Unterstützung für schulinterne Bestandsaufnahmen und vor allem für die Verbesserung der alltäglichen Arbeit und ihrer Bedingungen bereitzustellen. Wir brauchen Fallstudien von erfolgreichen Schulen, vor allem von Schulen, die unter schwierigen Bedingungen, also wider Erwarten, erfolgreich sind. Von deren Praxis können wir eher lernen, was wir konkret besser machen können, als aus dürren Testergebnissen.

Bildungsforscherin: Akzeptiert man, dass menschliche Erkenntnis immer beschränkt, dass menschliches Handeln immer unzulänglich ist, dann erreichen die neuen Leistungstests ein erfreulich hohes Niveau des Möglichen. Deshalb verdienen sie eine faire Chance als Versuch und als ein Element in einem größeren Verbund. Denn schon in

der IEA*-Studie, mehr noch aber in PISA* und bei VerA* nähern wir uns Prüfungsformen, die im schulischen Alltag in der Breite und Differenziertheit oft nicht erreicht werden: ein großes Spektrum anspruchsvoller Leistungen; offene neben geschlossenen Aufgaben; zwei Bewerter statt nur einer Person (wie im Unterricht).

Ertragreich werden diese Studien aber nur sein, wenn die Programme langfristig angelegt sind, sodass Entwicklungen erkennbar werden. Und nur dann können die Forscher/innen auch aus den ersten Studien lernen und ihre Instrumente kontinuierlich verbessern. Deshalb würde ich jetzt nicht das Land mit einem Testprogramm nach dem anderen überziehen, sondern lieber verschiedene Maßnahmen in überschaubarem Maßstab erproben. So sind auch am ehesten das Vertrauen und die Mitarbeit der Betroffenen zu gewinnen.

Schulaufsicht: Ich stimme dem zu, aber politisch ist das kaum zu machen. Der öffentliche Druck auf die Ministerien ist immens. Eigentlich bräuchten wir ein differenziertes Rechenschaftssystem, in dem Leistungsmessungen einen wichtigen, aber auch nur einen begrenzten Beitrag leisten können. Zudem ist zu bedenken, dass dieser Beitrag unterschiedlich ausfällt, je nachdem, ob die Messungen den Lernerfolg von Schüler/inne/n, den Lehrerfolg von Lehrer/inne/n oder die Wirksamkeit des Schulsystems insgesamt erfassen sollen. Mir kommt es darauf an, dass Fragen untersucht werden, zu denen es tatsächlich Handlungsoptionen gibt. Also müssten die Politiker/innen konkrete Entscheidungsfragen vorweg formulieren.

Außerdem finde ich es wichtig, dass die Interpretation der Ergebnisse nicht in der Hand eines einzelnen Forscherteams liegt, sondern dass durch konkurrierende Interpretationen offengelegt wird, wo die Daten in der Tat eindeutig und wo unterschiedliche Deutungen möglich sind. Es geht doch nicht an, dass die einen aus TIMSS folgern, Japan habe wegen seines problemlösenden Unterrichts so gut abgeschnitten und deshalb müsse eine anspruchsvollere Lernkultur entwickelt werden, während andere aus derselben Studie folgern, Japan habe so gut abgeschnitten, weil die privaten Nachhilfeschulen, die *Jukus*, nachmittags durch schlichtes Pauken den Lernerfolg gesichert hätten. Testprogramme sind keine Selbstläufer und Daten sprechen nicht für sich. Deshalb kommt der sozialen Kontrolle von Interpretationen und Entscheidungen eine Schlüsselfunktion zu.

Moderatorin: Ich versuche einen Ausblick. Bei Evaluation geht es offensichtlich um viel mehr als nur um Leistungstests. Vorhin war die Rede von der Pflicht zur Rechenschaft. Die betrifft doch alle:

Schüler/innen müssen sich um Klarheit über die eigenen Ziele und über den Erfolg ihrer Arbeit bemühen und in der Übernahme dieser Verantwortung gefordert und gefördert werden.

Lehrer/innen sollten Lernprozesse im Blick auf einzelne Schüler/innen und auf die Klasse insgesamt beobachten und auswerten, den eigenen Unterricht beobachten (lassen) und Erfahrungen im Gespräch mit Kollegen und Eltern durchdenken, um die

Wirkungen der eigenen Arbeit zu überprüfen. Das Kollegium sollte gemeinsam – im Blick auf die Schule als Ganze – ein Schulprogramm entwickeln und anhand fokussierter Bestandsaufnahmen (auch mithilfe von externen *critical friends*) regelmäßig fortschreiben.

Die *Schulaufsicht* sollte in den schulinternen Diskussionen den Fremdblick stärken, kritische Fragen stellen, Weiterentwicklung fordern, vor allem aber konstruktive Ansätze anerkennen und unterstützen.

Das *Ministerium* sollte mit Blick auf das Gesamtsystem allgemeine Problemstellen ausmachen, politische Prioritäten überprüfen, mehr Ressourcen bereitstellen und deren Verteilung ziel- und bedarfsorientiert organisieren.

Und noch eins: Vieles, was im Moment als Evaluation eines konkreten Systems oder einer einzelnen Maßnahme verkauft wird, ist eigentlich Grundlagenforschung. Zu den Zusammenhängen zwischen verschiedenen Faktoren im System gibt es doch schon eine Fülle von Publikationen und Daten, die man mit vergleichsweise geringerem Aufwand aufbereiten und gezielt auf aktuelle Fragen hin auswerten könnte, statt mit hohen Kosten Berge neuer Daten zu erheben. Wenn von Ökonomie die Rede ist, sollte dieser Anspruch auch für die Evaluation selbst gelten.

Schließlich: Müsste man nicht Schüler/innen, Eltern, Lehrer/innen und die Öffentlichkeit differenzierter nach ihren Ansprüchen an die Schule, nach ihren unterschiedlichen Erfahrungen und Kritikpunkten befragen? Damit würden die Wertfragen und die konkurrierenden Perspektiven auf Unterricht deutlicher, und man könnte den Blick zugleich auf unzulängliche Bedingungen der schulischen Arbeit erweitern. Es geht doch nicht nur darum, was *ist*, sondern auch darum, was sein *soll* – und diese Fragen sind normativ zu diskutieren.

Epilog des Autors: In dieser kurzen Diskussion sind fünf Schlüsselfragen deutlich geworden:

- Welchen Geltungsanspruch verbinden Forscher/innen mit »Evidenzbasierung« und wie kommunizieren sie ihn?
- In welcher Hinsicht ist dieser Anspruch in einzelnen Studien gerechtfertigt, in welcher nicht?
- Welche forschungsmethodischen Alternativen gibt es?
- Wo überziehen Interpreten in der öffentlichen Diskussion die Aussagekraft der Befunde?
- Welche eigentlich drängenden Fragen im Umfeld der Studien werden von ihnen nicht beantwortet?

Diese Fragen sollen in den kommenden Kapiteln vertieft werden.

2. Über das Spiel mit Zahlen hinaus[3] – Grundprobleme einer »Evidenzbasierung«

Im Herbst 2013 stellte die FDP im Landtag von NRW und in der Hamburger Bürgerschaft den Antrag, die Methode »Lesen durch Schreiben« für den Anfangsunterricht zu verbieten. Begründung: Die Forschung habe erwiesen, dass sie schädlich für die Entwicklung der Rechtschreibfähigkeit sei. »Evidenzbasiert« nennen sich solche bildungspolitischen Vorstöße. Unabhängig davon, ob die unterstellte Deutung der Forschungslage zutrifft, stellen sich einige grundsätzliche Fragen zu diesem Vorgehen.

Nach der Methode »Lesen durch Schreiben« verfassen Kinder von Anfang an eigene Texte. Sie verschriften Wörter, indem sie sich diese Laut für Laut vorsprechen und den Lauten passende Buchstaben zuordnen (»FATA KOMT«) – anfangs mithilfe einer Tabelle, in der diese durch Anlautbilder erschlossen werden können. So gewinnen die Kinder grundlegende Einsichten in die alphabetische Struktur des Schriftsystems. Orthografische Modelle und Anforderungen spielen im Unterricht erst später eine Rolle.

Geht man in die Schulen, findet man in der Tat »Lesen-durch-Schreiben«-Klassen, in denen Kinder gegen Ende der Grundschulzeit erhebliche Schwierigkeiten mit der Rechtschreibung haben. Aber man findet auch viele Lehrer/innen, die von der Methode und ihrer Wirksamkeit überzeugt sind und deren Klassen bei Rechtschreibvergleichen überdurchschnittlich abschneiden. Soll, ja kann man diesen Lehrer/inne/n verbieten, weiter nach »Lesen durch Schreiben« zu unterrichten? Festgestellt wurde ja allenfalls ein Mittelunterschied.

Anders gefragt: Gelten Erfolge in Einzelsituationen und persönliche Erfahrungen nichts mehr, wenn Großstudien mit standardisierten Tests zu dem Ergebnis kommen, dass andere Methoden *im Durchschnitt* erfolgreicher sind?

2.1 Sind Forschungsergebnisse »evident«?

»Offenkundig«, »klar«, unbestreitbar« – in diesem Sinne wird leicht missverstanden, was als Ergebnis empirischer Forschung in die Diskussion von Alltagsproblemen eingebracht wird. Dabei vermischen sich verschiedene Bedeutungsebenen des Begriffs:

---

3 Der Titel dieses Kapitels ist eine dankbare Erinnerung an den gleichnamigen Sammelband von Hamilton et al. (1977): Beyond the numbers game. Macmillan: London (→ »Beyond the numbers game« 2008), und damit an meine Kolleginnen und Kollegen in Norwich/England und Urbana-Champaign/USA, bei denen ich Anfang der 70er Jahre Theorie und Handwerk der Evaluation lernen durfte – nicht als Technik, sondern als eine soziale Praxis, deren Probleme und Risiken sie sehr differenziert durchdacht haben. Leider ist der damals erreichte Reflexionsstand 40 Jahre später nicht mehr selbstverständlich.

»Evidenz« als ein anderes Wort für »empirische, auf die systematische Erhebung und Auswertung von Daten gestützte Belege« (Beywl/Zierer 2013, XVII) und »evident« als inhaltliche Wertung der gewonnenen Erkenntnis. Für den Bildungsbereich nennt Altrichter (2010, S. 220) jene Entscheidungen evidenzbasiert, die »auf der Basis von geprüften Informationen fallen und in ihrer Umsetzung empirisch evaluiert werden«. Den Anspruch von verbindlicher Geltung oder gar Offensichtlichkeit verbinden seriöse Forscher/innen nicht damit.

In der Medizin werden schon länger bestimmte Formen der Forschung propagiert, deren Befunde als forschungsmethodisch abgesicherte »Evidenz« das ärztliche Handeln und insbesondere den Einsatz von Medikamenten auf ein höheres Niveau heben sollen: experimentell angelegte Längsschnittstudien* mit zufällig ausgewählten Versuchs- und Kontrollgruppen, die doppelt-blind durchgeführt werden, das heißt weder Mediziner noch Patienten (und teilweise nicht einmal die Forscher) wissen, wem das zu überprüfende Medikament oder ein Placebo verabreicht wurde.

Dieser sogenannte Goldstandard der Forschungsmethodik soll nun auch in der Pädagogik gelten. Denn es gibt skeptische Rückfragen an die bisherige Schul- und Unterrichtsforschung, wie sie z. B. Biesta (2011, S. 95) in seinem Beitrag über »evidenzbasierte Bildungsforschung« referiert (auch wenn er breit verwendet wird, ist dieser Begriff eigentlich unsinnig: Evidenzbasiert können Politik, Verwaltung und Praxis handeln, Forschung *produziert* Evidenz). Biesta verweist auf »ernsthafte Zweifel an der Qualität und Relevanz der Bildungsforschung. (...) dass die Bildungsforschung keine Antworten auf Fragen gebe, die von staatlicher Seite an die Entwicklung bildungspolitischer Programme gestellt werden, dass sie professionellen Pädagogen keine klaren Richtlinien für ihre Arbeit zur Verfügung stelle, dass sie bruchstückhaft, nicht-kumulativ und methodisch fehlerhaft sowie oftmals tendenziös und politisch motiviert sei«. Und programmatisch heißt es im Rahmenprogramm zur Förderung der empirischen Bildungsforschung (BMBF 2008, S. 7):

> »Seit den ›Konstanzer Beschlüssen‹ nimmt Deutschland wieder an internationalen Vergleichsuntersuchungen teil. Die dadurch politisch eingeleitete sogenannte ›empirische Wende‹ in der Bildungspolitik setzt zwingend auch eine ›empirische Wende‹ in der Bildungsforschung voraus, weil sonst die Erkenntnisse und Instrumente nicht verfügbar sind, die für eine evidenzbasierte Steuerung benötigt werden. Allerdings setzt umgekehrt die konsequente Entwicklung der empirischen Bildungsforschung die systematische Etablierung einer evidenzbasierten Systemsteuerung auf den unterschiedlichen Ebenen des Bildungssystems voraus, einschließlich eines darauf bezogenen Qualitätsmanagements.«

Einerseits soll durch die »empirische Wende« das Niveau der Forschung selbst angehoben werden. Andererseits scheint das Paradigma deshalb attraktiv, weil es eine Alternative verspricht zur Fehleranfälligkeit der Alltagserfahrung, die situationsbedingt variieren kann und zudem personenabhängig wahrgenommen und gedeutet wird.

Denn was Lehrerin A in ihrer Schule gelingt, muss als Methode nicht an anderer Stelle genauso erfolgreich sein: im Rahmen einer anderen Pädagogik, mit weniger engagierten oder kompetenten Kollegen, mit einer anderen Schülerschaft, bei schlechterer Ausstattung der Schule und ...und ...und. Es gibt viele Bedingungen, die den Transfer von Ideen beeinträchtigen können. Ein erfolgreicher Einzelfall kann keinen Modellcharakter beanspruchen, er kann lediglich Potenziale aufzeigen. Insofern erscheint es sinnvoll, neue Konzepte oder Methoden unter vielfältigen Bedingungen zu prüfen und zu versuchen, den Einfluss der Randbedingungen so weit wie möglich zu kontrollieren.

Zu kontrollieren versucht man in diesem Paradigma auch einen zweiten Aspekt: die Subjektivität der Wahrnehmung. Das Beispiel der viel diskutierten Zifernnoten zeigt, wie heikel es sein kann, sich auf das persönliche Urteil von Praktikerinnen zu verlassen (s. Kap. 6.1). Leistungen in anderen Bereichen, Sympathie, unterschiedliche Kriterien für die Beurteilung und deren mehr oder weniger strenge Anwendung beeinflussen, wie eine Prüfung bewertet wird.

2.2 Von der Medizin lernen?

Es scheint also plausibel, diese Person- und Situationsabhängigkeit praktischer Erfahrung zu überwinden, indem man Daten über standardisierte Instrumente erhebt – und dies in großen, möglichst nach dem Zufallsprinzip ausgewählten Stichproben*. So wie man es aus der Medizin kennt, die häufig als Vorbild für eine Professionalisierung der Pädagogik gepriesen wird.

Ärzte und Schulreformer teilen ein Problem: Sie müssen aus beobachtbaren Symptomen erschließen, ob ihr »Patient« wirklich krank ist und welche tiefer liegenden Störungen es sind, die sein Unwohlsein verursachen. Erst aus einer umfassenden Diagnose können sie wirksame Maßnahmen zur Überwindung der Krankheit entwickeln.

Die meisten Patienten fühlen selbst, ob sie Fieber haben oder nicht. Thermometer sind aber ein wichtiges Hilfsmittel, um präziser festzustellen, ob die Körpertemperatur dem Standard entspricht oder tatsächlich erhöht ist. Für Ärzte ist sie ein bedeutsamer Krankheitsindikator – interpretierbar allerdings nur im Kontext weiterer Symptome. Denn die Grundtemperatur verschiedener Menschen ist unterschiedlich. Manche bekommen sogar, wenn sie sehr krank sind, kein Fieber. Bei anderen erhöht sich die Temperatur schon bei leichter Anstrengung. Das heißt: Fieber ist nur ein *Symptom* – und es ist nur *ein* Symptom.

Ärzte messen deshalb nicht nur die Temperatur, sie erheben je nach Konstellation eine Fülle weiterer Daten, um mögliche Ursachen für die erhöhte Temperatur festzustellen. Gute Ärzte verzichten im Regelfall darauf, Fieber senkende Medikamente zu verschreiben, da diese nur die störenden Symptome möglichst rasch verschwinden lassen. Oft könnte dies sogar riskant sein, weil eine nur oberflächliche Anpassung an die Standardtemperatur verhindert, die eigentliche Krankheit wahrzunehmen.

Auch in der Pädagogik nutzen wir »Thermometer«, besser bekannt unter dem Namen »Leistungstests«. Aber auch hier gilt: Fieber zu messen, ist noch keine Diagnose – und Fieber zu senken, keine Therapie. Leistungstests können ein wichtiges Hilfsmittel sein, um Problemen des Bildungssystems auf die Spur zu kommen. Aber nicht jede erhöhte Temperatur ist Anzeichen für ernste Probleme. Für die Frage, ob das Bildungswesen insgesamt oder ob eine Schule »gesund« ist, bedarf es einer Fülle weiterer Indikatoren. Umgekehrt muss eine Annäherung der Testergebnisse an vorgegebene Standards nicht bedeuten, dass das untersuchte System wirklich besser geworden ist, gibt es doch auch in der Pädagogik fiebersenkende Medikamente. Das bekannteste wurde dank PISA und Co. aus den USA auch nach Deutschland importiert: Teaching to the Test. Verlage werben seit Jahren mit Mottos wie »PISA-fit« für Arbeitshefte, in denen Aufgabentypen aus den großen Leistungsstudien zum Modell für Übungsaufgaben generell werden.

Dass man mit verschiedenen Thermometern unterschiedliche Werte bekommt, sollte zwar nicht sein, wird aber innerhalb enger Bandbreiten akzeptiert. Eine Eichung fällt aufgrund der technischen Normierbarkeit vergleichsweise leicht. In der Bildungsforschung gibt es dagegen erhebliche Messunterschiede – je nachdem, welchen Test man einsetzt. Bei PISA-2006 hat sich sogar die für Mediziner erstaunliche Situation ergeben, dass die deutschen und die internationalen Autoren an demselben »Thermometer« unterschiedliche Temperaturen ablasen. Spätestens hier wird deutlich, wie naiv diejenigen sind, die glauben, man könne psychische Merkmale mit vergleichbarer Genauigkeit messen wie physische. Tests sind heikle Instrumente und deshalb noch vorsichtiger zu handhaben als ein Quecksilber-Thermometer. Jede Beobachtung, jede Befragung, jeder Test erfasst nur Indikatoren für das eigentlich Gesuchte. Die Bedeutung der jeweils erfassten Daten muss deshalb interpretativ erschlossen werden.

In den letzten Jahren wurde viel von der empirischen Wende der Erziehungswissenschaft gesprochen. Übersehen wird dabei oft, dass Heinrich Roth (1962) schon vor über 50 Jahren eine »realistische Wende« eingeläutet und dafür *auch* eine stärkere Nutzung kontrollierter Forschungsdesigns und standardisierter Untersuchungsverfahren gefordert hat. Dass die Pädagogik einige schmerzhafte Erfahrungen mit der einseitigen Orientierung an diesem Forschungsstil gemacht hat und dass andererseits auch qualitative Daten empirisch gewonnen werden, wurde oft vergessen. Es lohnt sich also, genauer hinzuschauen, was Bildungsforschung für Lehrer/innen leistet – und was umgekehrt diese zur Bildungsforschung beitragen können, ja müssen, damit sie für den Unterrichtsalltag nützlich werden kann. Denn auch aus der Medizin kommen warnende Stimmen:

> »Der Hausarzt kennt seine Pappenheimer. Er weiß genau, was ihnen fehlt. Frau Müller zum Beispiel. Die ist übergewichtig, hat ein Schilddrüsenproblem, Bluthochdruck und nimmt die Pille. Der Fall ist unübersichtlich, und der Doktor hadert mit der Therapie für die Patientin – deshalb sucht er Rat bei der Wissenschaft. Frau Müller aber kommt in den fernen Universitäten nicht vor. (...) Beim Versuch, die Statistik immer wasserdichter zu machen, geht der wirkliche Patient öfter verloren. (...) [Die Studien] werden in Kliniken durchgeführt und nicht in Praxen, dort sind Mehrfachkranke wie Frau Müller die Regel, nicht die Ausnahme.« (Albrecht 2005, S. 25)

In der Medizin hat der Anspruch der Evidenzbasierung eine etwas längere Geschichte als in der Pädagogik. Seit seiner Einführung Anfang der 1990er Jahre ist er intensiv und kritisch diskutiert und in wichtigen Punkten relativiert worden. So heißt es in der unter Medizinern viel zitierten Definition von Sackett u.a. (2000, S. 1): »Evidenzbasierte Medizin (EBM) bedeutet die Verbindung der besten Forschungsbefunde mit klinischer Erfahrung und den Werten des Patienten (...)« *(Übersetzung des Autors).*

Die Verantwortung für die Abwägung hat danach der Arzt vor Ort. Dessen praktische Erfahrung, aber auch die Sichtweise des Patienten stehen gleichrangig neben den Forschungsbefunden. Dies ist eine wichtige Relativierung des Geltungsanspruchs von Ergebnissen aus experimentellen Studien und Reihenuntersuchungen, die sich zunehmend auch in der Bildungsforschung findet. So gibt es entschiedene Stellungnahmen der zuständigen Fachgesellschaften gegen eine Bevorzugung sogenannter quantitativer Ansätze als Gold-, Silber- und Bronzestandard (z.B. von der AERA 2003).

Im Blick auf den dennoch vielfach propagierten Goldstandard stellt sich vor allem die Frage, welche Folgen die Kontrolle pädagogischer Interventionen und ihrer Randbedingungen sowie die methodische Standardisierung von Erhebung und Auswertung auf den Erkenntnisstatus der Befunde und auf ihren Nutzen für die Praxis haben. Denn nicht für alle Fragen sind experimentelle Designs der angemessene Forschungstyp; und nicht in allen Situationen sind sie machbar oder zumutbar.

2.3 Untersuchungsdesign: viele Fälle oder reiche Daten?

Entscheidungen über das »Design« einer Studie betreffen den Aufbau und die Organisation einer Untersuchung. Während es auf der methodischen Ebene um die Auswahl von Instrumenten zur Erhebung und Auswertung von Daten geht, zielt das Design auf die Auswahl der Fälle oder Stichproben* und auf die Anlage der Untersuchung, z.B. die Kontrolle von Randbedingungen in einem Experiment oder die Planung eines Längsschnitts*.

Dabei ist es in der Bildungsforschung wie im wirklichen Leben: Man kann nicht alles haben. Entweder bezieht man eine große Zahl von Fällen ein – dann kann man an diesen nur zu wenigen Aspekten Daten erheben. Oder man versucht ein breites Spektrum an Informationen zu erfassen – dann muss man sich aus forschungsökonomischen Gründen auf wenige Fälle beschränken. In der Sprache der Fotografie und des Filmens gesprochen: Weitwinkel oder Tele. Beides macht Sinn, aber beides gleichzeitig geht nicht.

Konkret bedeutet das erhebliche Einschränkungen für die Aussagekraft von Großerhebungen zu pädagogischen Interventionen – weil die Umsetzung für solche Studien eine Blackbox bleibt. Erfasst wird »Output«, z.B. die Leseleistungen der Schüler/innen am Ende eines Schuljahrs, vielleicht auch noch deren Voraussetzungen am Anfang. Aber was dazwischen passiert und wie der Unterricht konkret aussieht, bleibt im Dunkeln. Hinzu kommt, dass die konkreten Randbedingungen nicht erfasst werden. De-

ren Einfluss kann lediglich (durch eine Zufallsstichprobe) neutralisiert werden – sie werden damit aber zu einem nicht mehr erkennbaren »grauen Kontext«. Das ist auch eine Schwäche von Experimenten: Sie sind stark im Nachweis, *dass* ein Lerneffekt A die Folge einer Maßnahme B ist, aber schwach in der Erklärung, *wie* dieser Zusammenhang zustande kommt und was getan werden muss, um diese Potenziale unter den variierenden Alltagsbedingungen optimal zu entfalten.

Bei den internationalen Leistungsvergleichen – und zum Teil auch bei der Evaluation von Programmen oder Methoden – werden sogar nur die Ergebnisse am Ende verglichen. Diese können aber auch von unterschiedlichen Voraussetzungen am Anfang abhängen, etwa kognitiven Unterschieden zwischen den Schüler/inne/n oder Differenzen in deren sozialer Herkunft. Der Goldstandard fordert deshalb zu Recht, auf eine Vergleichbarkeit der Eingangsbedingungen zu achten. Aber reicht das?

Zahlen sprechen nicht für sich

In den 1950er Jahren wurde folgender Befund aus einer empirischen Studie kolportiert: Kinder aus Familien, die einen Staubsauger besaßen, lasen besser als Kinder aus Familien ohne Staubsauger. Trotz der deutlichen Korrelation* wäre aber niemand auf die Idee gekommen, Lese- und Rechtschreibschwierigkeiten (LRS) durch das Verschenken von Staubsaugern an die Eltern der betroffenen Schüler/innen anzugehen. Jeder vernünftige Mensch interpretiert den Staubsaugerbesitz als Indikator für die sozioökonomische Lage der Familie. Aber so vernünftig diese Deutung auch ist – die Daten geben sie nicht vor. Es ist unser Hintergrundwissen, das diese Interpretation trägt, nicht allein die Statistik, auf die sie sich bezieht. Zahlen sprechen nicht für sich. Auch Bildungsforscher sehen ein »Schlüsselproblem (...) in der ›Übersetzung‹ von Befunden in Evidenz: Dies ist ein Prozess der Forschungssynthese, der zugleich eine Bewertung der Belastbarkeit unterschiedlicher Arten von Forschungsergebnissen umfasst. Wir sprechen also von evaluativer Forschungssynthese, um hervorzuheben, dass es nicht bloß um eine Befundaggregation gehen kann« (Bromme u. a. 2014, S. 18).

Damit wird deutlich: Nicht nur die Alltagserfahrung von Praktikern ist personengebunden, sondern auch die angeblich so »objektive« Bildungsforschung à la PISA. Pointiert gesagt: Wenn die OECD-Studien einen Wert haben, dann weniger deshalb, weil sie mit standardisierten Instrumenten eine große Zahl von Fällen erfassen, sondern weil sich in den Forschungsteams kluge Bildungswissenschaftler mit einem fundierten Hintergrundwissen zusammengefunden haben. Dank ihrer Erfahrung wissen sie die Zahlen – aus ihrer Sicht! – sinnvoll zu deuten. Umso mehr wünschte man sich, dass schon in den Forschungsberichten selbst die Mehrdeutigkeit von Daten sichtbar gemacht würde. So könnten alternative Interpretationen explizit gegeneinandergestellt werden, um die scheinbare Expertenautorität zu relativieren.

Wie ist es beispielsweise, wenn in einer Studie herauskommt, dass ein hoher TV-Konsum mit niedrigen Leseleistungen oder mit hoher Aggressivität korreliert? Neigen wir hier nicht dazu, eine direkte Kausalität zwischen den beiden Variablen zu unterstellen? Nach dem Staubsaugerbeispiel sollten wir vorsichtig sein. Beispielsweise fand sich in der internationalen Lesestudie der IEA* von 1991 kein Beleg dafür, dass Kinder aus Ländern mit einer durchschnittlich längeren Fernsehdauer schlechter lesen als Kinder aus Ländern mit einer geringeren Fernsehdauer. Aus dem PISA-Vorzeigeland Finnland wurden sogar deutlich überdurchschnittliche Fernsehzeiten berichtet (→ »Kultur und Kontext« 1995).

Hätte man die IEA*-Studie schon Anfang der 60er Jahre durchgeführt, wäre vielleicht sogar eine positive Beziehung zwischen Fernsehdauer und Lesekompetenz festgestellt worden. Aber daraus zu folgern, Fernsehen fördere die Lesefähigkeit, wäre genauso kurzschlüssig gewesen wie eine Verteilung von Staubsaugern an LRS-Kinder. Denn möglicherweise waren seinerzeit sowohl Staubsauger- als auch Fernseherbesitz Indikatoren eines dritten Faktors: des sozio-ökonomischen Status der Familie – sagt unser Hintergrundwissen.

Dieses Hintergrundwissen hilft uns oft, die Widersprüche zwischen Daten aus unterschiedlichen Zeiten aufzulösen. Fernsehen am Anfang der 60er Jahre bedeutet etwas anderes als Fernsehen Anfang der 90er. Befunde humanwissenschaftlicher Forschung haben ein schnelles Verfallsdatum. Nicht nur, weil die ältere Forschung methodisch weniger raffiniert war, sondern auch, weil sich der Gegenstand selbst verändert. Man denke nur an die Studien zu Geschlechterdifferenzen in sozialen Interaktionen aus den 70er Jahren. Diese lassen sich nicht unbesehen auf die heutige Situation übertragen, da sich die gesellschaftlichen Normen, aber auch die tatsächlichen Wahrnehmungsmuster und vor allem das Verhalten selbst erheblich verändert haben – auch als Folge der viel diskutierten Befunde solcher Studien.

Und aus mehreren Gründen ist auch die Interpretation von Daten fragil: Die befragten, beobachteten oder getesteten Personen können dieselbe Aufgabe unterschiedlich verstehen, die Auswerter/innen klassifizieren die Lösung, ohne diesen Hintergrund zu kennen, die statistische Verrechnung kann nach unterschiedlichen Verfahren durchgeführt und die Ergebnisse können unterschiedlich dargestellt werden.

Diese Einsicht hat durchaus auch gesellschaftspolitische Bedeutung. Wenn die Kriminalitätsrate unter Menschen anderer Staatsbürgerschaft höher ist als unter denen mit deutschem Pass, wird schnell die Parole laut »Ausländer raus«. Wenn man andere Faktoren einbezieht, zeigt sich aber, dass Ausländer beispielsweise häufiger unter erschwerten Bedingungen leben und dass der Anteil Jugendlicher unter ihnen überproportional groß ist – beides Teilgruppen, in denen die Rate der Rechtsverletzungen generell höher ist. Solche Bedingungen kann man wiederum nur kontrollieren, wenn man sie auch erhoben hat. Das heißt, man muss schon vor Beginn der Studie wissen oder zumindest vermuten, welche Drittfaktoren bedeutsam sein können – und diese müssen messbar sein. Theoretische Annahmen und ihre Gewichtung spielen also eine große Rolle bei der Anlage einer Studie und damit bei der Möglichkeit, die Daten differenziert zu interpretieren.

Aussagen der Bildungsforschung gelten nur kontextbezogen

Wie gezeigt, verändert sich in den Humanwissenschaften der Gegenstand auch durch die Forschung selbst. Schon in den Untersuchungen können Beobachtung und Befragung das Verhalten der Betroffenen beeinflussen. Zudem entziehen manche Forscher/innen ihren Befunden selbst die Geltungsgrundlage – infolge der Veröffentlichung ihrer Ergebnisse und deren Wirkung auf die Handelnden.

Dazu ein konkretes Beispiel: Als Bildungsforscher herausfanden, dass der Umfang der Buchstabenkenntnis vor Schulbeginn die Leseleistung nach ein, zwei Schuljahren am besten vorhersagt, fingen Eltern und Kindergärten an, Kinder im Aufsagen des ABC und im Benennen bzw. Wiedererkennen der Buchstaben zu trainieren. Leider blieb der erwartete Erfolg aus.

Wie der Staubsauger in der Soziologie ist nämlich auch die Buchstabenkenntnis nur ein Indikator. Zur Zeit der Erstuntersuchung konnte man an ihr festmachen, wie umfangreich die Schrifterfahrung der Kinder und wie hoch ihr Niveau bei der kognitiven Verarbeitung dieser Erfahrung war. Das anschließend propagierte Training des Verhaltens vermittelte aber genau diese Erfahrungen nicht. Die kognitiven Tiefenstrukturen blieben meist unberührt. Damit änderte sich die Bedeutung des Symptoms – und sein Wert als Prädiktor des schulischen Schriftspracherwerbs. Entscheidend ist nämlich, ob die Buchstabenkenntnis Folge eines aktiven Umgangs mit Schrift ist oder ob sie bloß antrainiert wurde. Diese Tiefenstruktur lässt sich am Oberflächenverhalten in einem Test aber nicht einfach ablesen. Sie muss erschlossen werden – getragen von den Theorien der Interpreten.

In der Sozialforschung haben wir also mit Reaktionen auf unsere Untersuchungsaktivitäten zu rechnen – kurzfristig, indem die Beteiligten unsere Fragen oder Aufgaben aus ihrer individuellen Sicht interpretieren, langfristig, indem sie ihr Verhalten ändern, in der Schule z. B. durch ein Teaching to the Test. Menschen reagieren nicht mechanisch, sondern mit Berechnung, aber genau deshalb nicht berechenbar.

Können wir uns denn wenigstens für zehn oder fünf Jahre auf die Gültigkeit empirischer Befunde verlassen? Leider selbst das nicht.

Noch einmal zurück zum Lese- und Fernsehverhalten von Kindern. Die IEA*-Studie von 1991 hat nicht nur Unterschiede im TV-Konsum zwischen Ländern, sondern auch innerhalb einzelner Länder mit der Leseleistung korreliert. Das Ergebnis war besonders irritierend. Es gab Länder, in denen die Leseleistung mit wachsender Fernsehhäufigkeit abnahm, und andere, in denen Kinder umso besser lasen, je häufiger sie fernsahen.

Zum Teil ist das mit dem unterschiedlichen Entwicklungsstand der jeweiligen Gesellschaften erklärbar. Schwellenländer waren in ökonomisch und kulturell vergleichbarer Situation wie Industrieländer 30 oder 50 Jahre vorher. Sozioökonomisch hatte der Fernseherbesitz also eine unterschiedliche Bedeutung.

Aber es sind noch andere Erklärungen denkbar und nötig, wenn man an das oben zitierte Beispiel Finnland denkt. Denn dort werden – anders als in Deutschland und in den USA – viele Filme in der Originalsprache und nur mit Untertiteln in der Natio-

nalsprache gezeigt. Es lohnt sich also für Kinder, lesen zu lernen, um die Filme besser zu verstehen.

Zugleich bietet ihnen das Fernsehen eine effektive Lerngelegenheit, wie Laborexperimente gezeigt haben: Rasches Lesen unter hoch motivierenden Umständen und gestützt von einem Bildkontext fördert das Lesenlernen jüngerer Kinder. Andere Studien legen außerdem nahe, dass auch die Inhalte der Filme einen Unterschied machen – und vor allem die Situation, in der die Kinder fernsehen: allein oder mit ihren Eltern, stumm oder im Gespräch. Dass dies auch für das Lösen von Tests gilt, haben die Analysen von Aufgaben in den zentralen Vergleichsarbeiten für Klasse 3 (VerA*) eindrucksvoll gezeigt (s. Kap. 5.2 und 6.2).

Die Bedeutung desselben Verhaltens wechselt, sie ist kultur-, ja sogar biografieabhängig. Dies ist ein grundsätzliches Problem psychologischer Forschung, deren Befunde zu über 95 Prozent aus westlichen Ländern stammen und zudem in den meisten Fällen aus Experimenten mit Studenten. Diese sogenannten WEIRDs (»Western, Educated, Industrialized, Rich and Democratic«) stellen eine doppelt ausgelesene Gruppe dar, die die Weltbevölkerung nicht repräsentiert. Denn wie Henrich u.a. (2010) gezeigt haben, liegen die Ergebnisse der WEIRDs in Studien, die auch andere Kulturen einbezogen haben, eher am Rande des Spektrums, ob es um Einstellungen, Wahrnehmungen oder Denkstrategien ging. Im Bildungsbereich zeigen sich bedeutsame Milieuunterschiede sowohl auf der Makroebene (Strukturen und Traditionen des Bildungssystems) als auch auf der Mikroebene (»Kultur des Klassenzimmers«). Forschungsbefunde haben deshalb nur eine begrenzte Reichweite und ein rasches Verfallsdatum.

Damit wird ein weiteres Problem der experimentell orientierten Forschung deutlich, das für pädagogische Interventionen besonders bedeutsam ist. Das als Stärke gehandelte Bemühen um Kontrolle möglichst vieler Bedingungen ist zugleich eine Schwäche, da die Differenz zwischen Forschungs- und Alltagssituation wächst. Damit entsteht ein Konflikt zwischen den Ansprüchen auf forschungsmethodische Rigorosität und auf pädagogische Relevanz, der nur schwer auszubalancieren ist. Durch die Künstlichkeit der Untersuchungssituation, etwa im psychologischen Labor oder gar in der Röhre der Neurowissenschaftler, verlieren die Ergebnisse ihre ökologische Validität: Sie lassen sich nur unter Vorbehalt auf Alltagssituationen übertragen. Aber auch der wissenschaftlich begleitete Modellversuch im Feld findet unter besonderen Bedingungen statt, die einen Transfer von Erfahrungen und Erkenntnissen erschweren. So sind es eher besonders engagierte und kompetente Lehrer/innen gewesen, die sich in den Versuchen vor dem Jahr 2000 auf die Anforderungen der Integration behinderter Kinder eingelassen haben, und bestimmte Eltern, die für ihre Kinder solche Versuche erkämpft haben. Auch heute noch dürften die (immer noch wenigen) Integrationsschulen unter diesen Aspekten nicht repräsentativ sein.

Um dem Vorwurf fehlender Übertragbarkeit zu begegnen, sind Großprojekte der letzten 15 Jahre in Regelschulen durchgeführt worden. Meist wurden Daten allerdings nur punktuell erhoben, also nicht im Längsschnitt*. Das ist problematisch, weil Be-

ziehungen, die zwischen gleichzeitig erhobenen Faktoren gefunden wurden, nicht so leicht als Wirkungszusammenhänge interpretiert werden können: Korrelation* bedeutet nicht schon Kausalität. Forschungslogisch könnte diese nur durch Experimente bewiesen werden, deren Befunde sich aber – wie gesagt – nur begrenzt auf Alltagsbedingungen übertragen lassen. In Feldstudien wiederum lassen sich Ursache-Wirkungs-Beziehungen mit Längsschnittdaten* eher plausibel machen als bei Querschnitterhebungen*. Selbst beim Vergleich von zwei Querschnitten (z.B. des Leseverhaltens und des Fernsehkonsums von 7- und 16-Jährigen) besteht die Gefahr, dass Unterschiede nicht auf einen Lernzuwachs zurückzuführen sind, sondern mit anderen Merkmalen der beiden Gruppen zusammenhängen.

Wenn der Umfang des TV-Konsums mit der Leseleistung korreliert, heißt das also nicht, dass der TV-Konsum Ursache von Leistungsproblemen sein muss, sprich: dass von der Einschränkung der Fernsehhäufigkeit eine Verbesserung der Lesekompetenz zu erwarten ist. Es kann auch sein, dass schwache Leser/innen das Medium Buch meiden und auf das Fernsehen ausweichen. Es könnte aber auch sein, dass sowohl hoher Fernsehkonsum als auch schwache Leseleistung Folge einer dritten Bedingung sind, nämlich mangelnder Anregung und Unterstützung im Elternhaus.

Wer wegen der Probleme von Querschnittvergleichen auf Längsschnitte setzt, sieht sich allerdings mit neuen Schwierigkeiten konfrontiert. Damit meine ich nicht nur die hohen forschungslogistischen Anforderungen wie den Erhalt der Stichprobe* über längere Zeit, damit eventuelle Dropouts* deren Zusammensetzung nicht systematisch verzerren. Längsschnittstudien mit wiederholten Erhebungen in derselben Population* haben auch inhaltliche Probleme. Denn es verändern sich ja nicht nur die Personen, sondern auch die Umwelt. Die Entwicklung des Medienverhaltens von 16-Jährigen in unserem Beispiel kann also nicht einfach auf ihr Verhalten mit sechs Jahren zurückbezogen werden, sondern muss berücksichtigen, dass sich parallel zu der individuellen Entwicklung auch die Umstände verändert haben, die zusätzlich einen aktuellen Einfluss auf den heutigen Buch- und TV-Konsum ausüben. Wenn 16-Jährige 2010 mehr lesen und weniger fernsehen, als sie dies im Jahre 2000 im Alter von sechs Jahren getan haben, kann dies ein Alterseffekt sein oder an einer gesellschaftlichen Veränderung der Mediengewohnheiten generell liegen.

Und noch eine Schwierigkeit ist zu beachten: Die Aussagekraft von Längsschnittstudien ist in vielen Fällen umso höher, je größer der zeitliche Abstand, bis die Effekte erhoben werden. Wenn sich eine Pädagogik – wie das eingangs erwähnte »Lesen durch Schreiben« – die Förderung von Selbstständigkeit auf die Fahnen geschrieben hat, interessiert nicht nur, ob die Schüler/innen im Unterricht selbstständig arbeiten, sondern auch, ob sie das noch nach der Schulzeit tun: im Beruf, zu Hause, bei ihrer Teilnahme am öffentlichen Leben. Solche Absolventenstudien haben nicht nur das oben bereits erwähnte Problem, die Repräsentativität der Stichprobe über zehn, zwanzig Jahre zu erhalten. Sie können auch kaum mehr den Einfluss von Drittfaktoren kontrollieren. In solchen Zeiträumen verändert sich nicht nur das gesellschaftliche, sondern auch das persönliche Umfeld, sodass beobachtete Verhaltens-

unterschiede nicht ohne Weiteres auf Interventionen in der Schulzeit zurückgeführt werden können – und selbst wenn diese die Ursache wären, kann nicht erwartet werden, dass sie unter den veränderten Bedingungen die gleichen Effekte hätten. Erneut zeigt sich: Die Halbwertszeit von Befunden der Sozialforschung ist in unserer schnelllebigen Zeit sehr kurz.

Zoom-Dilemma in der Bildungsforschung

Die Reichweite von Befunden der Bildungsforschung ist aber 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. Wie schon gesagt: 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 pädagogische Programme ganz unterschiedlich um (→ »Veränderungen des Curriculum auf dem Weg vom Autor zum Kind« 1978). 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« (1982/2001) eine ganz andere Bedeutung als im Vier-Säulen-Modell des Spracherfahrungsansatzes (Brinkmann u. a. 2008 ff.) oder als in einem Fibellehrgang. 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 Curriculums auf dem Papier in ein Curriculum in Aktion. In den Studien, auf die sich FDP und CDU in dem Eingangsbeispiel bezogen haben, wurden die Umsetzung und ihre Randbedingungen nicht erfasst. Damit ist für ihre Verbotsanträge der Anspruch einer »Evidenzbasierung« höchst fragwürdig. Es stellt sich darüber hinaus grundsätzlich die Frage nach dem Alltagswert von Studien, die versuchen, den Einfluss der Randbedingungen möglichst weitgehend auszuklammern. Denn nach Zufall können Lehrer/innen und Schüler/innen nur im Labor einem Programm zugewiesen werden. In der Praxis entscheiden sich Pädagogen und Eltern hingegen aus bestimmten Gründen für ein didaktisches Angebot. Ihre besondere Motivation, Kompetenz oder soziale Herkunft verzerrt dann aber die Ergebnisse.

Zurzeit werden Kindergärten und Grundschulen mit Programmen zum Training der »phonologischen Bewusstheit« überschwemmt. Die Kinder werden darin trainiert, nicht nur auf die Bedeutung von Wörtern, sondern auch auf ihre Lautform zu ach-

ten. Reimen, Zerlegen von Wörtern in Einzellaute und deren anschließende Synthese sollen das Lesen- und Schreibenlernen fördern. Verwiesen wird auf Studien zur Vorhersagekraft phonologischer Testleistungen für spätere Lese- und Rechtschreibschwierigkeiten sowie auf positive Evaluationen von Trainingsprogrammen.

Im Durchschnitt lassen die Befunde aus diesen Untersuchungen die Versprechungen der Autoren plausibel erscheinen. Aber der Teufel steckt auch hier im Detail. Viele Studien wurden in den USA durchgeführt. Dort wird vielerorts noch analog zur alten Ganzheitsmethode ein »whole language approach« gepflegt, der sich in deutschen Schulen fast nirgendwo mehr findet. Bei uns lernen die Kinder unabhängig von der konkreten Methode schon relativ früh die Funktion der Buchstaben im Wort kennen.

Doch auch die in Deutschland durchgeführten Studien haben nur eine begrenze Aussagekraft. Lehrer arbeiten didaktisch-methodisch sehr unterschiedlich. Viele von ihnen lassen die Kinder eigene Wörter Laut für Laut mithilfe der bereits erwähnten Anlauttabelle verschriften. Kinder, die nach dieser Methode schreiben und lesen lernen, brauchen kein vorgängiges Phonologie-Training – sie erhalten es »on the job«. »Lernen im Gebrauch« hat die Berliner Sprachdidaktikerin Barbara Kochan (1995) dieses Phänomen genannt. Was als Vorbereitung eines – heute verpönten – naiv ganzheitlichen Unterrichts durchaus Sinn gemacht hat, ist unter veränderten Umständen Zeitvergeudung.

Ein grundlegendes Problem humanwissenschaftlicher Forschung ist: Sie kann nicht einzelne Verhaltensweisen genau erfassen und zugleich ihren Kontext einbeziehen und damit ihre Bedeutung erschließen (»Zoom-Dilemma«). Unabhängig davon stellt sich die Frage: Haben statistisch signifikante Unterschiede auch eine praktische Bedeutung? Lohnt es, wegen eines kleinen, wenn auch statistisch nicht zufälligen Nachteils Schulbücher einzustampfen, Lehrer/innen umzuschulen, in teure Fortbildungen zu investieren und Kolleg/inn/en zu verärgern, die mit einer Alternativmethode durchaus erfolgreich arbeiten? Und müsste man nicht auch Effekte der oben genannten Methoden in anderen Dimensionen als der Rechtschreibung mit bedenken: Vorteile beim Lesenlernen oder bei der Schreibmotivation beispielsweise?

Statistische Signifikanz sichert nicht inhaltliche Bedeutsamkeit

Damit sind wir bei einem zentralen Problem der Bildungsforschung – jedenfalls wie sie zurzeit in den Hochschulen und in der Forschungspolitik favorisiert wird. Generalisierbarkeit gilt als das Markenzeichen von Großuntersuchungen wie TIMSS*, PISA* und IGLU*. Aber dieses Konzept ist genauso problematisch wie die Scheinklarheit der »objektiven« Staubsauger-Daten.

Die Ergebnisse von Repräsentativstudien sind Durchschnittsaussagen mit einer erheblichen Streuung. Für politische Entscheidungen, die Tausende von Lehrern und Schülern betreffen, bieten solche Daten eine hilfreiche Grundlage. Sie minimieren das

Fehlerrisiko bei flächendeckenden Maßnahmen. Aber was ist mit Lehrern, die es mit einzelnen Klassen oder Schülern zu tun haben?

Die Großforscher konzentrieren sich auf die Frage, wie aus einer Vielzahl von Einzelfällen Allgemeinaussagen zu gewinnen sind. Dabei ist die Verdichtung von Informationen noch relativ einfach: Man erhält Mittelwerte*, Korrelationen* und andere Kennwerte für die Stichprobe*, und man kann die Ergebnisse nach Untergruppen differenzieren. Deren Nutzen für diejenigen, die entscheiden und handeln müssen, hängt aber wesentlich davon ab, wie weit die Einzelwerte um den Mittelwert* streuen.

Beim Zusammenhang zwischen zwei Faktoren kommt es auf die Stärke der Korrelation* an – nicht nur darauf, ob sie statistisch »signifikant« ist. Die Signifikanzprüfung* hilft nur einzuschätzen, wie hoch das Risiko ist, Ergebnisse aus der untersuchten Stichprobe auf andere Stichproben aus derselben Grundgesamtheit* zu übertragen. Über die Stärke von Zusammenhängen sagt sie allein nichts aus.

In der viel zitierten Schulstudie SCHOLASTIK des Max-Planck-Instituts für Psychologie in München (Weinert/Helmke 1997) werden schon Korrelationen von r = .30 als inhaltlich bedeutsam interpretiert. Sie sind aber nur wegen der großen Stichprobe »signifikant«, ihre inhaltliche Bedeutsamkeit ist damit noch nicht bewiesen. Abgesehen davon signalisieren selbst Werte von r = .60 eine eher lockere Beziehung. Bei Fallanalysen fanden die SCHOLASTIK-Forscher »bizarre Merkmalsprofile« erfolgreicher Lehrer/innen in den untersuchten Dimensionen von Unterricht. Das aber hinderte sie nicht an folgendem Pauschalurteil: »Zum ›Entsetzen vieler Reformpädagogen‹, wie Professor Weinert sagt, erwiesen sich jene Lehrer als überdurchschnittlich erfolgreich, die einen zielgerichteten und straff strukturierten Unterricht bevorzugen und sich nicht in erster Linie als Betreuer autonomer Lerngruppen begreifen. Letztere nennen ihren Stil gemeinhin ›offenen Unterricht‹ ...« – so die Zusammenfassung von Heike Schmoll in der Frankfurter Allgemeinen vom 9.6.1997.

Obwohl die statistischen Trendwerte – wie in diesem Fall zugunsten einer straffen Unterrichtsführung – recht gering ausfallen, scheint ihr Sog stärker zu sein als die Zentrifugalkräfte der individuellen Streuungen um diese Mittelwerte* herum. Wen wundert es, wenn angesichts solcher Kunstfehler in der Forschungsdiskussion auch Medien und Politiker den Kredit statistischer Aussagen oft weit überziehen.

Es reicht eben nicht aus, für die Lerneffekte Unterschiede zwischen den Mittelwerten* verschiedener Interventionen nachzuweisen. Man muss sich auch die Streuung innerhalb eines Programms oder einer Methode anschauen. Durch die Konzept- und Kontextabhängigkeit der Umsetzung ist sie meist so groß, dass sich die Verteilungen der Effekte beider Gruppen erheblich überlappen. Selbst wenn Intervention A im Durchschnitt weniger erfolgreich ist als Intervention B, wird es viele A-Klassen geben, die oberhalb des B-Durchschnitts liegen – auch wenn die Mittelwertunterschiede im statistischen Sinne »signifikant« sind. Und je größer die Stichproben, desto wahrscheinlicher erhält man auch bei kleinen Mittelwertunterschieden eine »Signifikanz*« im statistischen Sinne. Statistische Signifikanz bedeutet deshalb nicht, dass die gefundenen Unterschiede auch in der Sache »signifikant«, das heißt praktisch relevant sind.

Wenn Schüler/innen nach einem Unterricht mit Lehrmethode A auf 100 Wörter im Durchschnitt 12,2 Rechtschreibfehler machen, bei Methode B aber nur 11,8 Fehler, dann wird selbst so ein kleiner Unterschied bei zureichend großen Stichproben statistisch signifikant. Kurz: Statistische Signifikanz heißt nicht schon inhaltliche Bedeutsamkeit.

Seriöse Statistiker/innen kennen die Grenzen ihrer Kennwerte und weisen diese aus. Als weiteres Maß werden deshalb oft Effektstärken* berichtet. Sie berücksichtigen nicht nur, wie weit die Mittelwerte* auseinander liegen, sondern auch, wie stark sich die Werte der verglichenen Gruppen überlappen: Je weniger sich die Verteilungen decken, desto höher die Effektstärke*. In sogenannten Meta-Analysen wird dieses Maß genutzt, um die Wirkung von Methoden über verschiedene Studien hinweg zu vergleichen. Im Forschungsüberblick von Hattie (2013) wird eine Effektstärke* von 0.4 als Untergrenze für bedeutsame Unterschiede angenommen. Aber was bedeutet dieser Wert? Zur Veranschaulichung wird er inzwischen oft in Schuljahre bzw. deren Anteile übersetzt. Dass eine Gruppe einen Vorsprung von einem halben Schuljahr habe, macht die entsprechende Mittelwertdifferenz verständlicher. Nur ist auch diese Aussage interpretationsbedürftig. Im Lesen macht eine Leistungsdifferenz von einem halben Schuljahr auf Klassenstufe 1 einen großen Unterschied in der tatsächlichen Kompetenz. Aber kann ein durchschnittlicher 16-Jähriger wirklich bedeutsam besser lesen als ein durchschnittlicher 15-Jähriger? In anderen Fächern sind – je nach inhaltlichen Schwerpunkten des Lehrplans über die betreffende Zeit hinweg – ebenfalls mal größere oder kleinere oder eben gar keine Fortschritte zu erwarten.

Bei all diesen Diskussionen über technische Details wird aber oft eine grundlegende Frage übersehen: Was ist überhaupt die Einheit, die in einer Studie verrechnet wird? Die von den Medien und den politischen Parteien im Sommer 2013 hochgespielte Studie von Steinig u. a. (2009; 2014) zu den Rechtschreibleistungen von Viertklässler/innen 1972, 2002 und 2012 bezog für jeden Termin ein paar hundert Schüler/innen ein. Nach Klassen gezählt sind dies aber nur ein Dutzend »Einheiten«. Bedenkt man, wie leicht zu einem Termin zwei bis drei besonders kompetente Lehrer/innen den Durchschnitt nach oben (oder im entgegengesetzten Fall nach unten) drücken können, wird deutlich, dass man eine große Zahl an Klassen (und nicht nur an Schüler/inne/n) braucht, um zumindest statistisch signifikante Unterschiede zu erhalten. Auch ist es erst mit einer großen Zahl von Klassen möglich, Mehrebenen-Analysen durchzuführen, über die Methodeneinflüsse bzw. -wirkungen nach Lehrer- und Schülerebene aufgeschlüsselt werden können. Da das Konzept der einzelnen Lehrperson und die Art der Umsetzung so bedeutsam sind für die Effekte einer Methode, muss man also genau hinschauen, welche Einheiten in den statistischen Kennwerten verrechnet wurden. Unabhängig davon bleibt das folgende Problem:

Mittelwerte helfen wenig bei Fallentscheidungen

In unserem Beispiel »Ausländer raus« (s. S. 37) spielten Unterschiede in der durchschnittlichen Kriminalitätsrate eine Rolle. Aber darf man aus Unterschieden in den Mittelwerten* auf die Gruppe als Ganze verallgemeinern? Betrachten wir einmal die Gruppe der »jungen Männer« in unserer Gesellschaft: Unter ihnen ist die Kriminalitätsrate im Durchschnitt höher als unter Männern über 25 und bei Frauen generell. Würde deshalb jemand auf die Idee kommen zu fordern »junge Männer raus«?

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 ihnen nicht folgt. Ob sie aber auf den Einzelfall zutreffen, kann man vorher nicht sagen. Die einzelne Lehrerin 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 (vgl. die anschaulichen Beispiele in DiSalvo 2014). 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 PISA und SCHOLASTIK nicht. Aber gerade für Entscheidungen dieser Art brauchen und erwarten Pädagogen Hilfe von der Forschung.

Um einen Überblick über die verschiedenen Aspekte von Unterricht bemüht sich die Meta-Meta-Analyse von Hattie (2013), in der für einzelne Methoden wie »Gruppenarbeit«, »direkte Instruktion« oder »entdeckendes Lernen« Befunde aus den jeweiligen Meta-Analysen vieler Einzelstudien verglichen werden – Analysen dritter Ordnung sozusagen (s. Kap. 3). Aber mit der Verrechnung von Mittel- und Streuungswerten aus so vielen Untersuchungen verschärfen sich die oben genannten Probleme des »grauen Kontexts«: Differenzen zwischen Studien und zwischen zusammenfassenden Meta-Analysen verschiedener Studien verschwinden in den Mittelwerten* höherer Ordnung – und damit auch die Gründe für diese Unterschiede. Bei Hattie werden sie über seine statistische Meta-Meta-Analyse als forschungsmethodische »Schmutzeffekte« neutralisiert, für Praktiker/innen geraten damit aber Bedingungen aus dem Blick, die für ihre spezifische Handlungssituation hoch bedeutsam sein können.

Noch schwieriger ist die Analyse von Lernentwicklungen. Verändern sich die Leistungen der Experimentalgruppe in einer Studie überdurchschnittlich, wird dieser Lernzuwachs gern als Ausweis für die Qualität des untersuchten Programms interpretiert. Wenn eine Gruppe im Diktat durchschnittlich nur noch fünf Fehler statt sieben macht, können sich hinter dieser Verbesserung aber unterschiedliche individuelle Entwicklungen verbergen: Manche Kinder machen nur noch zwei Fehler statt vorher acht, andere stagnieren und wieder andere machen mehr Fehler als bisher. Diese individuellen Unterschiede sind für Lehrer das entscheidende Merkmal, wenn sie einzelne Kinder gezielt fördern wollen.

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 anzuwenden. Die Einsicht, dass die Anwendung allgemeiner Aussagen auf einen spezifischen Fall das zentrale Problem sowohl bei Normen als auch bei empirischen Befunden ist, findet in der Erziehungswissenschaft und vor allem in der Rezeption durch Medien wie auch durch Politiker/innen zu wenig Aufmerksamkeit.

Jurist/inn/en greifen bei der Lösung von Konflikten nicht nur auf Gesetze, also allgemeine Normen zurück, sondern auch auf Fallentscheidungen anderer Gerichte, besonders ausgeprägt im angelsächsischen *Case Law*. Auch deutsche Gerichte argumentieren mit den Präzedenzfällen des BGH und der Oberlandesgerichte. Mediziner nutzen nicht nur experimentelle Befunde, sondern auch Fallanalysen. Hier sei nur an Klassiker wie Sigmund Freud und an die eindrucksvollen Fallgeschichten der Neurologen Oliver Sacks (1987), Alexander Lurija (1991) und Harold Klawans (2005) erinnert. Dies ist eine Forschung, in der Menschen noch als Personen erkennbar sind. Die Befunde verdeutlichen, dass Menschen bei gleichem neurologischen Befund mit ihrem Handicap ganz unterschiedlich umgehen. Einzelne Merkmale determinieren nicht die Entwicklung und das Verhalten einer Person. Menschen sind nicht berechenbar.

Die Nützlichkeit von Forschung für die Unterrichtspraxis hängt nicht davon ab, ob sie verallgemeinerbar, sondern ob sie in der aktuellen Situation nutzbar ist. Erkenntnis bewährt sich in der Anwendung auf den Einzelfall.

Induktiv-deduktive Schlüsse oder Analogien über Fallvergleiche

Damit wird ein anderer Aspekt des Forschungsdesigns angesprochen, nämlich die Frage, wie Erfahrung oder Erkenntnis aus einer Situation in eine andere übertragen werden kann. In der evidenzbasierten Bildungsforschung wird dieser Transferanspruch an den Begriff der Verallgemeinerung gebunden – eine fatale Verkürzung des Problems. Handlungsorientierte Disziplinen wie Medizin, Jurisprudenz und eben auch Pädagogik kennen und nutzen nämlich zwei verschiedene Wege, um Erfahrung oder Erkenntnis auf eine neue Situation zu übertragen (s. Abbildung): den induktiv-deduktiven Weg über Verallgemeinerungen und den Weg der Analogie von Fall zu Fall (→ »Fallstudien in der Pädagogik« 1982).

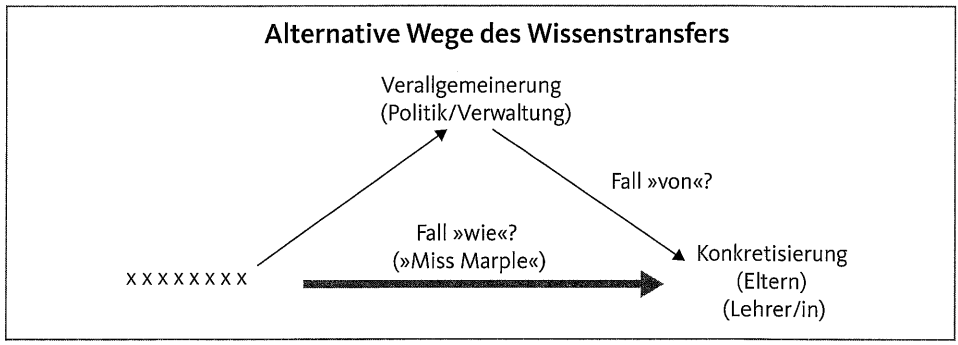

Abb. 1: Alternative Wege des Wissenstransfers

Im statistischen Modell werden Kennwerte einzelner Fälle gemittelt und in Regelform verallgemeinert. Sie sind Wahrscheinlichkeitsaussagen, das heißt, sie können keine allgemeine Gültigkeit für neue Situationen beanspruchen, aber immerhin für eine größere Zahl die Risiken einer Übertragung benennen. Da sie die zu erwartenden Häufigkeiten von Ereignissen oder Beziehungen angeben, erleichtern sie flächendeckende Entscheidungen, beispielsweise in der Bildungspolitik. Schwieriger ist die Situation von Eltern oder Lehrer/inne/n, die es mit Einzelfällen zu tun haben. Sie müssen prüfen, ob ihre konkrete Situation den abstrakt beschriebenen Mustern entspricht oder eher eine Ausnahme von der Regel darstellt. Anders als im Verhältnis von Physik und Bautechnik lassen sich Allgemeinaussagen nicht einfach *anwenden*, sie müssen kontextbezogen *interpretiert* werden. »Evidenzbasierung« schaltet also Subjektivität nicht aus. Insofern kann sie keine höhere Geltung beanspruchen als die zweite Form des Transfers: die Analogie von Fall zu Fall. Deren Unschärfen sind ähnlich, bedingt durch die Einschätzung, ob eine neue Situation in den wesentlichen Punkten einer bekannten ähnelt oder ob es bedeutsame Unterschiede gibt.

Die eine Form des Transfers ist also nicht höherwertig als die andere. Beide ergänzen sich. Wichtige Theorien sind aus sorgfältigen Fallstudien entstanden – man denke nur an die Tiefenpsychologie von Sigmund Freud (1954) und an die Entwicklungspsychologie Jean Piagets (1980).

Komplementär ist das Verhältnis der beiden Denkweisen nicht nur in der Jurisprudenz, die einerseits allgemeine Gesetze und andererseits höchstrichterliche Urteile als Bezugspunkte für Gerichtsentscheidungen nutzt, sondern auch in der Medizin. Pharmaziefirmen hilft es zu wissen, dass ein Medikament in 95 Prozent der Fälle positiv wirkt und nur in 5 Prozent heftige Nebenwirkungen hat. Für einzelne Patienten ist dagegen wichtig, ob sie zu den 95 oder zu den 5 Prozent gehören. Arzt oder Ärztin stehen vor der schwierigen Aufgabe, diese Frage zu entscheiden. Deshalb gibt es für sie – neben den Großuntersuchungen zu bestimmten Medikamenten und Therapien – seit Jahren eine Datenbank, um die Erfahrung von Praktikern zu dokumentieren und für Kollegen nutzbar zu machen:

»Jede medizinische Erkenntnis beginnt mit einem Einzelfall; jede Behandlung hat das Individuum zum Ziel. Doch in der Medizin herrscht heute die Diktatur der randomisierten kontrollierten Studie. Ärzte teilen große Patientengruppen verschiedenen Behandlungsformen zu und extrahieren aus den Ergebnissen die beste Therapie. Was nicht in diesem Verfahren erprobt ist, gilt nichts im Zeitalter der evidenzbasierten Medizin.
›Jeder Fall ist wichtig‹, schreibt dagegen der frühere Chefredakteur des British Medical Journal, Richard Smith. Seit vergangener Woche gibt er das Cases Journal heraus und verschafft damit seltenen Einzelfällen ein Forum (www.jmedicalcasereports.com). (...) Dort können künftig die Ärzte fündig werden – ähnlich wie Computerbenutzer auch bei exotischen technischen Problemen darauf hoffen dürfen, eine Lösung im Netz zu finden. Damit wird die medizinische Forschung um eine wichtige Komponente reicher: Denn im bisherigen System der evidenzbasierten Medizin wird die Erkenntnis, die ein Hausarzt in seiner Praxis aus einem vertrackten Einzelfall gewonnen hat, schlicht ignoriert. Umgekehrt sind die aus Großstudien gewonnenen Leitlinien für die Behandlung oft nur schwer auf den einzelnen Patienten im Sprechzimmer anwendbar.« (Albrecht 2008, S. 33)

In Einzelfallstudien kann mehr von dem erhalten bleiben, was man wissen muss, um einen neuen Fall zu verstehen, als im Modelldenken der abstrahierten Variablen. Die Argumentationsform der Fallstudie ist nicht induktiv-deduktiv, sondern beruht auf Analogie, auf einem Denken von Fall zu Fall. Damit kommt sie den Anforderungen des pädagogischen Alltags, aber auch dem Denkstil von Praktikern entgegen. Darum enthält etwa die Core-Hom-Datenbank homöopathischer Forschung neben Meta-Analysen von Großstudien auch Fallstudien. Fallberichte sind hilfreich, weil sie analoges Denken ermöglichen: die Deutung einer neuen Situation im Licht vergangener Fälle.

Ich habe dieses Vorgehen auch als Miss-Marple-Stil bezeichnet. In den Romanen von Agatha Christie löst Miss Marple die Kriminalfälle im Rückgriff auf ihre persönliche Alltagserfahrung mit analogen Einzelfällen: »So war das damals auch bei Onkel John, als er ...« Forschung unterscheidet sich von der Alltagserfahrung dadurch, dass sie ihre Annahmen systematisch prüft. Aber die Qualität einer Theorie kann nicht nur nach der Zahl der Fälle bestimmt werden, in der sie sich als erklärungskräftig oder vorhersagestark erweist. Sie kann und muss ihre Geltung bereichsspezifisch begründen. Damit gilt sie immer nur vorläufig.

Die Übertragung von Erfahrung auf konkrete Einzelsituationen, nicht ihre Verallgemeinerung – das ist das zentrale Problem für Praktiker. Kluge Statistiker relativieren die Versprechen ihrer Studien entsprechend. In der Forschungspraxis gewinnt aber rasch Verbindlichkeit, was allenfalls ein Mosaikstein bei der Deutung von Situationen und Verhaltensweisen sein kann.

Statistische Aussagen sind insofern nicht End-, sondern Ausgangspunkte pädagogischer Forschung – oder wie es der englische Bildungsforscher Lawrence Stenhouse (1975) schon vor 40 Jahren formuliert hat: Nicht die Fallstudien erzeugen Hypothesen für Großstudien, sondern umgekehrt liefern die Repräsentativerhebungen Folien für die Analyse von Einzelfällen – seien es einzelne Personen, Institutionen oder Programme. Sie analysieren Beziehungen zwischen wenigen Merkmalen an vielen Fällen als Einstieg in die sorgfältige Untersuchung vieler Merkmale des Einzelfalls. Ergebnisse

der Forschung über die Wirksamkeit pädagogischer Maßnahmen können nur deren Potenzial bzw. ihre Risiken und wahrscheinliche Bedingungen für ihre Realisierung benennen. Allgemeinaussagen zu ihrer Qualität stehen immer unter kontextabhängigen Vorbehalten.

Damit ist deutlich, dass die Anwendung von Allgemeinaussagen auf neue Fälle kein technischer Vorgang ist. Sie bedarf eigener Forschungskompetenz, nämlich der Fähigkeit, relevante Bedingungen für die Übertragbarkeit von Befunden und Modellen auf den konkreten Fall zu erkunden (vgl. das »Modifikationsproblem« bei Neuweg 2015, S. 36, 216 f.). Was aber tut die Bildungsforschung, was tut die Bildungspolitik, um diesen zentralen Schritt der Erkenntnisgewinnung zu unterstützen? Die Fähigkeit, Regeln situationsbezogen zu interpretieren, ist nicht schon in den Regeln enthalten, kann nicht durch sie vermittelt werden (Neuweg 2015, S. 237). Auf ihre Entwicklung müssen Ressourcen konzentriert werden, wenn sich die Qualität von Unterricht verbessern soll. Statt Evaluation zunehmend an Spezialisten außerhalb der Schule zu delegieren, muss die Evaluationskompetenz in den Schulen selbst gestärkt werden (s. Kap. 5.3).

Fehlende Offenheit für Besonderheiten »wider Erwarten«

Albrecht (2005) hat die Orientierung der medizinischen Forschung an Standardpatienten kritisiert. »Dreimal täglich eine Tablette« – so oder ähnlich stehe es in den Beipackzetteln zu Medikamenten. Kann es wirklich sein, dass die Pillen immer gleich zu dosieren sind, unabhängig von Geschlecht, Alter, Gewicht, Blutdruck, Ernährung und aktuellem Stress? Nur Hausärzte, die über viel Erfahrung verfügen, die ihre Patienten persönlich kennen und die sich Zeit für Gespräche mit ihnen nehmen, können diese Bedingungen in Rechnung stellen und Medikamente individuell dosieren.

Jeder Mensch ist anders. Leider interessiert sich die Forschung mehr für die Mittelwerte* als für die breite Streuung der individuellen Besonderheiten um sie herum. Das gilt nicht nur für die Medizin. Mühlhausen (1994) hat für die Unterrichtsplanung gefordert, sie müsse offen bleiben für Überraschungen, denn pädagogische Prozesse sind nicht in dem Maße vorhersehbar und steuerbar wie technische Abläufe. Aus der Häufigkeit von Wirkungsmustern kann man nicht deren Gültigkeit im Einzelfall folgern. Forschung und Ausbildung haben einen Nachholbedarf im Bereich kontextbezogener Fallanalysen, die auf individuelle Unterschiede aufmerksam machen und die zudem unterschiedliche Perspektiven in deren Deutung einbeziehen.

Allgemeine Normen haben auch in anderen Bereichen unserer Gesellschaft Konjunktur, zurzeit aber besonders im Schulwesen. Regelstandards, die von allen Schülern dieselbe Leistung zu demselben Termin fordern, und Kompetenztests, die landesweit abgeprüft werden, versprechen einfache Lösungen. Aber Kinder kommen mit Erfahrungs- und Entwicklungsunterschieden von drei bis vier Jahren in die Schule. Was kann da ein gleicher Unterricht für alle bewirken, wie er heute noch in vielen Schulen üblich ist – und durch die Standardisierung eher wieder zunehmen wird?

Dass Bildung, Gesundheit oder Gerechtigkeit etwas mit individuellen Besonderheiten, mit persönlichen Beziehungen zu tun haben, gerät leicht in Vergessenheit. Standardisierte Behandlung des vermessenen Menschen – eine Sackgasse nicht nur in der Medizin. Früher lernten gute Praktiker/innen aus der sorgfältigen Analyse kumulierter Einzelfälle. Im Sinne der Evidenzbasierung gilt als Rechtfertigung für eine Maßnahme, wenn sie der Mehrheit der »Bedürftigen« hilft. Was sie für Minderheiten bedeutet, denen die Maßnahme vielleicht sogar schadet, und wie man beide Gruppen bzw. die noch viel komplizierteren Unterfälle unterscheidet, verdiente die gleiche Aufmerksamkeit – in der Ausbildung und in der Forschung.

Nun bilden Großstudien wie PISA bei der Auswertung ihrer Daten durchaus Untergruppen. Leistungen werden beispielsweise nach sozialer Schicht und dem Migrationshintergrund unterschieden. Aber diese Differenzierungen werden angebunden an Oberflächenmerkmale, die objektiv erfasst werden können, beispielsweise an den Schulabschluss oder die berufliche Position der Eltern. Vernachlässigt wird die immer noch immense Streuung auch innerhalb dieser Untergruppen. Selbst die viel zitierte Zahl der Bücher im Haushalt und die Häufigkeit des Vorlesens im Vorschulalter erfassen nicht die Qualität solcher Anregungspotenziale: Intensivere Fallstudien zeigen etwa, dass es auf die Art des Vorlesens ankommt. Deren Wirkung wiederum hängt von der subjektiven Bedeutung ab, die die Beteiligten ihr zuweisen. Erlebt ein Kind das Vorlesen als Zuwendung oder als Zwang? Hier stoßen standardisierte Instrumente an prinzipielle Grenzen des naturwissenschaftlichen Paradigmas. Nicht die Häufigkeit eines Merkmals, sondern erst die Qualität seiner Ausprägung entscheidet über seine Wirkung. Qualität aber hängt von den Wahrnehmungen der Beteiligten ab und ist in ihrer Bedeutung für sie nur interpretativ zu erfassen.

Diese Anmerkungen stellen den Sinn von Großstudien nicht grundsätzlich infrage. Sie mindern aber die Wucht, mit der diese das Erfahrungswissen und die »Intelligenz der Praxis« (Reinhard Kahl) mundtot zu machen drohen. Empirische Untersuchungen liefern Argumentationsbausteine. Wie die Sachverständigen vor Gericht. Aber auch dort sind es die Richter, die die Expertenaussagen mit Blick auf die besonderen Umstände des Einzelfalls »würdigen« – genau so, wie Sackett es in seiner Definition »evidenzbasierter Medizin« vom Arzt gefordert hat.

Die Spannung zwischen Einzelfall- und Gruppenaussagen ist in der Geschichte der Wissenschaften immer wieder kontrovers diskutiert worden. Wir brauchen beides – und zwar gleichberechtigt –, wollen wir handlungsfähig werden: Vereinfachungen für den Überblick bzw. für Entscheidungen unter Zeitdruck und Reichtum im Detail, um dem Einzelfall gerecht zu werden.

Dabei ist ein grundsätzlicher Vorbehalt zu beachten, auf den Karl Popper (1982) aufmerksam gemacht hat: Selbst wenn wir Tausende von Belegen für eine Regel haben (z. B. dass Schwäne weiß seien) – aus Erfahrungen in der Vergangenheit können wir nie mit Sicherheit auf die Zukunft schließen. Schwarze Schwäne sind schon gesichtet worden – und es ist nicht auszuschließen, dass man irgendwann einem roten begegnet. Taleb (2014, S. 140 f.) veranschaulicht diese Grenze der Empirie an einer

Metapher von Bertrand Russell, dem sogenannten Truthahn-Problem: Amerikanische Truthähne machen Tag für Tag die Erfahrung, dass der Metzger kommt und sie füttert. Die statistischen Belege sprechen also eindeutig dafür, dass Metzger Truthähnen wohlgesonnen sind – bis zum Erntedankfest, an dem diese scheinbar sichere Erkenntnis durch ein einzelnes, aber existenzielles Ereignis über den Haufen geworfen wird. Analog dürfen sich auch Pädagogen nicht auf bewährte Regeln verlassen. Sie müssen darauf eingestellt sein, mit dem Unvorhergesehenen zurechtzukommen, z. B. mit »Überraschungen im Unterricht« (Mühlhausen 1994).

Von dieser Diskussion der Forschungslogik zu unterscheiden ist die Frage des methodischen Zugangs, die im Anschluss an Dilthey (1957, S. 144) gerne auf seine Formel von 1921 bezogen wird: »Die Natur erklären wir, das Seelenleben verstehen wir.« Damit kommen wir zu einer weiteren Kontroverse, die zwar häufig unter dem Etikett »quantitativ« vs. »qualitativ« mit der Design-Diskussion (Fallstudie vs. Stichprobenuntersuchung) in einen Topf gerührt wird, aber eigenständige Probleme aufwirft.

2.4 Untersuchungsmethode: eindeutige Daten oder bedeutungsvolle Informationen?

Der »Goldstandard« verlangt nicht nur eine bestimmte Anlage der Untersuchung (»Design«), er setzt auch auf bestimmte Verfahren der Erhebung und Auswertung von Daten. Besonders die vielen empirischen Befunde zur Fehleranfälligkeit des Lehrerurteils bei der Bewertung von Schülerleistungen lassen 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 (vgl. Arbeitsgruppe Primarstufe 2006 und Kap. 6.1):

- Verschiedene Lehrer/innen beurteilen dieselbe Arbeit – selbst in Mathematik – auf der üblichen 6er-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: Standardisierung. 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.

Das Problem der Mehrdeutigkeit von Verhalten

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 auseinanderfallen. 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 Klasse 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 könnte man 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: »Jetzt wissen Sie wirklich, wo ihr Kind steht!«, als die Vergleichsarbeiten VerA* in Klasse 3 flächendeckend eingeführt wurden (vgl. zu VerA Kap. 5 und 6). 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, anhand einer vergleichsweise einfachen Aufgabe. 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 ggf. 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.

Kinder können schon die Aufgabe unterschiedlich verstehen – und die Auswertung kann leicht verkennen, was sich hinter der 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. Denn »Bildung läßt sich weder auf Wissen reduzieren, noch läßt sie sich von außen erzwingen. Sie setzt die Aktivität des ›Sich-Bildens‹ voraus« (Heymann 1997, S. 8). Insofern reicht für eine bildungstheoretische Begründung von PISA & Co. nicht der Hinweis, dass Basiskompetenzen die Tür zur Allgemeinbildung öffnen. Die Qualität grundlegender schriftsprachlicher und mathematischer Kompetenzen hängt auch davon ab, *wie,* das heißt über welche konkreten Erfahrungen sie erworben wurden, in welche persönliche Gesamtkompetenz sie eingebunden sind.

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 schnitten Jungen erwartungsgemäß im Durchschnitt schlechter ab als Mädchen (vgl. die Beiträge in Richter/Brügelmann 1994, S. 110 ff., 121 ff.). In der Regel nahm 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 der 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.

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 Stu-

dien 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 (s. Kap. 3.1). So entstehen Sammelbegriffe wie »lernwegsorientierte Ansätze«, die in sich ähnlich heterogen sind wie auf der anderen Seite »lehrgangsorientierte« Ansätze. Damit zeigt sich erneut die Interpretationsbedürftigkeit von Ergebnissen – und die Frage, wer die Deutungsmacht hat bzw. welche Möglichkeiten für eine unabhängige Urteilsbildung eröffnet werden.

Wünschenswert sind weder eine Expertokratie noch eine Bürokratie; notwendig sind – im Sinne einer demokratiedienlichen 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 (s. Kap. 5.3) als auch bei der Leistungsbeurteilung (s. Kap. 6.3). Dialogische Verfahren bieten keine Sicherheit für die Richtigkeit von Deutungen, aber Teilsichten können infrage gestellt, Autoritätsansprüche gebrochen werden. Im übertragenen Sinn könnte man analog zur Physik von einer »Unschärferelation« sprechen (Bauersfeld 1972, S. 243). Keine Methode kann gleichzeitig technische Genauigkeit/Objektivität und kulturelle Bedeutsamkeit/Gültigkeit von Beobachtungen erreichen.

Wie bei der Relativierung des Geltungsanspruchs von Großstudien gilt auch hier: Tests sollte man wegen der genannten Schwächen nicht *weg*werfen. Aber Lehrer/innen und Eltern sollten sich ihren Ergebnissen auch nicht *unter*werfen. 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.

Personenabhängigkeit technischer Entscheidungen

Nun mag man meinen, mit subjektiven Einflüssen habe es ein Ende, wenn die Daten als Zahlen kodiert sind und nur noch statistisch verrechnet werden müssen. Doch auch dann stehen Entscheidungen an, in die (theoretische) Annahmen und (persönliche) Einschätzungen eingehen.

Berechnet man Korrelationen* zwischen zwei Variablen A und B, so vermutet man einen linearen Zusammenhang: je größer A, desto ausgeprägter oder (bei einer negativen Korrelation) desto weniger ausgeprägt B. Findet man keine oder nur eine niedrige Korrelation, bedeutet das aber noch nicht, dass es keinen Zusammenhang gibt. Es kann sein, dass die Instrumente nicht valide sind oder nicht sensibel genug, um die gemeinten Bedingungen zu erfassen. Abgesehen davon kann es sein, dass sich zwei gegenläufige Trends aufheben: In der einen Teilgruppe gibt es eine positive Beziehung (viel Druck steigert die Lernleistung), in der anderen eine negative (starker Druck löst Angst aus und hemmt das Lernen). Oder der Zusammenhang ist nicht linear. Wenn

man beispielsweise die tägliche Dauer des TV-Konsums auf die Lesekompetenz bezieht, kann trotz eines fehlenden Zusammenhangs im Bereich von null bis drei Fernsehstunden pro Tag die Leseleistung sprunghaft abnehmen, wenn der TV-Konsum drei oder mehr Stunden pro Tag beträgt. Solche qualitativen Sprünge sind plausibel, weil der menschliche Organismus und psychische Funktionen darauf eingestellt sind, Schwankungen in Umwelteinflüssen in begrenztem Umfang zu kompensieren.

Der Statistiker muss also entscheiden, welche Art der Beziehung er für möglich hält (z. B. graduell oder gestuft, einsinnig oder gegenläufig) – und er muss entscheiden, welche Modellierung am besten passt. Bei der Annahme qualitativer Sprünge stellt sich zudem ein weiteres Problem: Wo legt er die Schnitte auf dem Kontinuum der Leistungen, um diese in Niveaus zu gruppieren? Die Bildungsforschung interessiert sich besonders für die Teilgruppe der besonders schwachen Schüler/innen. Da Leistungen sich graduell unterscheiden, braucht man Kriterien, um Schwellenwerte festzulegen. Je nachdem, wo man sie legt, kommt man beispielsweise zu sehr unterschiedlichen Anteilen von »leseschwachen Schüler/inne/n«: Bei PISA waren es rund 25 Prozent. Die leo.-Studie hat für die Bestimmung des Anteils erwachsener Analphabeten drei Gruppen benannt, deren konkrete Kompetenz inhaltlich definiert wurde (vgl. Grotlüschen/Riekmann 2012). Dadurch wurde eine hilfreiche Differenzierung der sonst oft sehr pauschalen Urteile über Anteile von Risikogruppen ermöglicht. Die meisten Studien begnügen sich hingegen mit einem Schnitt. Und weil sich die Bevorzugung eines Schwellenwerts gegenüber anderen inhaltlich nur schwer begründen lässt, ziehen sich viele auf statistische Kriterien zurück, indem sie bestimmte Streuungen um den Mittelwert* als »normal« unterstellen und formal die unteren 15 Prozent oder die unteren 5 Prozent als »schwach« bezeichnen. Das aber sind willkürliche Entscheidungen – und sie haben eine verhängnisvolle Eigenschaft: Gleichgültig wie gut ein Bildungssystem oder eine Unterrichtsmethode ist – es gibt immer »untere« 5 Prozent bzw. 15 Prozent.

Solche Anteile verschieben sich zusätzlich, je nachdem, in welchen Einheiten Verhaltensweisen oder Leistungen vorher kodiert wurden. Nehmen wir das Beispiel der Rechtschreibleistung: Fehler in einem Text können ausgezählt werden als »falsche Wörter« oder als »falsche Grapheme«, also einzelne Buchstaben(gruppen) (im Deutschen durchschnittlich sieben pro Wort). Angenommen, in einem Diktat wurden durchschnittlich 25 von 100 Wörtern falsch geschrieben, dann beträgt der Anteil falsch geschriebener Wörter 25 Prozent, der Anteil falscher Grapheme könnte aber bei nur 3 bis 4 Prozent liegen. Wer mangelnde Rechtschreibkompetenzen öffentlichkeitswirksam beklagen will, wird einen Fehleranteil von 25 Prozent berichten – durchaus korrekt, aber eben nur eine Seite der Medaille.

In diesem Abschnitt habe ich nur einige Beispiele für Spielräume genannt, innerhalb deren Forscher/innen Entscheidungen treffen (müssen), ohne dass ihre Annahmen für den Leser der Ergebnisse immer transparent gemacht werden (können). In den vorhergehenden Abschnitten wurde schon auf andere Untersuchungsmerkmale hingewiesen, die ebenfalls von den Forscher/inne/n verändert werden können, z. B.

die Stichprobengröße oder die Auswertung nach Schüler/inne/n vs. Klassen. Je komplexer das Design und je differenzierter die statistischen Verfahren, desto bedeutsamer werden solche Entscheidungen, erst recht auf der Ebene von Meta-Analysen (s. Kap. 3.1) – und desto schwieriger sind sie von Laien (und oft schon von den Forscher/inne/n selbst) nachzuvollziehen, z. B. bei Faktoren- oder Mehrebenen-Analysen. Umso größer wird damit auch die Verantwortung der Forscher/innen, die eben nicht nur berichten, sondern immer auch bewerten.

2.5 Maßstab der Bewertung: starke Wirksamkeit oder hohe Wichtigkeit?

Wir haben gesehen, wie schwierig es im pädagogischen Feld ist, schon einen einfachen Wirkungszusammenhang zu »sichern«. Bei konkreten Entscheidungen geht es aber immer um mehrere Kriterien – und die Befunde können sich widersprechen. Rechtschreibkönnen im Diktat deckt sich nicht mit Rechtschreibkompetenz in eigenen Texten (vgl. Brügelmann u. a. 1994; Fay 2010). Migrantenkinder haben in der BeLESEN-Studie zwar in Rechtschreibtests besser abgeschnitten, wenn sie mit einem Fibellehrgang unterrichtet wurden, als wenn sie nach der Methode »Lesen durch Schreiben« gelernt haben; ihre Leseleistungen waren aber im zweiten Fall besser (Schründer-Lenzen/Merkens 2006). Und was bewirken diese unterschiedlichen Methoden in Bereichen wie Freude am Schreiben und Lesen oder Selbstständigkeit? Zielkonflikte lassen sich empirisch nicht entscheiden – sie spielen aber bei pädagogischen Entscheidungen durchweg eine Rolle: kurzfristiger Erfolg in einer Prüfung vs. Entwicklung langfristig wirksamen Wissens und Könnens; fachliche vs. soziale Ziele; inhaltliche vs. methodische Kompetenzen.

Wer nur nach der Effektivität pädagogischer bzw. didaktischer Konzepte fragt, also einen möglichst hohen Wirkungsgrad zum Maßstab nimmt, verkürzt die Diskussion über pädagogische Konzepte unzulässig – ganz unabhängig davon, ob die Maße und Messverfahren angemessen sind. Und auch die Effizienz von Maßnahmen, das heißt ein optimaler Ertrag im Verhältnis zum Aufwand, ist ein nachgeordnetes Kriterium. So bedeutsam diese ökonomischen Aspekte sind – im Vordergrund müssten normative Überlegungen stehen: Wie soll der soziale Raum aussehen, in dem unsere Kinder und Jugendlichen heranwachsen? Nach welchen Grundsätzen wollen wir das Verhältnis der Generationen gestalten, und welche Pädagogik, welche didaktischen Konzepte werden diesen Ansprüchen gerecht? (→ »Pädagogik von gestern – in der Schule heute – für ein Leben morgen?« 2010)

Deshalb ist zu fragen, ob für die Evaluation von Praxismaßnahmen dieselben Kriterien gelten sollen (und können) wie für Grundlagenforschung. Evaluation umfasst die genaue Beschreibung von Sachverhalten, die Analyse von Beziehungen zwischen den betrachteten Faktoren und die Bewertung von Handlungsoptionen. Um Entscheidungen vorzubereiten, muss eine Evaluation auch Wissen bereitstellen. Sie zielt aber

nicht auf allgemeingültige Erkenntnis, sondern auf Entscheidungen in konkreten Fällen: Sollen in Niedersachsen die Förderschulen geschlossen, soll in Bayern wieder das neunjährige Gymnasium statt oder neben G8 eingeführt werden, soll die Fachkonferenz Maya in die nächste Klasse versetzen und Edy Ritalin bekommen? Evaluation ist gebunden an konkrete Bedingungen – und sie steht unter Zeitdruck.

Zugleich ist deutlich geworden, wie komplex das pädagogische Feld ist, mit welchen prinzipiellen Unwägbarkeiten bei Entscheidungen zu rechnen ist. Der Statistiker Gigerenzer (2007) hat an verschiedenen Beispielen aus Medizin, Wirtschaft und Bildung gezeigt, dass explizites Regelwissen den Anforderungen an handlungsleitende Vorhersagen seltener gerecht wird als erfahrungsbasierte Faustregeln oder Bauchentscheidungen (»unbewusste Intelligenz«). Bezogen auf pädagogisches Handeln hat Neuweg (2015) dieses Problem intensiv unter dem Stichwort »implizites Wissen« diskutiert. Im Anschluss an Polanyi (1958; 1966) und sein Konzept des »personal (tacit) knowledge« zeigt er drei Grenzen für Wissen in Form expliziter Regeln auf (Neuweg 2015, S. 28 ff.):

- die unzureichende Verbalisierbarkeit und Formalisierbarkeit der Expertise von Könner/inne/n (»Explikationsproblem«);
- die Schwierigkeit, dieses Wissen so darzustellen und zu vermitteln, dass Neulinge es verstehen und nutzen können (»Instruktionsproblem«);
- die fehlende Steuerungskraft expliziten Wissens, weil Handeln in komplexen Situationen durch zu viel Reflexivität behindert werden kann und eine strikte Regelanwendung den Besonderheiten unterschiedlicher Situationen nicht gerecht wird (»Modifikationsproblem«).

Lässt sich unter diesen Voraussetzungen das naturwissenschaftliche Paradigma einer Evidenzbasierung im Sinne des Goldstandards unbesehen auf Schule und Unterricht übertragen? Schön (1983) kritisiert die Vorstellung, komplexe Aufgaben, wie sie für didaktisches Handeln charakteristisch sind, ließen sich durch die Anwendung expliziten Regelwissens lösen. Stattdessen fordert er den »reflective practitioner«. Pädagogische Situationen sind eben nicht technisch beschreibbar und didaktische Interventionen nicht technologisch planbar und umsetzbar, wie es das Steuerungsparadigma voraussetzt. Dessen ökonomische Intentionen bzw. (Neben-)Wirkungen sind ein eigenes Thema und ebenfalls Gegenstand einer heftigen Kontroverse, die ich im Rahmen dieses Buches allerdings nicht angemessen diskutieren kann.

2.6 Wege aus den Dilemmata: Anregungen und Forderungen

Im Sinne einer Zwischenbilanz fasse ich nochmals die wesentlichen Aussagen und Thesen dieses Kapitels stichwortartig zusammen:

- Befunde in den Humanwissenschaften sind in hohem Maß situationsabhängig. Die Bedeutung eines Faktors verändert sich von Fall zu Fall mit dem Einfluss anderer Faktoren (Interaktionseffekte).

- 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.

Damit stellt sich die Frage: Was kann seitens der Wissenschaft getan werden, um die erforderliche Forschung vor Ort zu unterstützen?

Vor 25 Jahren haben Hopf u. a. für das Max-Planck-Institut für Bildungsforschung eine Studie über den Zustand der Grundschulen in Deutschland veröffentlicht (in: Baumert u. a. 1980, Bd. 2., Kap. 9). Die Forscher besuchten – jeweils zu mehreren – ausgewählte Schulen, sie führten Gespräche mit Verwaltungsbeamten und Experten aus Wissenschaft und Verbänden, sie werteten Statistiken und andere empirische Studien aus. Die Ergebnisse publizierten sie in einem Bericht, dessen Facettenreichtum und Realitätsnähe spürbar macht, was mit dem Paradigmenwechsel in der aktuellen Bildungsforschung verloren gegangen ist. Statistisch verfeinerte Verfahren sind hilfreich für spezielle Forschungsfragen und auch für manche politischen Entscheidungen. Zur Verbesserung des pädagogischen Alltags tragen sie wenig bei. Die Grundprobleme standardisierter Großstudien: Sie können nur wenige Faktoren erfassen, ihre Kategorien sind zu abstrakt und individuelle Bedeutungen gehen verloren. Was nützt eine Forschung, auf deren Ergebnisse sich wegen ihrer zu globalen Begrifflichkeit – »integrative Schulsysteme«, »offener Unterricht«, »Migrationshintergrund« – ganz kontroverse Positionen berufen können, wie bei PISA oder gar Hatties (2013) Meta-Meta-Analyse?

Darum kann man auf interpretative Fallstudien nicht verzichten. Allerdings sind viele besorgt, dass sie einen Rückfall in die Beliebigkeit subjektiver Deutungen bedeuten. Das Problem besteht in der Tat. Über eine technische Präzisierung der Methoden lässt es sich nicht lösen. Dies hat auch die Jurisprudenz schmerzhaft lernen müssen. Ihre Lösung: Verfeinerung der sozialen Kontrollen bei der Erhebung des Sachstands und bei der Auslegung von Normen. Die Ausweisung getrennter Rollen im Verfahren, eine Besetzung des Gerichts mit mehreren Richtern, die Einrichtung von Instanzenzügen sind Beispiele für solche sozialen Mechanismen (→ »Towards checks and balances« 1974; Brügelmann 2007). Wo Objektivität durch technische Präzisierung der Untersuchung nicht erreichbar ist, müssen die Auswirkungen von Subjektivität transparent und diskutierbar gemacht werden. Forschung ist auch ein politisches Geschäft. Damit werden andere Kriterien relevant:

Relative Einschätzungen statt allgemeiner Urteile

Allgemeinaussagen über die Wirksamkeit und Bewertungen von Programmen oder Methoden als insgesamt »besser« oder »schlechter« sind angesichts der Konzept- und Kontextabhängigkeit von Wirkungen wenig hilfreich. Zu fordern, es dürften nur Methoden oder Programme mit empirischem Gütesiegel (»evidenzbasiert«) in den Schulen eingesetzt werden, überzieht den Kredit der Forschung. Denn redlicherweise müsste man nach deren Evaluation sagen: »In den untersuchten Klassen hat sich gezeigt, dass die Intervention im Durchschnitt erfolgreich war – ob das dort morgen noch so sein wird und wie die Ergebnisse andernorts ausfallen würden, wissen wir nicht.« Wohl aber können und sollten die Potenziale und Risiken der untersuchten Intervention benannt werden. Dies ist in bloßen Input-Output-Untersuchungen nicht möglich, die weder die vermittelnden Prozesse noch die unterschiedlichen Bedingungen) erfassen (→ »Kultur und Kontext« 1995 und → »›Black box‹ Kognition und ›grey context‹ Kultur« 2003). Es geht um komplexe Konstellationen. Bei der Untersuchung eines Programms auf seinem → »Weg vom Autor zum Kind« (1978) ist z. B. zu bedenken:

- Wie haben es die Entwickler verschiedenen Referenzgruppen verkauft: gegenüber dem Sponsor vielleicht als anspruchsvolle Innovation und gegenüber der Praxis als einfache Lösung ihrer Alltagsprobleme?
- Haben die Lehrer/innen die Intentionen des Programms so verstanden, wie die Entwickler sie gemeint hatten?
- Wie haben sie es jeweils umgesetzt?
- Wie haben verschiedene Schüler/innen diese konkretisierten Lernangebote genutzt?

Effekte lassen sich einem Programm auf dem Papier also nicht direkt zurechnen, da es sich bei der Umsetzung mehrfach verwandelt. Um diese Veränderungen zu erfassen, sind weitere Anforderungen zu erfüllen:

Mehrperspektivität bei der Erhebung und Deutung von Daten

Wie im Verlauf dieses Kapitels immer wieder deutlich geworden ist, sind Mittelwerte* nur begrenzt aussagefähig. Mindestens ebenso interessant ist die Streuung um den Durchschnitt. Sie ist nicht nur die Folge besonderer Bedingungen, sondern auch unterschiedlicher Deutungen. Da diese den Umgang mit Programmen oder Methoden bestimmen, sind qualitative Einschätzungen der Sichtweisen verschiedener Beteiligter wichtig. Sie sind zumindest exemplarisch zu erheben.

Auch die Daten von Großstudien sind mehrdeutig. Deshalb ist vor ihrer Veröffentlichung darauf zu achten, dass nicht einzelne Gruppen die Deutungsmacht behalten. Dafür ist es wichtig, dass Untersuchungen öffentlich dokumentiert und damit für Sekundäranalysen zugänglich sind. Aber das reicht nicht. Vor allem bei bildungspolitisch wirksamen Untersuchungen wie PISA gilt das Gesetz des ersten Eindrucks: Wenn die medial wirkungsvoll inszenierte Analyse der Autoren (»Ranglisten«) erst einmal in der Welt ist, wird es schwierig, diese Deutungen im Nachhinein zu korrigieren. Es sollte zum Standard guter Forschung gehören, dass sie die

Mehrdeutigkeit ihrer Befunde selbst zum Thema macht, z.B. indem schon für die Erstveröffentlichung Kommentatoren eingeladen werden, die verschiedene Disziplinen oder unterschiedliche theoretische Positionen in derselben Disziplin vertreten. Dies führt zum nächsten Punkt:

Komplementarität von Methoden

Der US-amerikanische Bildungsforscher David C. Berliner (2002) hat die Bildungsforschung als »the hardest science of all« gewürdigt. Bei der Übersetzung angelsächsischer Forschungsansätze ins Deutsche ist die tiefere Bedeutung dieses Satzes wohl leider verloren gegangen: Bildungsforschung lebt nicht von »harten Daten« – sie ist ein »hartes Geschäft«. Dafür brauchen wir auch Studien, die dem Standardisierungsparadigma folgen und die Zusammenhänge zwischen ausgewählten Variablen an großen Stichproben* untersuchen. Aber als Norm für die Wissenschaftlichkeit von Bildungsforschung führen sie in eine Sackgasse. Das sehen auch reflektiertere Vertreter des sogenannten quantitativen Paradigmas.

Repräsentative Stichprobenerhebungen und Fallanalysen, standardisierte Methoden und interpretative Verfahren können und müssen sich ergänzen. Dafür reicht nicht das gönnerhafte Zugeständnis: »Fallstudien können als Voruntersuchungen hilfreich sein.« Wer sie nur als pädagogische Folklore sieht, verkennt den Eigenwert dieses Erkenntniszugangs, wie er in Kapitel 2.3 begründet wurde. In empirischen Studien müssten beide Zugänge genutzt und die Befunde aus großen Stichproben und aus Fallstudien möglichst über Brückeninstrumente aufeinander bezogen werden. Es geht also um mehr als nur eine friedliche Koexistenz in verschiedenen Projekten, auch nicht um ein bloß additives Nebeneinander innerhalb eines Projekts, sondern um eine inhaltliche Verknüpfung.

Bildungspolitik und Forschungsförderung bevorzugen zurzeit allerdings einseitig ein Paradigma. Damit geht die Komplementarität der Zugänge verloren. Deren Notwendigkeit wurde bereits früher, z.B. in den 70er Jahren, differenziert diskutiert. Leider beginnen die Literaturverzeichnisse vieler aktueller Veröffentlichungen erst mit Titeln aus diesem Jahrtausend. Forschungsergebnisse zu einzelnen Verhaltensbereichen können rasch veralten. Grundeinsichten in die Natur dieser Forschung nicht.

Differenzierung und Trennung von Rollen

Wenn eine technische Perfektionierung der Methoden ihre Grenzen hat, wird eine soziale Kontrolle des Forschungsprozesses notwendig. Innerhalb einer Studie kann man z.B. die Verantwortung für die Erhebung und die Auswertung von Informationen trennen. Bei der Publikation von Befunden können unterschiedliche Interpretationen nebeneinandergestellt werden. Bisher bleiben solche Differenzen oft verborgen – etwa bei dem anonymen »Peer-Review« vor der Veröffentlichung von Studien.

Aber auch die Arbeitsteilung zwischen Wissenschaft und Praxis ist zu überdenken. Letztere ist mehr als die bloß technische »Anwendung« von Erkenntnissen, das heißt das Verhältnis ist nicht hierarchisch, sondern kooperativ zu sehen. Die Pädagogen

als »reflective practitioner« (Argyris/Schön 1978) sind auf Wissenschaft als Anreger und als kritisches Gegenüber angewiesen. Aber ebenso hat diese Wissenschaft die Pädagogen als »experimental colleagues« (Stenhouse 1975) ernst zu nehmen, die ihre Entscheidungen im Blick auf die Bedingungen vor Ort zu begründen und zu evaluieren haben. Man darf nicht kompetente Forscher/innen mit inkompetenten Praktiker/inne/n vergleichen, will man den Wert von Empirie begründen. Lehrer/innen haben eine eigenständige Aufgabe und bringen spezifische Kompetenzen ein.

Es geht aber auch um Aufteilung von Macht. Wenden wir uns also der institutionellen Organisation von Verantwortlichkeiten zu.

2.7 Balance von Ebenen und Rollen: ein differenziertes System von Rechenschaftsformen

Verantwortung rollengerecht verteilen

Evaluations- und Rechenschaftsformen müssen im Blick auf die pädagogische Entwicklung des Schulalltags differenziert werden. Damit wird dann auch deutlich, wie unterschiedlich die Anforderungen an Verfahren und Instrumente der Evaluation sind.

Ein differenziertes System der Qualitätsentwicklung richtet Maßnahmen der Beobachtung und Überprüfung auf die Ebenen aus, auf denen spezifische Verantwortlichkeiten für eine Verbesserung des Unterrichts bzw. seiner Bedingungen liegen. Solche Entscheidungsträger sind …

- Schüler/innen und Eltern mit der gemeinsamen Verantwortung für die Lebensplanung der Kinder, insbesondere für eine sinnvolle Nutzung der individuellen Entwicklungsmöglichkeiten;
- die Pädagogen, differenziert im Blick
  - auf die Entwicklung einzelner Schüler/innen und
  - auf ihre eigene Arbeit in und mit der Klasse;
- die Schule bzw. die verschiedenen Gremien innerhalb der Schule – dabei heute besonders wichtig: kooperativ arbeitende Teams (aus verschiedenen Professionen), die eine besondere Verantwortung für bestimmte Lern-, Alters- oder Fachgruppen haben;
- die in Verwaltung und Politik tätigen Personen, differenziert nach regionalen Behörden, Ministerium und Parlament;
- Institutionen und Personen der Aus-, Fort- und Weiterbildung zur (Weiter-)Entwicklung und Überprüfung von pädagogischen und didaktischen Kompetenzen.

Erst eine Differenzierung nach solchen Zuständigkeiten und die Vernetzung der jeweiligen Aktivitäten ermöglichen es, die Daten bereitzustellen, die auf den verschiedenen Ebenen gebraucht werden, um Entscheidungen abzusichern und zu überprüfen (→ »Formen der Rechenschaft« 1999). Dabei ist im Blick zu behalten, dass sich konkrete

Evaluationsmaßnahmen auf Probleme richten sollten, für deren Lösung die evaluierte Einrichtung oder Person auch zuständig ist, und dass es auf der betreffenden Ebene tatsächlich Handlungsmöglichkeiten zur Verbesserung der Situation gibt bzw. die Möglichkeit, Unterstützung von außen einzufordern.

Evaluation muss helfen, den Unterricht zu verbessern

Die Qualität von Unterricht zu fördern ist eine anspruchsvolle und konfliktträchtige Aufgabe, deren Erfüllung von der Bildungspolitik und -verwaltung differenzierte Maßnahmen verlangt. Ihre Beiträge können darin bestehen, ...

- Engagement und Erfolge von Lehrer/inne/n anzuerkennen und sie zur weiteren Entwicklung ihres Unterrichts zu ermutigen;
- diejenigen herauszufordern, die ihren Bildungs- und Erziehungsauftrag mit größerer Sorgfalt und Kompetenz wahrnehmen müssten, und die Einhaltung normativer Vorgaben einzufordern;
- allen Lehrer/inne/n dort Unterstützung, Beratung und Fortbildung anzubieten, wo sie sich überfordert fühlen, den Anforderungen gerecht zu werden;
- die erforderlichen personellen und materiellen Rahmenbedingungen zu sichern.

Evaluation muss so organisiert werden, dass die erhofften Impulse und Einsichten nicht durch ungewollte Nebenwirkungen zunichtegemacht werden. Sie sind deshalb auszurichten auf spezifische Informationsbedürfnisse und Anforderungen in konkreten Handlungsfeldern. Schulen und Lehrer/innen sind darin zu unterstützen, dass sie ihre Schüler/innen, z.B. durch ein besseres Verständnis von Lernprozessen und -schwierigkeiten, bei der Entwicklung ihrer individuellen Fähigkeiten gezielt fordern und fördern können.

Um statistische Kennwerte und Bezugsnormen aus repräsentativen Untersuchungen mit Leben zu füllen, sind dichte, situations- und prozessbezogene Fallstudien hilfreich. Vor allem aber muss die Kapazität und Kompetenz zur (Selbst-)Evaluation auf Dauer im System selbst installiert werden. Schulen müssen »lernende Systeme« werden. Dafür brauchen Lehrer/innen einerseits methodische Kompetenzen und alltagstaugliche Verfahren. Andererseits profitieren sie von externen Sichtweisen und von der Unterstützung durch externe Kompetenz und Ressourcen. So hat sich bewährt, dass ein Kollegium Beobachter als »kritische Freunde« einlädt, um die Selbstwahrnehmung mit einem Blick von außen zu konfrontieren (s. Kap. 5.3). Auf der Materialebene sind kommentierte Aufgabenbeispiele und Orientierungswerte aus repräsentativen Stichproben* hilfreicher als punktuelle Vergleichsarbeiten für alle Klassen, die auf die besondere Situation vor Ort (Eingangsvoraussetzungen, Unterrichtsschwerpunkte, Rahmenbedingungen) nicht Bezug nehmen können.

Letztlich müssen alle Maßnahmen dazu beitragen, die Entwicklung von Schulen und konkrete Verbesserungen des Unterrichts in Gang zu bringen, das heißt die »Leh-

renden so zu stärken, dass sie die Lernenden stärken« können (Heide Bambach). Deshalb sollten sich die Anstrengungen der Evaluation und die finanziellen Mittel darauf konzentrieren, Modelle und Voraussetzungen gelingenden Unterrichts zu untersuchen und zu dokumentieren. Insbesondere Fallstudien von »wider Erwarten erfolgreichen« Personen bzw. Einrichtungen können helfen, lernförderliche Bedingungen zu erkennen und analoge Entwicklungen an anderen Orten anzuregen und zu unterstützen. Auf zwei Ebenen sind solche Untersuchungen wichtig:

- Fallstudien von Schüler/inne/n, die sich trotz ungünstiger Voraussetzungen (z.B. individueller Behinderung, bildungsfernem Milieu, sozialer Benachteiligung, Zurückstellung am Schulanfang, Wiederholung einer Klasse) in der Schule erfolgreich entwickeln;
- Fallstudien von Schulen, die trotz eines schwierigen Umfelds (z.B. sozialer Brennpunkt, hoher Anteil an Kindern mit einer anderen Muttersprache, fehlende Altersmischung im Kollegium) erfolgreich arbeiten.

Im Vordergrund sollten also Maßnahmen zur Entwicklung von Qualität stehen. Eine bloß deskriptive Bestandsaufnahme weithin bekannter Problembereiche bringt wenig. Vorrang vor einem nur punktuellen Vergleich von Effekten hat die Analyse von Prozessen und Bedingungen.

Notwendig sind langfristig angelegte Entwicklungskonzepte statt kurzatmiger Reaktionen auf politische Stimmungslagen, und das heißt vor allem: Entwicklung von Beobachtungs- und Förderhilfen, um Probleme von Lehrer/inne/n und Schüler/inne/n in ihrer konkreten Arbeitssituation aufklären und bewältigen zu können – und die begleitende Unterstützung und Evaluation solcher Maßnahmen (s. Kap. 5.3 und 6.3).

3. Hattie und der Zauber großer Zahlen

3.1 Was bringen Meta-Analysen der Unterrichtsforschung?

Vor einigen Jahren 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: »Visible Learning« von John Hattie (2009, dt. 2013); Daten zu mehr als 130 Merkmalen von Schule und Unterricht aus über 50 000 Basisstudien. Ist damit wirklich endgültig geklärt, was »guten Unterricht« ausmacht?

Die empirische Bildungsforschung – vor allem in den angelsächsischen Ländern – hat in den vergangenen 20 bis 30 Jahren eine Unzahl von Befunden aus Einzelprojekten zu verzeichnen, die kaum jemand mehr überblicken kann. Angesichts dieser zunehmenden Fülle von Detailstudien hat die Bedeutung sogenannter Meta-Analysen zugenommen. Um zu sehen, ob es eine Beziehung zwischen Faktoren wie z. B. dem »entdeckenden Lernen« und dem Sachwissen von Schüler/inne/n gibt und wie stark sie ausgeprägt ist, verrechnet man statistische Effekte über verschiedene Studien hinweg. Man tut dies, obwohl die Daten unter verschiedenen Bedingungen, in verschiedenen Kontexten (Bildungssystemen, Sprachen, Kulturen ...) und mit unterschiedlichen Instrumenten gewonnen worden sind. Hattie (2013, S. 13 f., 16) sieht diese Probleme durchaus, aber er schätzt ihre Auswirkungen geringer ein als andere. Über die Verdichtung der Befunde aus vielen Studien hofft er 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.

Hatties Zusammenschau wird auch in Deutschland immer häufiger zitiert, um bestimmte Methoden oder Ansätze als mehr oder weniger sinnvoll zu bewerten – in der Wissenschaft wie auch in der öffentlichen Diskussion. Denn Hattie ist noch einen Schritt weitergegangen als die Kollegen vor ihm: Er hat die Werte aus über 800 Meta-Analysen zum Schulerfolg auf einer noch höheren Ebene in einer Meta-Meta-Analyse weiter verdichtet und findet in der Tat einige beeindruckende Zusammenhänge. Hier die Plätze 1 bis 10 der angeblich einflussreichsten Faktoren (s. ergänzend unten die Auswahlübersicht 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

Abb. 2: Hattie-Ranking (ES = Effektstärke; ab 0,5 gilt ein Effekt gemeinhin als beachtenswert – der Anschaulichkeit halber oft gleichgesetzt mit dem Vorsprung von etwa einem Schuljahr; Hattie setzt die Schwelle bei einer ES von 0,4)*

Die formal aufgelisteten Merkmale lassen allerdings Spannungen erkennen, die den Sinn von Ranglisten, wie Hattie sie publiziert, infrage stellen. Zum Beispiel wirke positiv, wenn Lehrer/innen klare Anweisungen bei der Aufgabenstellung geben, aber auch wenn Schüler/innen anderen Schüler/inne/n etwas beibringen oder wenn sie sich gar selbst beurteilen dürfen. Insofern sind additive Listen von Unterrichtsmerkmalen ohne eine gemeinsame theoretische Rahmung wenig hilfreich. Dass Studien nach demselben Ansatz noch einmal von anderen Forscher/inne/n durchgeführt werden, kommt in den Erziehungswissenschaften selten vor. Nach Makel/Plucker (2014, S. 310 f.) trifft das auf deutlich weniger als 1 Prozent der publizierten Studien zu. Wurde die Untersuchung nicht von denselben Autoren wiederholt, werden die Befunde nur in der Hälfte der Fälle voll bestätigt. Das kann an methodischen Schwächen liegen, aber auch daran, wie die untersuchten Konzepte jeweils verstanden und umgesetzt wurden.

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, aus denen als »gleich« angesehene Variablen miteinander verrechnet werden, als auch bei der Übersetzung in andere Sprachen (s. die Originalbegriffe in Klammern) und schließlich in der individuellen Auslegung der stark verknappten »Etiketten« durch einzelne Leser/innen:

ES	
1.4	Selbstbeurteilung
1.3	Piaget-orientierte Programme
.9	förderorientierte Beurteilungen
.9	Einüben von methodischen Fertigkeiten (»micro teaching«)
.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 der Effekt sogar ganz (-.02). Andererseits steigt er, wenn die Schüler das Vorgehen im Programm selbst steuern konnten auf .49, und wenn sie mit einem Partner arbeiteten sogar auf .96.

Damit deutet sich schon auf dieser noch sehr allgemeinen Ebene an, dass nicht die Mittelwerte* das eigentlich interessante Datum sind, 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? Während Hattie Abweichungen vom Mittelwert sozusagen als Messfehler betrachtet, verweisen sie für Pädagogen vor Ort auf den Einfluss bedeutsamer Drittfaktoren.

Diese Konzept- und Kontextabhängigkeit von Wirkungen einer Methode zeigt sich auch bei anderen Interventionen. So hat nach Hattie ein auf Problemaufgaben 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 auf .75, für den Erwerb von Basiswissen ist die Beziehung dagegen sogar negativ, das heißt ein traditioneller Unterricht ist in dieser Hinsicht erfolgreicher.

Entsprechend relativieren Hattie und auch vorsichtige Interpreten einzelne Aussagen bzw. ihren Geltungsbereich. Trotzdem operieren die meisten mit den Ranglisten, als spiegelten sie aussagekräftige Merkmale von Unterricht. Und die machen Furore. So wird vielerorts die Rückkehr zu traditionellen Lehrmethoden gefordert, denn offener Unterricht sei nicht wirksamer als traditioneller Unterricht (d = .01). Aber was

bei Hattie bzw. in den von ihm zusammengefassten Meta-Studien konkret als »offen« erfasst worden ist, kann nicht mehr nachvollzogen werden. Aus anderen Analysen wissen wir, dass sich in diesem Topf ganz unterschiedliche Ansätze befinden. Und so überrascht es nicht, dass – entgegen den verbreiteten Plädoyers für mehr Frontalunterricht – bei Hattie als »sehr wirksame« Faktoren auch folgende auftauchen:

.90	»formative assessment« (inhaltsbezogene Rückmeldungen)
.82	Diskussionen zwischen Schüler/inne/n im Unterricht
.74	reziprokes Lernen (wechselseitige Unterstützung der Schüler/inn/en untereinander)
.69	»meta-cognitive strategies« (über das eigene Lernen nachdenken)
.54	kooperatives statt konkurrenzorientiertes Lernen
.47	Überzeugung der Schüler/innen von ihrer Selbstwirksamkeit

Das sind allesamt Merkmale eines Unterrichts, der Schüler/inne/n Raum für eigene Lernwege gibt und ihnen Verantwortung für das eigene Lernen zugesteht. Das Bild ist also viel komplizierter, als herausgegriffene Einzelmerkmale vermuten lassen – gerade im Hinblick auf die Diskussion über lehrer- vs. schülerzentrierten Unterricht.

Schon die SCHOLASTIK-Studie von Weinert/Helmke (1997) operierte mit allgemeinen Merkmalen wie »Strukturiertheit«, die – ohne Kenntnis der Originalskalen – ganz Unterschiedliches bedeuten können und entsprechend unterschiedlich gedeutet werden: inhaltliche Systematik, methodische Kleinschrittigkeit, Kontrolle von Verhaltensregeln. Zudem können Strukturen von der Lehrperson vorgegeben oder mit den Schüler/inne/n ausgehandelt werden: Regeln für Arbeits- und Verkehrsformen, gemeinsame Projektplanung, individuelle Lernverträge ... In der didaktischen Diskussion werden Vorteile von Strukturiertheit jedoch meist als Argumente für Lehrerzentrierung verbucht, ohne dass – erst recht in Meta-Studien wie der von Hattie – klar ist, wie der Gegenpol in den einzelnen Basis-Studien aussah.

Für die Einschätzung des Erfolgs einer Unterrichtsmethode kommt es also erstens darauf an, an welchem *Kriterium* man diesen misst. Dann ist zu prüfen, in welcher *Form* und unter welchen *Bedingungen* die Methode umgesetzt wird. Die für einen Überblick über die Forschung und für bildungspolitische Entscheidungen verdienstvolle Verdichtung der Daten in Hatties Meta-Meta-Analyse hat den Nachteil, dass die Daten durch die Ablösung vom konkreten Kontext weitgehend ihre Bedeutung für die Praxis verlieren. Denn die Zahlen sind mehrdeutig. Rolff (2013) vergleicht die Hattie-Studie und ihre Rezeption sogar mit einem Rorschach-Test. Was bedeutet zum Beispiel eine mittlere Korrelation* oder Effektstärke* im konkreten Fall: einen leicht positiven Einfluss in allen Fällen *oder* einen starken in einigen und gar keinen in anderen *oder* einen hohen Einfluss im Zusammenwirken mit Faktor X, aber einen niedrigen im Zusammenwirken mit Variable Y? Ganz abgesehen von dem bereits diskutierten Problem, dass das gleichzeitige Auftreten von zwei Faktoren nicht gleichbedeutend ist mit einer inhaltlichen Beziehung: Statistische Korrelation ist nicht gleichzusetzen mit inhaltlicher Kausalität.

Das heißt: Selbst wenn 99 von 100 Studien eine Überlegenheit von Methode B zeigen, dürfte das kein Grund sein, einer Lehrerin die Nutzung von Methode A, mit der sie erfolgreich arbeitet, zu untersagen (vgl. das Beispiel der Computernutzung: einige Ausreißer unter den Meta-Analysen weisen besonders positive 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 Stichproben* dann eine »Beweislast«, das heißt eine Begründungspflicht für Abweichungen – sie können diese aber nicht abwerten oder gar ausschließen. Schon Weinert/Helmke (1997, S. 48) fanden unter ihren erfolgreichsten Lehrer/inne/n »bizarre Merkmalsprofile« in den untersuchten Dimensionen, also selbst auf dieser abstrakten Ebene gerade nicht eine Idealform, die dem Durchschnitt entsprach.

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* und Handlungsanweisungen nicht von einer Ebene auf die andere übertragen: Sie müssen »rekontextualisiert« werden (vgl. Fend 2008). Insofern gerät die vertraute Grundfigur der Forschung und ihr Verhältnis 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-)Meta-Analysen können dafür hilfreiche Hypothesen liefern – aber keine Vorschriften.

Für die wissenschaftsinterne Diskussion bedeuten die genannten Einschränkungen, dass Allgemeinurteile nur als Wahrscheinlichkeitsaussagen formuliert werden dürfen. »Evidenz« sollte beim Umgang mit diesen Unsicherheiten helfen. Zu beschreiben sind Potenziale und Risiken, und diese sind möglichst differenziert auf bedeutsame Bedingungen zu beziehen. Die zunehmende Abstraktion der Begriffe in Meta-Analysen muss deshalb ausbalanciert werden durch dichte Fallstudien, wie das folgende Kapitel verdeutlicht.

3.2 Das Kind zwischen Hattie und PISA[4]

Frage: In der Öffentlichkeit wird die Diskussion über Bildung stark von den PISA-Ergebnissen bestimmt. Wie stehen Sie dazu?

Behnken: Als erstes: Wir lehnen die von PISA & CO verwendeten Verfahren nicht ab. Wir setzen sie für bestimmte Zwecke selbst ein. Die repräsentativen Kinder- und Jugendbefragungen »NRW-Kids« und »LERNenBILDung« beispielsweise waren nur mit standardisierten Fragebögen möglich.

Brügelmann: Leistungstests wiederum waren das zentrale Medium in unserem Projekt LUST, der »Lese-Untersuchung mit dem Stolperwörter-Test« über die Grundschulzeit hinweg ...

Heymann: ...und in unserem Projekt SCHLAU, in dem wir die Leistungsentwicklung während der Ferien im Vergleich zur Schulzeit untersuchen.

Frage: Dennoch sind Sie kritisch? Was macht PISA denn falsch bzw. was ist problematisch?

Behnken: Nun, zu dieser Frage bzw. zu zahlreichen Einzelaspekten der konkreten PISA-Kritik sind jüngst wichtige Beiträge erschienen. Problematisch ist übergreifend sicherlich, dass PISA, eigentlich ein Forschungsprojekt, zu einem Programm für Forschung, gar zu einem Paradigma für Bildungsforschung geworden ist. Dabei greift dieses Paradigma zu kurz, um die Alleinvertretung einer Forschungsrichtung sein zu können.

Frage: Was machen Sie denn anders als PISA?

Heymann: Man sollte vielleicht genauer von »PISA & Co.« sprechen. Es gibt ja inzwischen viele internationale und deutsche Studien, die Schülerleistungen untersuchen. Genau da liegt aber auch das Problem: Kinder und Jugendliche werden nur in ihrer Schülerrolle erfasst, zusätzlich eingeschränkt auf ihre Leistungen in wenigen ausgewählten Fächern. Uns interessiert das auch – aber wir verstehen die Person nicht nur als Variablenbündel. Im Projekt »Lernbiografien im schulischen und außerschulischen Kontext« (LISA & KO) versuchen wir sie anschaulich und erfahrungsnah zu porträtie-

4 Die Gesprächspartner in diesem Interview (aufgezeichnet von Michael Hellermann 2008) sind Imbke Behnken (vom »Siegener Zentrum für Kindheits- und Jugendforschung«), Hans Brügelmann und Hans Werner Heymann (damals beide Leiter des Siegener Lehr- und Forschungsprojekts »Lernbiographien im schulischen und außerschulischen Kontext« → »Lernbiografien im schulischen und außerschulischen Kontext« 2012 und Wagener 2012).

ren, erweitert auf die Biografie des Einzelnen und eingebettet in seine Lebenswelt. Es geht uns also um Relativierung und Ergänzung der PISA-Sicht.

Brügelmann: Und das bedeutet, dass sich auch die Methoden verändern, z. B. stärker narrative Formen einbeziehen müssen. PISA & Co. arbeiten vor allem mit standardisierten Instrumenten: mit Tests und mit Fragebögen. Beide sind nützliche Instrumente, aber sie schränken den Blick auch ein. Das ist wie in den Naturwissenschaften: Mit einem Mikroskop sehen wir die Welt anders als mit einem Fernrohr. Man muss sich klarmachen, welche Ausschnitte man mit einer bestimmten Methode erfasst und was man alles nicht in den Blick bekommt, wenn man diesen Ausschnitt scharf erfassen will.

Frage: Was heißt das konkret für Ihre Untersuchungen?

Heymann: Nehmen wir Tests. Die Standardisierung von Aufgaben erlaubt einen ökonomischen Einsatz in großen Gruppen. Der Preis ist aber, dass nur bestimmte Aufgabentypen eingesetzt werden können, dass eine künstliche Situation geschaffen werden muss, um die Vergleichbarkeit zu sichern, dass nur wenig Zeit zur Verfügung steht, sodass man sich auf wenige Indikatoren beschränken muss. Man bekommt also recht präzise Daten – aber wie aussagekräftig sind sie? In unserem Projekt »Lernbiografien im schulischen und außerschulischen Kontext«, in dem einzelne Kinder und Jugendliche über mehrere Monate begleitet werden, stellen wir immer wieder Differenzen fest zwischen Testergebnissen und einschlägigen Leistungen in anderen Situationen, die an der Aussagefähigkeit punktueller Tests zweifeln lassen.

Brügelmann: Interessant ist auch, was die Studierenden sagen, nachdem sie ein Kind über viele Wochen hinweg beobachtet und befragt, nachdem sie Tests durchgeführt und andere Personen zu ihren Eindrücken über das Kind befragt haben. Immer wieder lesen wir in den forschungsmethodischen Reflexionen zu den Porträts, die sie schreiben: »Am meisten über die besonderen Kompetenzen, Interessen und Aktivitäten des Kindes haben wir aus den informellen Begegnungen mit ihm gelernt.« Hier zeigt sich die Bedeutung des Forschers, vielleicht das wichtigste »Forschungsinstrument« neben ausgearbeiteten und standardisierten Instrumenten.

Frage: Aber ist es nicht riskant, sich auf das Urteil von einzelnen Beobachtern zu verlassen? Man weiß doch beispielsweise von den Noten in der Schule, wie subjektiv und wenig vergleichbar die sind. Treiben Sie den Teufel nicht mit dem Beelzebub aus?

Behnken: Sie haben völlig Recht. Die Person des Beobachters ist ein besonders sensibles, aber auch besonders fehleranfälliges Instrument. Darum sind uns zwei Prinzipien für unsere Studien besonders wichtig: Kombination komplementärer Methoden und

Multiperspektivität. Wir setzen ja durchaus auch standardisierte Tests, Beobachtungsraster und Fragebögen ein …

Heymann: …aber daneben eben auch informelle Beobachtungen, offene Interviews und Alltagsaufgaben, z.B. Würfelspiele, in denen wir beobachten können, wie vertraut ein Vorschulkind mit Zahlen ist. Besonders wichtig ist uns aber, die beteiligten Personen als Experten ihrer Lebenswelt ernst zu nehmen. Ihre Wahrnehmungen und ihre Deutung dieser Wahrnehmungen sind für uns der Schlüssel zu einem besseren Verständnis einer Person.

Brügelmann: Wir gehen davon aus, dass menschliches Verhalten immer mehrdeutig ist. Standardisierte Instrumente unterstellen, dass Kommunikation eindeutig gemacht werden kann. Aber selbst, wenn Verständigung in vielen Fällen gelingt, kann man nie sicher sein. Intelligente Personen haben oft besondere Schwierigkeiten mit Tests oder Fragebögen. Sie überlegen, was der Verfasser der Aufgaben oder Fragen gemeint haben könnte, können ihre Vermutungen in dieser Ein-Weg-Kommunikation aber nicht validieren. Sie lösen deshalb Aufgaben trotz höherer Kompetenzen nicht zwingend erfolgreicher. Dies ist ein Grundproblem, das man technisch-methodisch nicht in den Griff bekommen kann. Wir suchen deshalb das direkte Gespräch mit den Betroffenen und wir legen Widersprüche zwischen verschiedenen Beteiligten oder zwischen unterschiedlichen Instrumenten offen, das heißt, wir ergänzen die technische durch eine soziale Kontrolle.

Behnken: Außerdem erfordern spezifische Erkenntnisinteressen auch ihre besonderen Methoden. An so etwas wie »kollektive Identitäten« von Gruppen, z.B. von arbeitslosen Jugendlichen oder von Migranten am Gymnasium, kommt man am ehesten über Gruppendiskussionen heran. Kinder lassen wir subjektive Landkarten zeichnen, um ihre Bewegungsräume und die Bedeutung verschiedener Orte zu erfassen. Oder wir fordern Kinder und Jugendliche auf, sich mit Gegenständen für ein Selbstporträt in Szene zu setzen, die wir dann fotografieren.

Frage: Aber noch einmal zurück zu Ihrem Projekt »Lernbiografien im schulischen und außerschulischen Kontext«. Das sind doch alles Einzelfälle. Inwiefern sind die Befunde überhaupt generalisierbar?

Brügelmann: Jeder Mensch ist ein Einzelfall. Verallgemeinerungen sind Vereinfachungen. Sie beruhen auf statistischen Aussagen, mit einem entsprechenden Fehlerrisiko; auch bei den statistisch bearbeiteten Studien mit großen Fallzahlen ist dies so. Die kann man nicht so einfach auf eine neue Situation »anwenden« wie das Fallgesetz in der Physik.

Heymann: Ziel von Bildungsforschern ist auch nicht notwendigerweise die Verallgemeinerung. Es ist natürlich sinnvoll zu überprüfen, ob es starke statistische Zusammenhänge gibt, beispielsweise zwischen Buchbesitz in der Familie und der Lesefähigkeit der Kinder. Jedoch: Es gilt weder, dass alle Kinder, die viele Bücher besitzen, gute Leser werden, noch dass allen Kindern, die keinen Zugang zu Büchern haben, der Erfolg in der Schriftsprache verwehrt sein muss. Hier gilt es wieder den Blick auf die Kinder zu lenken, die wider Erwarten erfolgreich oder nicht erfolgreich sind. Die scheinbare Verallgemeinerung aus quantitativer Forschung enttarnt sich somit entweder als zu kurz gegriffen (wenn zu wenige Faktoren berücksichtigt werden, z. B. nur der Besitz von Büchern anstatt der Umgang mit ihnen zur Kenntnis genommen wird) oder als nicht gültig für alle Gruppen.

Brügelmann: Bei PISA wird beispielsweise erfasst, wie oft Eltern ihren Kindern vorlesen. In einer Großstudie ist das ein nützlicher Indikator für Bildungsanregungen im Elternhaus. Aber wenn Sie Lehrerin sind, werden Sie merken, dass Vorlesen im Elternhaus für zwei Kinder bei gleicher Häufigkeit durchaus Unterschiedliches bedeuten kann, je nachdem wie die Eltern vorlesen und in welcher Atmosphäre. Diese Qualitäten aber entscheiden letztlich über die Bildungswirkung solcher Aktivitäten. Das zu wissen ist auch wichtig, wenn man Förderprogramme entwirft. Vorlesen als äußeres Verhalten reicht nicht. In den Humanwissenschaften ist es anders als in den Naturwissenschaften: Indikatoren sind nicht stabil, sie verändern ihre Bedeutung, wenn die Menschen lernen, worauf es ankommt, und das gewünschte Verhalten lediglich imitieren.

Frage: Heißt das, man kann in den Humanwissenschaften keine Allgemeinaussagen machen?

Brügelmann: Doch, das schon, aber hilfreich sind Verallgemeinerungen zunächst für Entscheidungen, die viele Personen betreffen: Behalten wir die Sonderschule bei, schaffen wir das Sitzenbleiben ab, sollen Migrantenkinder bilingual unterrichtet werden? Anders ist es bei Entscheidungen im Einzelfall: Soll Lena auf die Sonderschule oder in der Grundschule bleiben, soll sie freiwillig die Klasse wiederholen, und soll Ahmed auf Türkisch oder auf Deutsch lesen und schreiben lernen? Für solche Entscheidungen bieten Aussagen, wie sie bei PISA und in ähnlichen Studien gewonnen werden, nur Hypothesen, deren Geltung im Einzelfall jeweils neu zu überprüfen ist. Dafür spielen viele Bedingungen eine Rolle.

Frage: Aber Sie können doch auch nicht aus ein paar Fallstudien ableiten, wie ein neuer Fall zu entscheiden ist!

Heymann: Richtig, auch von Fall zu Fall lässt sich Erfahrung nicht direkt übertragen, genauso wenig wie auf dem induktiv-deduktiven Weg, also über Verallgemeinerungen.

Aber Fallstudien sind oft hilfreicher, weil sie Kontexte mitbenennen. Wer viele dicht beschriebene Einzelfälle kennt, der verfügt über ein breites Repertoire an »Brillen«, die ihm helfen, die Besonderheiten eines neuen Falles einzuschätzen. Für Pädagogen, die täglich mit solchen »Einzelfällen« arbeiten, führt diese Herangehensweise daher wesentlich weiter.

Brügelmann: Eine solche Doppelstrategie ist übrigens auch in anderen sozialen Berufen üblich: Richter wenden allgemeine Gesetze an – aber sie orientieren sich auch an den Fallentscheidungen höherer Gerichte. Und Ärzte nutzen einerseits die Ergebnisse von repräsentativen Doppelblindstudien aus der medizinischen Forschung, andererseits ihre klinische Erfahrung aus der Behandlung einzelner Patienten.

Behnken: Ein Beispiel aus unserer Arbeit: Allseits wird das Zerbrechen der Bindungen über mehrere Generationen hinweg beklagt. Die quantitativen Befunde aus der NRW-Kids-Befragung zeigen aber, dass Großeltern für Kinder eine sehr hohe Bedeutung haben – fast so hoch wie die Eltern. Die LISA & KO-Befunde wiederum erhellen, was das qualitativ bedeutet, nämlich sehr Unterschiedliches – von den Kindern, die ganz bei ihren Großeltern wohnen, über die, deren Großeltern in der gemeinsamen Wohnung oder zumindest in demselben Haus leben, bis hin zu Großeltern im Nachbarhaus, in derselben Straße, dem gleichen Ortsteil – oder ganz woanders, aber mit häufigen wechselseitigen Besuchen. Mal kochen die Großeltern mittags für die Kinder, mal gehen Großvater oder Großmutter gemeinsam mit einem Enkelkind demselben Hobby nach, andere Großeltern beaufsichtigen die Hausaufgaben, wieder andere fahren mit ihren Enkeln ohne die Eltern in Urlaub. Uns interessiert diese Vielfalt, die sich hinter dem Etikett »Großeltern sind wichtig« verbirgt, und uns interessiert, was dieses Etikett für die Beteiligten konkret bedeutet.

Frage: Gibt es sonst noch Besonderheiten Ihrer Studien?

Behnken: Das öffentliche Bild heutiger Kindheit ist geprägt von Untersuchungen in Großstädten. Tatsächlich lebt aber nur ein Drittel der Kinder in Städten mit mehr als 500 000 Einwohnern. Bedingt durch die Lage unserer Universität haben einige unserer Projekte Kindheit und Jugend im kleinstädtisch-ländlichen Raum untersucht. Damit erschließen wir ganz andere Lebenswelten: Viele dieser Kinder spielen draußen, sind oft mit Freunden zusammen … Auch sie wachsen unter anderen Bedingungen auf als Kinder in den 50er oder 70er Jahren; doch unterscheiden sich ihre Lebensbedingungen teilweise deutlich von den in den Großstadtstudien dargestellten Kindheiten.

Heymann: Eine Besonderheit des Projekts LISA & KO ist sicherlich auch seine zeitliche Anlage. Die heutige öffentliche Forschungsförderung ist auf kurzfristige Förderung, z. B. einen Zeitraum von drei bis maximal sechs Jahren angelegt. Unser Projekt untersucht Kindheiten nun bereits im achten Jahr. So sind zweite und dritte Besu-

che bei den Kindern, jeweils mit einem zeitlichen Abstand von etwa zwei Jahren, und damit die Erstellung von Längsschnitten* in solchen Einzelfallstudien möglich. Obwohl lange bekannt ist, dass viele Entwicklungen und Veränderungen nur sinnvoll in Längsschnittstudien* zu erfassen sind, greifen Forscher doch immer wieder auf Querschnittstudien* zurück, weil diese nur eine knappe Datenerhebung voraussetzen und Ergebnisse somit schneller zu publizieren sind. So auch PISA.

Frage: Zu Beginn unseres Gesprächs kritisierten Sie, dass in PISA nur Leistungen und darin nur Leistungen bestimmter Fächer erfasst würden. Wo ist das Problem? Sind die Leistungen in Literacy, Mathematik und Science nicht besonders wichtig? Sie erfassen doch z. B. bei LISA & KO unter anderem auch die Lesefähigkeit und die mathematischen Fähigkeiten!

Behnken: Zunächst stört uns die dominante Fokussierung auf Leistungen. Der schulische Auftrag beschränkt sich nicht auf die Erhöhung von fachlichen Leistungen. Im Gegenteil: Demokratische Kompetenzen sind genauso zu entwickeln wie Bildungsaspekte jenseits von Problemlösen und Lesekompetenz. Diese mögen zwar bedeutsam sein, jedoch sollte ein demokratischer Staat wie Deutschland auch daran interessiert sein, wie es beispielsweise um die politische Bildung seiner Jugendlichen steht. Die CIVIC-Studie hat eben dies erfasst und mindestens ebenso wichtige Ergebnisse hervorgebracht wie PISA. Wahrgenommen wurde dies in der Öffentlichkeit jedoch kaum. Anders als in den Hauptfächern gibt es keine bildungspolitischen Maßnahmen vergleichbarer Dimension zur Reduktion von Fremdenfeindlichkeit, die in Deutschland im internationalen Vergleich immerhin am stärksten ausgeprägt ist.

Heymann: Gleichzeitig trägt die starke Betonung von sicherlich bedeutsamen Kompetenzkomponenten zur weiteren Verengung des schulischen Curriculums bei. Andere Anteile von Allgemeinbildung werden zugunsten dieser als zentral betrachteten Kompetenzen zurückgestellt. Wenn wir auch bei LISA & KO die Schriftsprach- und Mathematikkompetenzen erfassen, so machen wir dies nicht allein wegen der »objektiven Wichtigkeit« dieser Bereiche, sondern auch, um die Bedeutung zu untersuchen, die sie für Kinder und Jugendliche bei der Entwicklung von Interessen und Perspektiven für das weitere Leben haben.

Frage: Was sind denn aus Ihrer Sicht die wichtigsten Ergebnisse Ihrer Studien?

Heymann: Die heute oft zitierte »veränderte Kindheit« ist ein wenig hilfreiches Stereotyp. Es verdeckt die große Bandbreite an Unterschieden zwischen den Kindern, von denen jedes einzigartig ist: als Person, in seinen Fähigkeiten und Leistungen, …

Brügelmann: …und nicht selten in seinem Verhalten in verschiedenen Situationen widersprüchlich, etwa in seinen Leistungen in der Schule und außerhalb, …

Heymann: ...aber widersprüchlich auch in der Wahrnehmung durch verschiedene Personen, in den Informationen aus verschiedenen Quellen: Selbsteinschätzung, Schulnoten, Tests.

Behnken: Ja, Kinder und Jugendliche entwickeln außerhalb der Schule Kompetenzen, die in der Schule nicht zählen. Schon Grundschüler/innen wissen in ihren Interessensgebieten oft mehr als ihre Lehrer/innen. Andere schreiben zu Hause Geschichten, Gedichte und Sachtexte – aber Aufsätze in der Schule nur ungern und mit mäßigem Erfolg. Noch weniger zählen Fertigkeiten, die nicht zum schulischen Curriculum gehören: Graffiti sprayen, selbst Musik komponieren und spielen, sportliche Kunstfertigkeiten und praktische Kompetenzen wie das Computern.

4. PISA & Co.: Nutzen und Grenzen von Leistungsvergleichen auf Systemebene

PISA* und weitere internationale wie nationale Leistungsstudien, z. B. IGLU/PIRLS*, TIMSS*, LAU* oder KESS, haben unser Repertoire zur Evaluation von Schule und Unterricht forschungsmethodisch bereichert – und die bildungspolitische Wirksamkeit empirischer Bildungsforschung nachhaltig gestärkt. Man darf aber nicht übersehen, dass ihre Stärken, die wir in der Tat nutzen sollten, unvermeidlich mit spezifischen Schwächen verbunden sind.

Zudem ist der durch PISA geförderte Forschungsstil inzwischen zum Modell für Bildungsforschung allgemein und für die Evaluation nicht nur des Bildungssystems, sondern auch einzelner Schulen, Lehrer/innen und Schüler/innen geworden, sodass man sich auch mit der paradigmatischen Wirkung dieses Projekts auseinandersetzen muss. Wenn dieser Modell-Effekt gemeint ist, spreche ich im Folgenden von PISA & Co. Sonst konzentrieren sich meine Anmerkungen auf PISA selbst.

4.1 Warum wir PISA & Co. brauchen, die Ergebnisse aber nur mit Vorsicht nutzen sollten

Methodisch kontrollierte Untersuchungen an repräsentativ ausgewählten Stichproben* verringern die Gefahr zufälliger Eindrücke. Das ist ein großer Vorzug von PISA & Co. gegenüber der persönlichen Erfahrung und gegenüber der Untersuchung von Einzelfällen. Je mehr Personen einbezogen werden, desto anfälliger werden Studien aber für logistische Probleme bis hin zu systematischer Verzerrung durch Nicht-Teilnahme bestimmter Gruppen. Außerdem gilt aus ökonomischen Gründen: je größer die Zahl der Teilnehmer/innen, desto weniger Aspekte können (genau) erfasst werden – und desto stärker muss ihre Erhebung fokussiert und formalisiert werden.

Die Formalisierung der Erhebungen und ihrer Auswertung verspricht wiederum weniger Subjektivität bei der Wahrnehmung von Sachverhalten und bei ihrer Deutung. Auch bietet Standardisierung – zumindest auf der Oberfläche von Fragebogenantworten und Aufgabenlösungen – einfachere Möglichkeiten für Vergleiche. Bezahlt wird die (scheinbare) Eindeutigkeit aber mit einem Verlust an kontextueller Bedeutung (»Wie hat der Schüler die Aufgabe verstanden?«) und inhaltlicher Tiefe (»Wie ist er zu seiner Lösung gekommen?«). Standardisierung verhindert ein Aushandeln von Bedeutungen und privilegiert den Zugang der Forscher/innen sowie ihre Interpretation der Antworten oder Lösungen (z. B. Wahl bzw. Gestaltung der Testaufgaben, Entscheidung für bestimmte statistische Rechenverfahren).

Bildungspolitisch bedeuten PISA & Co. einen Gewinn. Die Öffentlichkeit ist in eindringlicher Schärfe mit Problemen unseres Schulwesens konfrontiert worden, die

lange verdrängt wurden, auch wenn sie unter Experten durchaus bekannt waren. So gehören die in Deutschland lebenden 15-Jährigen – gemessen an den eingesetzten Tests – nicht zu den besonders erfolgreichen Schüler/inne/n. Das gilt – trotz Verbesserung einiger Vergleichswerte über die Jahre – nicht nur insgesamt, sondern auch für die Vorzeigeschule Gymnasium, wenn man diese mit den entsprechenden Teilgruppen in anderen Ländern vergleicht. Im internationalen Vergleich lagen die deutschen 15-Jährigen 2000 unter dem Durchschnitt, und auch 2009 und 2012 reichte es trotz relativer Positionsgewinne nur zu einem Platz im – je nach Fach – (oberen) Durchschnitt. Besondere Probleme bestehen nach wie vor im unteren Leistungsbereich bei – je nach Schwellenwert – 15 bis 25 Prozent der Schüler/innen, denen durch die Tests elementare Schwierigkeiten mit der Schriftsprache, Mathematik und den Naturwissenschaften attestiert wurden. Die Zugehörigkeit zu dieser sogenannten Risikogruppe steht zudem in einem engen Zusammenhang mit der Herkunft aus der sozialen Unterschicht und mit dem Migrationshintergrund. Insofern hat PISA lange tabuisierte Gründe für die Benachteiligung verschiedener Gruppen wieder in die Diskussion gebracht, etwa die Selektionsmechanismen des zwei- bis fünfgliedrigen Schulsystems. Auch die Streuung zwischen Leistungsspitze und Leistungsende ist – trotz einer vergleichsweise gering besetzten Spitzengruppe im internationalen Vergleich – besonders groß.

Obwohl all diese Probleme hier und da auch vorher schon wahrgenommen und diskutiert wurden: Erst die Aufmachung von PISA als Leistungsolympiade mit Ranglisten hat solchen Einschätzungen Durchschlagskraft verliehen. Diese mediale Inszenierung war wirkungsvoll, in ihr stecken aber auch Probleme. Vor allem wegen der Weiterungen auf andere Bereiche. Denn PISA ist ein, wenn nicht *das* bekannteste Projekt beim Paradigmenwechsel von der Input- zur Output-Steuerung (→ »Output statt Input« 2004).

Schon diese einleitenden Hinweise machen deutlich: Die Evaluation von bildungspolitischen Programmen und unterrichtspraktischen Aktivitäten stellt komplexe Anforderungen, denen kein methodischer Ansatz allein gerecht werden kann. Gewinne in einer Dimension sind nicht ohne Verluste in einer anderen zu haben. Meine Vorbehalte bedeuten deshalb keine grundsätzliche Absage an diesen Ansatz. Es geht mir vielmehr darum, die Autorität, die diesem Paradigma teils von den Forschern selbst, teils in der Rezeption der Befunde von außen zugesprochen wird, zu relativieren. Denn PISA ist inzwischen zum Symbol für einen Paradigmenwechsel in der Evaluation von Schule und Unterricht geworden, der sogenannten Output-Steuerung, die testbare Fachleistungen zum Maßstab für die Qualität von Schule und Unterricht macht.

PISA fair zu beurteilen ist also nicht leicht. Das fängt mit der Frage an, wer oder was eigentlich PISA ist. Verschiedene Beteiligte – z. B. aus dem internationalen Konsortium und aus der deutschen Forschergruppe – äußern sich durchaus unterschiedlich. Zudem muss man unterscheiden zwischen Methodenartikeln der Autoren und grundsätzlichen Selbstdarstellungen sowie konkreten Auswertungen und deren Kommentierung durch Dritte. Letztere wiederum finden sich in Fachbeiträgen oder in Artikeln und Interviews aus Alltagsmedien. Schließlich gehört zu PISA auch die Rezeption

und Nutzung durch Fachkollegen einerseits, durch Journalisten andererseits bis hin zu Bildungspolitik, Schulverwaltung, Pädagogen in der Praxis und interessierten Laien. Denn Testprogramme können je nach gesellschaftlichem und politischem Kontext ganz unterschiedliche (Neben-)Wirkungen entfalten.

Vor allem die Art der Präsentation hatte und hat Nebenwirkungen, die bedacht werden müssen, wenn man Funktion und Anspruch einer solchen Bildungsforschung diskutiert. Dabei sind verschiedene Ebenen zu unterscheiden. Ein besonderes Problem ergibt sich daraus, dass PISA mal als Forschungsprogramm, mal als Evaluationsstudie präsentiert wird – obwohl jeweils andere Anforderungen gelten. Auch die PISA-Forscher selbst haben diese Probleme der Selbstdarstellung bzw. ihrer Rezeption zunehmend wahrgenommen, weshalb sie den Geltungsbereich konkreter Aussagen und die Wirkung daran anschließender Maßnahmen in neueren Veröffentlichungen deutlich vorsichtiger formulieren. Andererseits spielt auch die OECD in der öffentlichen Darstellung mit unterschiedlichen Geltungsansprüchen.

Man kann das Potenzial und die Grenzen des Ansatzes hinter PISA insofern aus verschiedenen Blickwinkeln diskutieren. Als konkretes Projekt untersucht PISA alle drei Jahre die Fachleistungen am Ende der Pflichtschulzeit in sorgfältig ausgewählten Stichproben*. Analog zu TIMSS*, IGLU* und anderen internationalen Leistungsstudien hat PISA damit den begrenzten Anspruch, einen kleinen Ausschnitt der Wirkungen von Schule über verschiedene Länder hinweg zu vergleichen (s. dazu Kap. 4.2). Im Verbund von VerA* und anderen flächendeckenden Lernstandserhebungen haben die Bundesländer diesen Ansatz allerdings zu einem Programm des System-Monitoring bis hin zur Schul- und Unterrichtsevaluation gemacht – zum Teil sogar mit dem Anspruch, den Leistungsstand einzelner Schüler/innen bewerten zu können (s. dazu Kap. 5 und 6). Insofern ist Standardisierung für viele zum generellen Paradigma der Evaluation von Unterricht und individuellen Lernprozessen geworden. Wenn von PISA & Co. gesprochen wird, ist diese Modellwirkung gemeint (s. dazu Kap. 4.3).

Die Reichweite der Methoden und damit der Geltungsbereich der Befunde ist aber begrenzt: Statt der vielfach behaupteten Evidenz-Basierung handelt es sich um (Ein-)Schätzungen, die in hilfreicher, aber begrenzter Weise empirisch fundiert sind. Insofern plädiere ich am Ende dieses Kapitels für eine gleichgewichtige Ergänzung des Repertoires durch komplementäre Formen der Evaluation. Dabei sehe ich durchaus, dass manche Kollegen mit derselben Intention für eine Stärkung des Standardisierungsparadigmas argumentieren. Da diese in den letzten Jahren in großer Breite und mit starker Wirkung (siehe etwa die Besetzung von Lehrstühlen in der Erziehungswissenschaft) stattgefunden hat, sehe ich aber eher die Notwendigkeit, die Balance in umgekehrter Richtung auszutarieren.

4.2 Stärken, Reichweite und Schwächen von PISA

Reichen die Fachleistungen zum Ende der Pflichtschulzeit aus, um die Qualität des Schulwesens zu beurteilen?

Den Blick der Schule auf das zu wenden, was Schüler/innen aus ihr mitnehmen, ist wichtig. Der kontinentaleuropäischen Tradition entsprechend haben sich im deutschen Bildungssystem Steuerung und Kontrolle lange Zeit auf Input und Rahmenbedingungen wie Lehrpläne, Stundentafeln und Qualifikation der Lehrerschaft beschränkt. Pragmatisch verständlich ist auch, dass sich PISA auf sogenannte Basiskompetenzen konzentriert, und das in den Schlüsselbereichen wie Lesen und Mathematik. Zugleich sind aber die Einschränkungen zu beachten, die mit diesem Fokus einhergehen.

So zeigt PISA nicht, wie gut – oder auch nur wie erfolgreich – ein Schulsystem ist, sondern was 15-Jährige eines Landes in PISA-Tests leisten. Dies ist ein wichtiger Unterschied. Denn die PISA-Aufgaben sind bewusst nicht auf das jeweilige Curriculum abgestimmt, untersuchen also nicht, wie gut ein Schulsystem seine selbst gesetzten Ziele erreicht.

Die Frage, ob Ländercurricula die PISA-Kompetenzen zureichend abbilden, ist zunächst einmal normativ zu diskutieren: Setzt sich ein System die richtigen Ziele? Nur scheinbar selbstverständlich ist es beispielsweise zu prüfen, ob Schüler/innen lesen können. Denn PISA unterstellt ein bestimmtes Verständnis von Lesen: die sogenannte funktionale »Literacy«, sprich den Umgang mit Schrift im Alltag, vor allem die Nutzung von Sachinformationen.

Wesentliche Erfahrungen mit Inhalten des Unterrichts machen junge Menschen außerhalb der Schule. Sie betreffen nicht nur die Entwicklung der Lesefähigkeit, sondern auch Kompetenzen in den Sachfächern und in Sport und Musik. PISA selbst verweist auf die hohe Abhängigkeit des Schulerfolgs von der sozialen Herkunft, also auf die Bedeutung von Normen und Lernmöglichkeiten im Alltag. Insofern dürfen beobachtete Leistungsunterschiede nicht allein der Qualität von Schule zugerechnet werden, lassen sie sich doch auch auf externe Faktoren wie Familie und Medien zurückführen. Soweit empirisch gestützte Schätzungen dazu vorliegen, wie weit sich Leistungsunterschiede durch schulische Einflüsse erklären lassen, schwanken sie sehr stark: zwischen weniger als 5 Prozent und 35 bis 40 Prozent. Der zum Teil überraschend niedrige Anteil wird plausibel, wenn man Schätzungen zur Bedeutung des informellen Lernens hinzunimmt. So finden einer Auswertung der einschlägigen Forschung zufolge (Dohmen 2001, S. 7) zwei Drittel bis drei Viertel aller Lernprozesse außerhalb der Schule statt.

Neben dem Unterricht spielen Systembedingungen wie Stundentafel oder Halbtags- bzw. Ganztagsschule eine große Rolle für die Leistungsentwicklung der Schüler/innen. Hinzu kommen außerschulische Faktoren, beispielsweise die Bildungseinstellungen der Eltern und die Arbeitsmarktchancen der Absolventen. Unterschiede allein der öf-

fentlichen Schule zuzurechnen verkennt auch die unterschiedliche Rolle von zusätzlichen, privat bezahlten Fördereinrichtungen wie etwa der Jukus in Japan oder des individuellen Nachhilfeunterrichts in Deutschland.

Die Qualität von Schule kann nicht allein an den messbaren Wirkungen von Tests festgemacht werden. *Wie* etwas gelernt wird, hat Bedeutung für die Dauerhaftigkeit und den Transfer von Wissen bzw. Können und bestimmt zudem die Erfahrung von Schule als Lernraum. Prozessqualität und Rahmenbedingungen stellen deshalb eigene Merkmale einer guten Schule dar. Aus problematischen Fachleistungen in Tests darf nicht vorschnell auf Schwächen des Unterrichts, auf Schwächen von Schulen oder Schulsystemen geschlossen werden. Mit der Erhebung der Leistungen bei 15-Jährigen erfasst PISA erworbene Kompetenzen nur punktuell. Wie aussagekräftig diese als Indikatoren für längerfristige Leistungsmöglichkeiten sind, muss gesondert nachgewiesen werden – was bisher fehlt.

Zwischenfazit: Durch die Prominenz der Leistungsstudien gerät Schule als Ort der Persönlichkeitsentwicklung, des sozialen und politischen Lernens, als Raum, in dem eine Gesellschaft durch die Begegnung ihrer Teilkulturen und der Generationen zusammenwächst (→ »Pädagogik von gestern – in der Schule heute – für ein Leben morgen?« 2010), in der öffentlichen Diskussion ebenso aus dem Blick wie das Kind »hinter PISA« mit seinen lebensweltlichen Erfahrungen (s. Kap. 3.2). Zudem lenkt der Fokus auf Unterricht bzw. auf das Bildungssystem als Ursache von Problemen allzu leicht ab von der Bedeutung der Wirtschaftspolitik und den Anforderungen an die Sozialpolitik, die die Rahmenbedingungen des Aufwachsens (und Lernens) von jungen Menschen in unserer Gesellschaft bestimmen.

Erfassen die PISA-Tests die Fachleistungen in angemessener Form?

Studien zur Notengebung zeigen seit vielen Jahren, wie fehleranfällig das Lehrerurteil ist (s. Kap. 6.1). Insofern liegt es nahe, nach Alternativen zu suchen. Standardisierte Tests haben mehrere Vorteile: Fokussierung der Datenerhebung, Transparenz der Anforderungen, Kalibrierung* der Maßstäbe (durch Bezug auf Normstichproben), Unabhängigkeit von persönlichen Besonderheiten. Insofern sind standardisierte Tests in heuristischer Funktion (eben als »Thermometer«, s. Kap. 2.2) stärker als bisher in das Repertoire pädagogischer Leistungsbeurteilung einzubeziehen.

Aber standardisierte Tests bedeuten auch eine Einschränkung der Lernbeobachtung und Leistungsbeurteilung durch die Auswahl bestimmter Fächer, durch die Beschränkung auf bestimmte Bereiche in diesen Fächern und durch die Begrenzung auf bestimmte Aufgabenformate. Zudem führt die bewusst begrenzte Zeit für das Lösen der Aufgaben dazu, dass geringe Prüfungsangst, Vertrautheit mit den Testformaten, testökonomisches Arbeiten und Verzicht auf Reflexion »belohnt« werden und damit die fachliche Kompetenz überlagern. Mädchen und Jugendliche aus schulfernen Milieus

schneiden bei solchen Leistungssituationen tendenziell schlechter ab als in Situationen ohne starken Zeitdruck.

Die hohen Korrelationen* zwischen den Lösungshäufigkeiten in Mathematik und Naturwissenschaften einerseits und im Lesen andererseits verweisen auf eine mangelnde fachliche Trennschärfe der Aufgabentypen – schon allein wegen der sprachlichen Einkleidung der Aufgaben in allen Bereichen. Schließlich darf nicht übersehen werden, dass der forschungsmethodische Zugang das inhaltliche Bild des untersuchten Gegenstands bestimmt. Wählt man beispielsweise in einer Untersuchung Test(typ) A, erhält man ein anderes Ergebnis, als wenn man Test(typ) B einsetzt.

Es muss nachdenklich stimmen, dass die USA bei internationalen Vergleichen der Leseleistung in den letzten 15 Jahren auf vorderen, auf mittleren und auf hinteren Plätzen gelandet sind:

Kohorte Schulanfang	Rangplatz	von xx Nationen	Studie
1934–1982	7	8	IALS
1988	2	16	IEA-GS
1991	15	31	PISA-1
1994	15	29	PISA-2
1998	4-12[5]	35	PIRLS

Diese Studien wurden mit unterschiedlichen Verfahren an jeweils anderen Stichproben* durchgeführt. Dagegen hat sich im US-internen *National Assessment of Educational Progress* das Niveau der Leseleistung über 30 Jahre hinweg weder in der Grundschule noch auf der Sekundarstufe bedeutsam verändert (vgl. NCES 2010). Diese Diskrepanz ist (neben dem Einfluss unterschiedlich gezogener Stichproben) nur erklärbar durch den Einsatz verschiedener Tests.

Ähnlich verwirrend ist das Bild innerhalb Deutschlands, wenn man sich anschaut, welchen Teilgruppen mangelnde Lesefähigkeit attestiert wird. Selbst wenn man nur die internationalen Lesestudien von 1991 bis 2001 betrachtet, werden in den verschiedenen Auswertungen als »Risikogruppe« zwischen 2 Prozent und 25 Prozent der einbezogenen Jahrgänge benannt:

Studie	Alter	Schulanfang	Jahr	Anteil »Risiko«
IALS	16+	1934-1982	1993	14 %
IEA-Sek	14	1983	1991	2 %
PISA-I	15	1991	2000	25 %
IGLU/PIRLS	9	1998	2001	10 %

5 Die Abstände zwischen den Rangplätzen im Bereich 4 bis 12 waren so gering, dass eine Abstufung nicht möglich war.

Kann sich die Quote innerhalb von zehn Jahren von 2 Prozent auf 25 Prozent mehr als verzehnfacht haben – um dann wieder in sieben Jahren auf 10 Prozent abzusinken? Plausibler ist folgende Erklärung: Je nach Aufgabentyp und je nach Definition der Schwellenwerte ergeben sich ganz unterschiedliche Anteile. Beispielsweise zeigt unsere LUST-Studie in verschiedenen Schulbezirken in Nordrhein-Westfalen, dass am Ende der 4. Klasse selbst in der Gruppe der unteren 5 Prozent viele Kinder mehrere Sätze pro Minute lesen können (Brügelmann 2003, S. 13, 16).

Entsprechend ist auch die Rangverbesserung deutscher 15-Jähriger von PISA-2003 auf PISA-2006 diskutierbar: »Danach erreichen die 15-jährigen Deutschen in der neuen Pisa-Untersuchung Rang 13 – von diesmal 57 Teilnehmerländern. 2003 hatte Deutschland noch auf Platz 18 unter 40 Staaten gelegen. Laut OECD sind beide Tests wegen ihrer geänderten Aufgabenstruktur allerdings nicht vergleichbar« (www.spiegel.de/schulspiegel/wissen/0,1518,520291,00.html). Nun kann man sagen: Die Deutschen haben sich nicht verbessert, der neue Test bevorteile sie – oder aber: Die Deutschen waren schon 2003 besser (sprich: Das damalige Testverfahren habe bestimmte Stärken deutscher Schüler/innen ausgeblendet und sie deshalb benachteiligt).

Ratzka (2003) hat Aufgaben aus der 1995 in deutschen Grundschulen nicht durchgeführten TIMS-Studie nachträglich in einer eigenen Stichprobe eingesetzt – parallel mit zwei anderen Tests. Sie fand, dass selbst bei einer groben Aufteilung nach Leistungsniveaus nur 41 Prozent der Schüler/innen in allen drei Tests in derselben Gruppe landeten und dass sogar beim gleichem Aufgabentyp (Textaufgaben) viele Schüler/innen in verschiedenen Tests unterschiedliche Ergebnisse erzielten – ja, sogar in demselben Test (den TIMSS-Aufgaben für die Grundschule), je nachdem, ob die Aufgaben mit oder ohne Zeitdruck zu bearbeiten waren.

Zwischenfazit: Die aufgaben- und kontextabhängige Ausschnitthaftigkeit von Instrumenten und ihren Ergebnissen muss in der öffentlichen Diskussion bewusst gemacht und gehalten werden. PISA & Co. dürfen nicht mehr über *die* Lese- und Mathematikleistung oder gar *die* Lesekompetenz der deutschen Schüler/innen urteilen – geschweige denn so tun, als ob sie mit ihrer Sondierung ausgewählter Wirkungsausschnitte die Qualität des vorhergegangenen Unterrichts oder gar den individuellen Lernstand einzelner Schüler/innen erfassen könnten.

Sind die Auswertungs- und Darstellungsformen der erhobenen Daten aussagekräftig und verständlich?

PISA hat Probleme von Teilgruppen ins öffentliche Bewusstsein gerückt, die vorher zwar in Fachkreisen bekannt waren, in der Bildungspolitik aber nicht ernst genommen wurden. Das betrifft die (schrift-)sprachlichen Schwächen von Jungen, gilt aber auch für die Leistungsrückstände von Migranten und generell von Kindern und Jugendlichen aus Familien mit niedrigem sozioökonomischen Status.

Gleichzeitig sind folgende Einschränkungen der Ergebnisse zu bedenken, wenn man PISA sinnvoll nutzen will:

Die Itemanalyse nach Kompetenzstufen mit einem unterstellten eindimensionalen Aufbau wird der komplex verschachtelten Entwicklung von Wissen und Können nicht gerecht und schränkt zudem das Spektrum möglicher Aufgaben ein.

Falsch vs. richtig auf der Oberfläche ist interpretationsbedürftig im Blick auf die Tiefenstruktur. Ohne Kenntnis der Lösungswege sind die Aufgabenlösungen mehrdeutig. Ob jemand z. B. ein bestimmtes Wort oder einen Text schnell und mit Verständnis lesen kann, hängt von seinen inhaltlichen Interessen, von seinem persönlichen Weltwissen und seinem Wortschatz ab, sodass falsche wie richtige Aufgabenlösungen sich nicht eindeutig bestimmten Formalstufen zuordnen lassen.

Durch die große Zahl der getesteten Schüler/innen werden auch inhaltlich irrelevante Unterschiede statistisch signifikant (s. zur statistischen Signifikanz* vs. praktischen Bedeutung von Unterschieden Kap. 2.3).

Um Ergebnisse aus verschiedenen Untersuchungen besser vergleichen zu können, werden die Punktwerte in einem Test oft auf eine Standardskala umgerechnet, auf der das arithmetische Mittel bei 500 angesetzt wird und zwei Drittel der Fälle innerhalb der sogenannten Standardabweichung* von +/- 100 liegen. Durch diese Spreizung können Punktdifferenzen (z. B. bei PISA) von Laien leicht überschätzt werden. Auch eine Steigerung der durchschnittlichen Fehlerquote um 50 Prozent bedeutet in einem Diktat mit 50 Wörtern bei einer Zunahme von zwei auf vier Fehler etwas völlig anderes als bei einer Zunahme von zehn auf 20 Fehler.

Trotz der methodischen Sorgfalt der Forscher/innen bleiben so viele potenzielle Fehlerquellen unkontrolliert, dass deren Einfluss auf die Länder- und Untergruppen-Ergebnisse die Rangfolgen kräftig durcheinander schütteln kann. Konkret: Wenn 9 Punkte Unterschied im Ländervergleich als statistisch signifikant gelten, dann braucht diese Differenz bei den richtigen Antworten nur 2 Prozent zu betragen. Das bedeutet bei 26 Aufgaben (in PISA-Mathematik-2003) eine halbe Aufgabe oder auf 100 Aufgaben hochgerechnet einen Unterschied von zwei richtig gelösten Aufgaben. Ist das eine pädagogisch oder politisch relevante Differenz? Gaeth (2009) hat an einem fiktiven Beispiel illustriert, wie die Übertragung von einstelligen Rohwerten* in die PISA-Skala mit einem Mittelwert* von 500 zu einer Verzerrung der Wahrnehmung führen kann, wenn die Qualität der Werte und der Bezugsrahmen (Höchst- und Mindestwerte) nicht explizit gemacht werden. Die Übersetzung der Punktwerte in Alltagsbegriffe (z. B. »45 Punkte entsprechen etwa einem Schuljahr«) macht Kennwerte zwar anschaulicher, kann aber auch in die Irre führen (s. das Beispiel in Kap. 2.3).

Auch die Wahl von Schwellenwerten ist diskussionsbedürftig, so z. B. die Bildung des OECD-Mittelwerts aus den Ergebnissen der teilnehmenden Staaten und nicht der Schüler/innen oder die Schnitte auf der kontinuierlichen Leistungsskala bei der Bildung von Kompetenzstufen.

Zwischenfazit: Tests sind wie Präzisionsgeräte, die eine Messung von minimalen Differenzen erlauben. Diese führen zu klaren Ranglisten, deren Unterschiede damit aber noch nicht im Alltag bedeutsam sind. Entsprechend vorsichtig sollte man Unterschiede zwischen Bildungssystemen bei PISA interpretieren. Und: *Zahlen sprechen nicht für sich.* Schon die Wahl der statistischen Verfahren, die Darstellung von Testergebnissen, ihre Übersetzung in Schaubilder und Texte sind immer Akte der Interpretation und können bei entsprechender Absicht zur Manipulation der Wahrnehmung genutzt werden. Dass PISA an vielen Stellen stark ist, liegt auch an der hohen *theoretischen* Kompetenz der beteiligten Forscher/innen – und an dem Reichtum anderer Studien, die sie zur Interpretation ihrer Ergebnisse heranziehen.

Bilden Veränderungen in PISA-Tests tatsächliche Zuwächse und Verluste verlässlich ab?

Die Wiederholung von Erhebungen im Drei-Jahres-Rhythmus ermöglicht Reihenanalysen, in denen Veränderungen erkannt werden, jedenfalls wenn sie deutlich aus dem Rahmen fallen. Ein Wechsel in den Rangplätzen kann allerdings unterschiedliche Gründe haben:

Da die Rohwerte in den Tests durch Bezug auf die Ergebnisse anderer Länder in PISA-Punkte umgerechnet werden, können positive Entwicklungen trotzdem zu einer (relativen) Verschlechterung führen (und umgekehrt) – je nachdem, wie sich die gesamte Stichprobe* verändert.

Die zunehmende Vertrautheit mit den Testformaten verändert die Erfolgschancen – vor allem unter dem Zeitdruck der Tests. Das ist für Länder wie Deutschland bedeutsam, in denen standardisierte Tests traditionell eine geringe Rolle im Schulalltag gespielt haben. Bessere Testergebnisse müssen deshalb nicht zwingend eine Folge besserer fachlicher Kompetenzen sein.

Werden die Ergebnisse in den Tests von Beteiligten als bedeutsam für ihre eigene Situation oder Position wahrgenommen, wächst die Versuchung einer gezielten Vorbereitung. Verlage werben mit Etiketten wie »PISA-fit« für ihre Materialien, Lehrer/innen bauen gezielt Übungen in den Unterricht ein. Mit Veränderungen in den inhaltlichen Schwerpunkten können fachliche Schwächen ausgeglichen, mit der Gewöhnung an bestimmte Aufgabenformate vorhandene Kompetenzen stärker zur Geltung gebracht werden, ohne dass sie sich tatsächlich verbessern.

Zwischenfazit: Eine Wiederholung der Erhebungen alle drei Jahre ist zu kurzfristig angelegt (vgl. die bereits erwähnte Konstanz der Testleistungen im *National Assessment of Educational Progress* in den USA über mehr als 30 Jahre hinweg, obwohl sich in dieser Zeit das Schulsystem erheblich verändert hat). Diese Erhebungen binden zudem Aufmerksamkeit sowie fachliche und finanzielle Ressourcen, die an anderer Stelle für die Entwicklung von Schule und Unterricht oft bitter vermisst werden.

Wie bedeutsam ist die Qualität von Schule für die Lebenschancen des Einzelnen und für die wirtschaftliche Entwicklung einer Gesellschaft?

Eine wichtige Begründung für PISA ist die behauptete Bedeutung seiner Ergebnisse, also des untersuchten Outputs des Bildungssystems, für die wirtschaftliche Entwicklung eines Landes und für die Berufschancen seiner Bürger. Statistiken sowohl zu den Arbeitslosenquoten als auch zum Einkommensniveau bestätigen, dass Ausbildung und Bildung in der Tat wichtige Ressourcen für den individuellen Lebens- und Berufserfolg sind. Zumindest als Korrelation* lässt sich auch ein Zusammenhang zwischen allgemeinem Bildungsniveau und Wohlstand einer Gesellschaft feststellen. Aber die Frage, was Ursache und was Wirkung ist, lässt sich nicht so leicht beantworten. Konkret bestätigt sich in den PISA-Ergebnissen die behauptete Abhängigkeit des Wirtschaftswachstums und der Arbeitslosigkeit von Leistungspunkten nur eingeschränkt. Der Einfluss von Drittfaktoren, die mit dem (Aus-)Bildungsniveau eines Landes korrelieren, darf insofern nicht übersehen werden.

Die Zusammenhänge zwischen dem PISA-Ranking und der Wirtschaftskraft von Ländern sind nicht sehr hoch. Sie werden durch weitere Bedingungen und Faktoren stark relativiert. Dazu zählen etwa politische Veränderungen wie der Zusammenbruch der Sowjetunion und sein Einfluss auf die volkswirtschaftliche Entwicklung im Ostseeraum. Beispielsweise sind die Korrelationen* zwischen Indikatoren für die Qualität des Bildungssystems und Indikatoren für die Wirtschaftskraft eines Landes höher, wenn man Industrieländer mit Entwicklungsländern vergleicht, als wenn man nur von den Randbedingungen her ähnliche Länder einbezieht.

Auch auf der Individualebene belegen Untersuchungen, dass Zusammenhänge zwischen Schulerfolg bzw. bestimmten Fachleistungen und späterem Berufserfolg bestehen. Aber die soziale Herkunft, Migrationshintergrund und Geschlecht spielen nicht nur für den Schulerfolg, sondern auch für die berufliche Laufbahn nach wie vor eine große Rolle – von der Bedeutung der sozialen Beziehungen ganz zu schweigen.

Der Einfluss von (Aus-)Bildung wird also in hohem Maße durch andere Bedingungen moderiert. Zu diesen gehören ebenfalls Konjunkturen einer Branche oder demografische Veränderungen. Zu bedenken ist zudem die ökonomische Entwertung von Qualifikationen durch deren Zunahme in der Breite: Dasselbe Zertifikat, aber auch dieselbe Kompetenz verliert auf einem Markt mit knappen Angeboten an Bedeutung, wenn mehr Menschen über sie verfügen. Aktuelle Wertigkeiten bestimmter Qualifikationen können deshalb nicht linear fortgeschrieben werden, auch wegen der Veränderung der Anforderungsprofile. Vor allem sind die Zusammenhänge nicht linear: So nimmt das Risiko von Arbeitslosigkeit bei Abschlüssen an Fachschulen über Fachhochschulen zu Universitäten nicht ab, sondern sogar geringfügig zu. Ein aktuelles Beispiel für die hohe Bedeutung von ökonomischen Kontextbedingungen: In China hat 2013 ein Drittel der Hochschulabsolventen keine Stelle bekommen – aber von den Abgängern der Berufsschulen wurden nur 2,5 Prozent arbeitslos (Schmidt 2014).

Die ökologische Validität standardisierter Tests, also ihre Gültigkeit für den Erfolg im Alltag, erweist sich generell als problematisch. Schon die Prognosekraft für den weiteren Schulerfolg und erst recht die Alltagsrelevanz von Testleistungen für den Berufs- und Lebenserfolg außerhalb der Schule sind für PISA nicht überzeugend belegt. In Dänemark finden sich z.B. unter den 17 Prozent der nach PISA angeblich funktional leseunfähigen »Illiterates« immer noch 20 Prozent erfolgreiche Sekundarabsolventen (Dolin 2007, S. 114); und in Kanada erwerben sogar 62 Prozent der Schüler/innen auf Lesestufe 0 und 1 einen Highschool-Abschluss (Knighton/Bussière 2006, S. 21). Andere Beispiele für die Schwierigkeit, aus Testleistungen auf die Alltagskompetenz zu schließen, sind die Unterschiede der Rechtschreibleistungen in freien Texten gegenüber Diktaten und die niedrigen Testleistungen vieler berufsfähiger Handwerker und sogar Lehrer/innen gegenüber Grundschüler/inne/n in der Lese-Studie LUST (Brügelmann 2004).

Zwischenfazit: PISA-Daten sind als Indikatoren für individuelle Leistungsmöglichkeiten außerhalb der Testsituation und als Indikatoren für gesellschaftliche Entwicklungen nur schwer einzuschätzen. Unabhängig davon ist zu bedenken: Die Vermittlung von *Qualifikationen* kann man instrumentell für ihre wirtschaftliche Verwertbarkeit untersuchen, betrachten und planen. *Bildung* als eine bestimmte Form der Aneignung von Welt(wissen) hat einen Eigenwert unabhängig vom Nutzen in Leistungssituationen.

Erlauben Unterschiede zwischen PISA-Kennwerten Vergleiche zwischen verschiedenen Bildungssystemen oder zumindest zwischen Teilgruppen innerhalb eines Systems?

Interkulturelle Vergleiche sind wegen der unterschiedlichen Bedeutung der Aufgabeninhalte generell problematisch (s. Kap. 2.3). Die Herkunft von Testitems kann deren Schwierigkeit für Schüler/innen aus verschiedenen Ländern unterschiedlich beeinflussen (Vertrautheit des Testformats; Bekanntheit der Inhalte; Verständlichkeit der Übersetzungen; Länge der Sätze; Wortfrequenz »derselben« Wörter in verschiedenen Sprachen; unterschiedliche Reaktionen auf Zeitdruck und den Eindeutigkeitszwang bei Multiple-Choice-Aufgaben). Selbst statistisch signifikante Unterschiede können verschwinden, wenn sich allein durch die Übersetzung, etwa vom Englischen ins Französische, die Textmenge und damit der Zeitbedarf für das Lesen der Aufgabe um 10 bis 20 Prozent erhöht (vgl. Wuttke 2007a, S. 257).

Einen Einfluss auf die Ergebnisse kann auch haben, wie die Testsituation und ihre Bedeutung von den Teilnehmer/inne/n in den einzelnen Ländern wahrgenommen und bewertet wird. Nicht zu vernachlässigen ist zudem, dass die deutschen 15-Jährigen in der Regel erst die 9. Klasse besucht haben, während sie sich in erfolgreichen Bildungssystemen wie Finnland, Korea oder Japan bereits im 10. oder 11. Schuljahrgang befinden. Damit schwindet die Bedeutung der Punktdifferenzen weiter – vor allem, wenn man an ihnen die Leistungsfähigkeit unterschiedlicher Bildungssysteme festmachen will. Vor

diesem Hintergrund kann auch die Veränderung der Altersstruktur durch den Systemwechsel von G9 auf G8 einen Einfluss auf die Verbesserung der deutschen Ergebnisse seit 2000 gehabt haben, ohne dass der Unterricht besser geworden sein muss.

Die Stichproben* in den einzelnen Ländern sind nicht immer vergleichbar, vor allem was die Einbeziehung schwächerer Schüler/innen betrifft. So verweist Wuttke (2006a, S. 143) auf die – je nach Land – unterschiedlichen Schulbesuchsquoten mit 15 Jahren, auf die unterschiedliche Einbeziehung von Schüler/inne/n der untersten Leistungsgruppen, auf die unterschiedlichen Teilnahmequoten und auf nicht plausible Verteilungen in einzelnen Ländern.

Selbst innerhalb eines nationalen Bildungssystems sind Gruppenvergleiche nicht einfach. Kategorien wie »Migrationshintergrund« oder »Gesamtschule« lassen sich zwar für Deutschland als Variablen operationalisieren, haben aber je nach Bundesland eine ganz unterschiedliche Bedeutung. Besonders deutlich wird das beim Vergleich der Migrantenanteile und ihrer unterschiedlichen Zusammensetzung (damit auch ihrer Bedeutung für die Leistungsverteilung) in den verschiedenen Bundesländern.

Das lässt sich am Bundesländer-Vergleich »PISA-E« von 2000 verdeutlichen. Presse und Bildungspolitik handelten damals fast ausschließlich mit den Punktwerten für den Durchschnitt der 15-Jährigen insgesamt: 510 Punkte in Bayern, 482 in Nordrhein-Westfalen und 448 in Bremen – eine Differenz von fast einer ganzen Kompetenzstufe oder rund zwei Schuljahren. Bezieht man sich auf relevante Untergruppen, so verändert sich das Bild aber beträchtlich – und zwar je nach Kategorie unterschiedlich stark. Vergleicht man nur die Kinder deutscher Staatsangehöriger, schrumpft der Rückstand von Nordrhein-Westfalen auf Bayern von 28 auf 12 Punkte, der Bremer Rückstand von 62 auf 41 Punkte. Bezieht man sich nur auf die 9. Klasse in Gymnasien und rechnet man zusätzlich die höhere Besuchsquote in Nordrhein-Westfalen und Bremen heraus, vergleicht man also gleiche Anteile des Jahrgangs in der gleichen Schulform, dann liegt Bremen nur noch 18 Punkte zurück und Nordrhein-Westfalen sogar rund 20 Punkte vor Bayern. Auch die oft gerühmte Förderung von Migranten in Bayern erscheint nicht mehr wesentlich besser als in Nordrhein-Westfalen, wenn man deren unterschiedliche soziale Zusammensetzung in beiden Bundesländern berücksichtigt (nur 14 statt 41 Punkte Differenz). Zu bedenken sind zusätzlich die unterschiedlichen Anteile verschiedener Nationalitäten. In Nordrhein-Westfalen und Bremen ist zum Beispiel der Anteil von türkischen Schüler/inne/n besonders hoch, die im Durchschnitt schlechtere Ergebnisse erzielen als etwa Spanier/innen oder gar Iraner/innen. Damit verändert sich die Situation noch einmal.

Sind Erklärungen möglich – und Beschreibungen ohne Erklärungen hilfreich?

Korrelationen* aus Querschnittuntersuchungen* erlauben per se keine Erklärungen. Sie suggerieren jedoch oft Kausalitäten und werden auch so gelesen. Allerdings kann

im Querschnitt* nicht geklärt werden, ob erhobene Variablen Ursache oder nur Indikator für Drittfaktoren sind. Verwirrung stiften immer wieder Korrelationsstudien zu Klassengröße und Schülerleistung. In ihnen wird beispielsweise oft nicht berücksichtigt, dass Schulleitungen Klassen verkleinern, wenn in der Lerngruppe besonders viele schwache Schüler/innen sind.

Selbst wenn plausible Erklärungen angeboten werden, die zu den Daten passen, gibt es oft weitere bedeutsame Randbedingungen, die genauso erklärungskräftig wären. Bezogen auf die Differenzen bei PISA-E 2000 hätte man etwa auf die Zahl der Deutschstunden in Klasse 1 bis 9 verweisen können (z. B. über 2 000 in Sachsen vs. weniger als 1 450 in Nordrhein-Westfalen) oder auf die unterschiedlichen Quoten der Sozialhilfeempfänger unter den Eltern (z. B. 2 Prozent in Bayern vs. 14 Prozent in Bremen). Je nachdem, wie man diesen und weitere Faktoren gewichtet, purzeln die Rangplätze und Abstände der Länder heftig durcheinander. Zwischen den entsprechenden Alternativmodellen fällt eine Entscheidung oft schwer. Zwar werden bei PISA über Fragebögen Daten zu einer Fülle von Bedingungen erhoben, aber forschungsökonomisch bedingt in einer zu abstrakten Form, um die Qualität der Lebens- und Lernbedingungen zureichend zu erfassen.

Zwischenfazit: Sowohl für die internationalen wie für die nationalen Vergleiche werden viele Ressourcen absorbiert. Oft wären Sekundär- und Meta-Analysen vorhandener Studien aussagekräftiger als neue Querschnittstudien. Für eine methodisch abgesicherte und inhaltlich ergiebige Interpretation der PISA-Korrelationen* sind sie ohnehin zusätzlich notwendig, wie die vielfältigen Verweise auf solche Studien in den PISA-Bänden zeigen (z. B. zum Sitzenbleiben, zur Bedeutung kleiner Klassen oder zum Einfluss der Vorschulerziehung auf den späteren Schulerfolg). Die dort vorgelegten Interpretationen und Folgerungen beruhen insofern nur zum Teil auf den PISA-Befunden selbst. Durch inhaltlich überzeugende Folgerungen entsteht aber der Eindruck, diese Erkenntnisse seien dem spezifischen forschungsmethodischen PISA-Ansatz zu verdanken, sodass dadurch dessen Ausweitung auf andere Ebenen der Evaluation legitimiert, ja geboten erscheint.

4.3 Eignet sich PISA als Paradigma für die Evaluation von Schule und Unterricht generell?

Die Bedeutung standardisierter Verfahren nimmt in der Schullandschaft zu: verpflichtende Sprachstandserhebungen bei Vier- oder Fünfjährigen vor Schulbeginn, die landesweiten Vergleichsarbeiten »VerA« in den Jahrgängen 3 und 8, zentrale Prüfungen am Ende der verschiedenen Schulzweige, mit Computerprogrammen auswertbare Beobachtungsraster der »Qualitätsanalyse« bei Inspektionen. Zu beobachten ist ein Multiplikationseffekt des Evaluationsstils von Projekten wie PISA unter dem Etikett »evidence-based policy«. Deshalb lohnen sich ein zweiter Blick und eine grundsätzlichere Diskussion dieses Ansatzes.

Mein Fazit aus dem letzten Kapitel: Die angeblich »objektive« Bildungsforschung ist nicht schon deshalb wertvoll, weil sie mit standardisierten Instrumenten eine große Zahl von Fällen erfasst. Sie ist es erst, wenn die Forscher/innen über sinnstiftende Theorien verfügen. Dabei können sie dann auch zu unterschiedlichen Einschätzungen kommen, was für Laien nicht immer so leicht sichtbar wird wie in der Kontroverse über die Bewertung der besseren Ergebnisse der deutschen Schüler/innen in den Naturwissenschaften bei PISA-2006 gegenüber 2003 und 2000.

Das Problem verschärft sich dort, wo es mehrere Untersuchungen zu demselben Thema gibt. Denn dann gibt es meist auch unterschiedliche Ergebnisse – wie im Alltag, wenn unterschiedliche Wahrnehmungen und Interpretationen konkurrieren. Und wie im Alltag wird dann auch in der Forschung über Gründe für diese Differenzen diskutiert – zwar mit fachbezogenen und mit forschungsmethodischen Argumenten, strukturell aber gleich, eben mit Argumenten. Wie hoch deren Plausibilität eingeschätzt wird, hängt auch hier vom persönlichen Urteil der Beteiligten ab.

Wie unterschiedlich dabei Befunde auch von renommierten Bildungsforscher/inne/n gewichtet werden, hat Rolff (2015) an der Frage illustriert, ob die Struktur des Bildungssystems (dreigliedrig vs. Gesamtschule) für dessen Erfolg bedeutsam sei oder nicht. Er moniert als erstes, dass meist nur die Durchschnittsleistung als Kriterium berücksichtigt wird und z. B. nicht auch das Ausmaß der Chancengleichheit. Mit Bezug auf die Interpretation von Hatties Meta-Meta-Analyse (s. Kap. 3) zeigt er aber auch, dass die Befunde selbst für das erste Kriterium nur formal und damit inhaltlich nicht zutreffend interpretiert werden.

Schon House (1980, S. 81; meine Übers.) stellte fest: »Aus quantitativen Studien kann man oft wertvolle Einsichten gewinnen, aber sie sind gewöhnlich die Folge davon, dass der Forscher intuitiv richtig urteilt, und weniger daraus, dass er richtig gerechnet hat. Diese Erfolge werden oft der quantitativen Methodologie selbst zugerechnet statt dem Urteil. Kritik bezieht sich gewöhnlich auf die technische Qualität der mathematischen Analyse, nicht auf die Qualität des Urteils, die mit ihrer Formulierung und Interpretation verbunden ist.«

Umso dringlicher ist zu fordern, dass schon in den Forschungsberichten selbst die Interpretationsbedürftigkeit von Daten sichtbar gemacht wird: »Aus empirisch gewonnenen Daten unmittelbar praktisches Handeln oder Handlungsoptionen abzuleiten, ist aus meiner Sicht ein Problem missverstandener Evidenzbasierung. Forschungsbefunde geben Anlässe zu Kommunikation in Interpretations-, Selektions- und Übertragungsprozessen bei durchaus unterschiedlichen Logiken politischen, wissenschaftlichen und praktischen Handelns« (Baumert 2014, S. 22).

Wichtiger als eine weitere technische Perfektionierung der Erhebungen wird damit die soziale Kontrolle ihrer Interpretation (s. Kap. 2.6). In dieser Hinsicht könnte die empirische Bildungsforschung viel aus Erfahrungen im Rechtswesen lernen. Juristen haben über Jahrhunderte hinweg eine Tradition der Evaluation von menschlichem Verhalten und sozialen Situationen entwickelt, und zwar in zweierlei Hinsicht: als empirische Feststellung von Sachverhalten, z. B. Würdigung von Zeugenaussagen und

von Sachverständigengutachten, sowie als ihre normative Bewertung, z. B. Auswahl und Gewichtung von einschlägigen Normen.

Das Fazit der rechtswissenschaftlichen Diskussion: Diese beiden Aktivitäten lassen sich methodisch zwar strukturieren, aber sie bleiben an die beteiligten Personen gebunden und lassen sich von der Ausschnitthaftigkeit und Perspektivität jeder Erkenntnis und jedes Urteils nicht befreien. Um die Gefahr zu minimieren, dass deren unvermeidliche Subjektivität einseitig durchschlägt, schaffen die Prozessordnungen ein System von institutionellen Kontrollen – ein Vorbild für die Evaluation des Schulwesens und einzelner Bildungseinrichtungen (→ »Towards checks and balances« 1974). So müssten von Anfang an alternative Deutungen und Folgerungen explizit gegeneinander gestellt werden, um den Anschein von Expertenautorität zu relativieren und damit auch der Öffentlichkeit die Interpretationsbedürftigkeit der nur scheinbar eindeutigen Daten bewusst zu machen.

Standardisierte Tests als Kontrollelement im Programm der Output-Steuerung

PISA hat sich inzwischen von einem spezifischen Projekt zu einem Modell der Leistungsmessung über das System-Monitoring hinaus gemausert. Alle Bundesländer führen flächendeckende Lernstandserhebungen in der Grundschule und in der Sekundarstufe durch. Damit einher geht ein Wechsel zur sogenannten Output-Steuerung, der mit einem Versagen der in Europa traditionell üblichen input-orientierten Maßnahmen begründet wird. Diese Input-Steuerung erfolgte über Lehrpläne, über die Qualifikationsanforderungen an Lehrer/innen und über die Prozesskontrolle durch die Schulaufsicht. Demgegenüber definiert die Output-Steuerung erwartete Leistungen im Voraus und überprüft deren Erreichen mit standardisierten Kompetenztests. Die Forderung nach einem Wechsel von der Input- zur Output-Steuerung folgt einer Argumentationsfigur, die sich kurz so zusammenfassen lässt:

- Deutschlands 15-Jährige haben bei PISA schlecht abgeschnitten.
- Deutschland hat ein Input-System.
- Viele der erfolgreicheren Länder haben eine Output-Steuerung.
- Also ist das Input-System schuld an der »PISA-Katastrophe«.
- Damit die Schülerleistungen besser werden, muss Deutschland auch ein Output-System einführen.
- Dessen Effektivität hat sich darin erwiesen, dass die Leistungen der deutschen Schüler/innen im Vergleich zu 2000 besser geworden sind.

Damit wird der Vorwurf erhoben, bisherige Formen der Steuerung und Evaluation hätten keine Veränderungen bewirkt. Dem ist entgegenzuhalten, dass sich im Grundschulbereich schon zwischen 1980 und 2000 vergleichsweise viel verändert hat. Genau diese evolutionären Reformen von unten drohen nun durch PISA & Co. verloren zu

gehen. Dennoch behaupten Bos u.a. (2007b, S. 1): »Wie die Erfahrungen aus den Ländern, die in internationalen Schulleistungsstudien regelmäßig sehr gut abschneiden (z.B. Kanada, England, Finnland, die Niederlande und Schweden), zeigen, wirkt sich die regelmäßige ›Outputkontrolle‹ lang- und mittelfristig ausgesprochen förderlich auf die Kompetenzentwicklung der Schülerschaft aus.«

Für Deutschland schätzen andere Forscher/innen die Folgen des Systemwechsels zu einer Output-Steuerung sehr viel zurückhaltender ein. Denn wie verträgt sich die Einschätzung von Bos u.a. mit den doch recht unterschiedlichen Befunden zum Erfolg des deutschen Bildungssystems, wie wir sie etwa in den verschiedenen internationalen Leseuntersuchungen finden? So hat Deutschland mit Input-Steuerung 2001 und 2006 in der Grundschulstudie IGLU* deutlich besser abgeschnitten als 2000 und 2006 bei PISA am Ende der Pflichtschulzeit – und parallel dazu die USA mit Output-System ebenfalls in der Grundschule besser als in der Sekundarstufe. Zweitens: Die USA haben zwar mit ihrem schon lange etablierten Output-System bei PISA 2000 besser abgeschnitten als Deutschland mit seinem Input-System, doch 2001 bei IGLU liegen beide gleichauf. Zudem haben sich die Schülerleistungen in den USA trotz jährlicher Schulleistungstests in den vergangenen 30 Jahren kaum verbessert (vgl. NCES 2010).

Weiter: Wie ist zu erklären, dass in Kanada mit Output-System verschiedene Provinzen bei IGLU-2006 von 533 bis 560 Punkte streuen und damit teilweise vor und teilweise hinter Deutschland (548 Punkte) liegen? Und innerhalb Deutschlands gibt es größere Unterschiede zwischen den Bundesländern, die alle input-orientiert gesteuert werden, als zwischen Deutschland und Vergleichsländern mit Output-Steuerung. Nachdenklich stimmt auch, dass Schweden, England und die Niederlande mit Output-System bei IGLU 2001 zur Spitzengruppe zählen, 2006 dagegen trotz der nun schon länger etablierten Output-Steuerung zurückfallen, dass sich dagegen die deutsche Grundschule unter dem Input-System von der IEA*-Studie 1991 bis zu IGLU 2001 rechnerisch pro Jahr doppelt so stark verbessert wie von 2001 bis 2006 unter der allmählichen Umstellung auf ein Output-System (Bos u.a. 2007a, S. 152) – während sich der Punktedurchschnitt der Sekundarstufe von PISA 2000 bis 2006 im gleichen Umstellungsprozess fast gar nicht verändert.

Das Problem mit der eingangs skizzierten Argumentation ist insgesamt: Nicht alle Länder, die besser als Deutschland abgeschnitten haben, sind output-gesteuert. Länder, die besser abgeschnitten haben, unterscheiden sich auch in weiteren Merkmalen vom deutschen System (Dauer der gemeinsamen Schulzeit; Zeitpunkt, ab dem benotet wird, usw.). Zudem zeigen sich in Ländern, die bei PISA oder IGLU besser als Deutschland abgeschnitten haben, in anderen Bereichen Probleme, die mit der Output-Orientierung verbunden sind, sodass dem behaupteten Nutzen auf der einen Seite Kosten auf der anderen gegenüberstehen, die man sich beim Transfer des Systems mit »einkauft«.

Insofern scheint die Leistungsfähigkeit eines Systems nur sehr begrenzt mit dem Steuerungsmodell zusammenzuhängen bzw. in starkem Maße durch Drittfaktoren beeinflusst. Hinzu kommt, dass der Begriff der »Steuerung« ein technisches Verständnis

gesellschaftlicher Interventionen unterstellt, das der nur sehr begrenzten Planbarkeit und Kontrollierbarkeit solch komplexer Systeme wie des Bildungswesens nicht gerecht wird. Wir haben es mit einem Paradox zu tun (Taleb 2014, S. 26): Je mehr wir wissen, desto komplexere – technische wie soziale – Systeme entwickeln wir; je komplexer ein System, desto schwieriger wird es, sein Verhalten vorauszusagen. Das heißt: je höher unser Wissensstand, desto geringer unsere Fähigkeit zur Vorhersage und Steuerung der Umwelt.

Selbst Hermann Lange, als Hamburger Staatsrat einer der Protagonisten der »empirischen Wende« im Verhältnis von Erziehungswissenschaft und Bildungspolitik, hat darauf verwiesen, »dass die Steuerung eines komplexen Systems nach dem einfachen Muster von Regelung und Ausführung nicht funktioniert. In der Organisationsliteratur wird deshalb auch gelegentlich davon gesprochen, dass man komplexe soziale Systeme nicht ›steuern‹, sondern allenfalls ›segeln‹ könne« (2000, S. 20 f., zit. nach Tillmann 2008, S. 1).

Alternative Perspektiven

Wie schon in Kapitel 2.7 erläutert, ist ein nach Stufen differenziertes System der Rechenschaftspflichten und Evaluationsformen nötig (vgl. zu kulturell bedingt unterschiedlichen Modellen in den OECD-Ländern → »Experimental decision making and responsive accountability« 1980). Ideen für ein demokratisches Verfahren, in dem die Sicht der Betroffenen vor Ort stärker berücksichtigt wird, skizziert Retzl (2010, S. 360-363). Außerdem ist es wichtig, zentrale System-Monitorings (über punktuelle Leistungstests) durch andere Indikatoren für die Qualität des Bildungswesens zu ergänzen, wie es »Bildung auf einen Blick« der OECD (2007b und folgende Jahre) und innerhalb Deutschlands das Konsortium Bildungsberichterstattung (2006 ff.) versuchen. Das aber reicht nicht aus.

Für die Entwicklung einzelner Schulen ist eine umfassendere und kontextsensible Evaluation notwendig, und die individuelle Leistungsrückmeldung ist zu einer »dialogischen Lernbeobachtung« anhand förderorientierter Aufgaben zu entwickeln, die einen Blick unter die Oberfläche des Messbaren erlauben (s. zum Konzept einer »Pädagogischen Leistungskultur« Kap. 6.3). Standardisierte Tests und flächendeckende Erhebungen können – bei entsprechender Einordnung – einen bedeutsamen Beitrag zur Einschätzung der Fachleistungen von Schulen und Schüler/inne/n leisten. Insofern ist es wichtig, einerseits die latente Test-Aversion vieler Lehrer/innen zu überwinden und ihnen andererseits Pools von (fachdidaktisch kommentierten!) Testaufgaben mit repräsentativen Vergleichswerten zur Verfügung zu stellen. PISA & Co. müssen inhaltlich und methodisch durch komplementäre Formen informeller Evaluation ergänzt werden. Vor allem ist es notwendig, die Evaluation vor Ort als eigenständige Aufgabe mit besonderen Anforderungen und Möglichkeiten zu fördern.

Damit dies keine abstrakte Forderung bleibt, stelle ich in Kapitel 5.3 den reformpädagogischen Schulverbund »Blick über den Zaun« als Beispiel für eine Evaluation durch Peer-Review auf Schul- und Unterrichtsebene vor.

4.4 Fazit

Politisch bedeutet die Zuständigkeit für Evaluation Macht – und damit für die Betroffenen auch Abhängigkeit. Man darf sie nicht allein Experten überlassen, denn die mit ihr verknüpften Probleme sind technisch nicht zu lösen.

Schule ist zwar ein öffentlicher Raum, und dieser Status schließt ein, dass alle Beteiligten rechenschaftspflichtig sind. Dafür brauchen sie aber je nach Aufgabe unterschiedliche Verfahren und Unterstützung. Ein demokratisches Verständnis von Evaluation fordert, die Betroffenen durch externe Beurteilungen nicht zu entmündigen, sondern sie in ihrer persönlichen Evaluations- und Problemlösekompetenz zu stärken.

Dezentrale Evaluation kann bildungspolitisch orientierte Studien wie PISA nicht ersetzen. Beide Ansätze haben ihre spezifische Funktion in einem umfassenderen Rechenschaftssystem. Dabei sind mir abschließend drei Punkte wichtig:

- Der PISA-Stil muss auf Systemevaluation begrenzt werden und darf nicht zum Paradigma für Evaluation und Bildungsforschung generell werden. Insbesondere sind der Sinn und die Form flächendeckender jährlicher Lernstandserhebungen zu überdenken (s. Kap. 5.2).
- Auch auf der Systemebene muss die Scheinpräzision von Zahlen aus Erhebungen mit Leistungstests immer wieder bewusst gemacht und ihre mehrperspektivische Deutung schon bei der Veröffentlichung gesichert werden (vgl. unseren Vorschlag für die Einrichtung einer → »Evaluationsstelle für nutzerorientierte Bildungsforschung« 2006).
- Die Ressourcen für Evaluation dürfen nicht auf die zentrale Evaluation konzentriert werden. Wir brauchen eine vergleichbare politische, wissenschaftliche und finanzielle Unterstützung für die Entwicklung der Evaluationskompetenz vor Ort – und dafür, Folgerungen aus den Ergebnissen ziehen zu können, z. B. für die Entwicklung schwacher Schulen und die Förderung von Kindern mit besonderen Schwierigkeiten.

PISA & Co. können unser Evaluationsrepertoire nur bereichern, wenn wir das Instrumentarium nicht überfordern.

5. Evaluation von Schule und Unterricht

5.1 Qualität und die Kunst, den Erfolg von Unterricht zu bestimmen

Meine zentrale Botschaft vorweg: Die Qualität von Schule lässt sich nur sehr begrenzt über die Evaluation ihrer »Produkte« bestimmen – vor allem wenn wir die Wirkungen so erfassen, wie dies gemeinhin geschieht: kurzfristig, ausschnitthaft und quantifiziert in Form von Testwerten oder gar Ziffernnoten.

Unterricht und Schulleben sind pädagogische Prozesse. Pädagogische Prozesse aber haben einen Eigenwert. Dieser besteht unabhängig von der Qualität der Produkte und ihrer Messbarkeit. Anders herum können Bedingungen und Maßnahmen nicht direkt bestimmten Wirkungen zugerechnet oder gar technologisch geplant werden. Um Unterricht und Schule angemessen beurteilen zu können, brauchen wir deshalb – neben tauglichen Messinstrumenten (auf deren technische Präzision sich die gegenwärtige Diskussion konzentriert) – plausible Begründungen für die Kriterien zur Beurteilung der Qualität pädagogischer Prozesse und ihrer Wirkungen sowie institutionell verfasste Verfahren zu ihrer fairen Bewertung.

Vor allem der output-orientierte Ansatz wird der Grundidee von Bildung nur sehr unzureichend gerecht. Kaum jemand käme auf den Gedanken, den Wert der Demokratie allein nach Kriterien ökonomischer Effektivität und Effizienz zu bestimmen. Die Entscheidung für eine demokratische Staatsform ist eine Wertentscheidung, die im Respekt vor dem Selbstbestimmungsrecht der Person begründet ist, die nicht für fremde Ziele instrumentalisiert werden darf. Nicht anders verhält es sich mit Bildung.

Damit sind wir bei einem zentralen Thema der Evaluationsdiskussion in den 70er Jahren. Bob Stake formulierte schon 1967 selbstkritisch: »Ein moderner Pädagoge kann sich wenig auf formale Evaluation verlassen, weil ihre Ergebnisse selten die von ihm gestellten Fragen beantworten« (Stake 1972, S. 93). Seine (und heute auch meine) Kritik: Wir konzentrieren unsere Instrumente auf die Dimension »Beschreibung«, operationalisieren unsere Erwartungen möglichst präzise und perfektionieren die Instrumente zu ihrer Überprüfung in der Realität, vernachlässigen aber die Auseinandersetzung über Normen und die Klärung von Verfahren zu ihrer Anwendung auf konkrete Sachverhalte.

Oft wird übersehen, dass Evaluation einen anderen Kontext darstellt als Grundlagenforschung. Qualität ist kein deskriptiv zu beschreibendes Faktum, sondern hat mit Werten zu tun, die eben auch die Prozesse und Bedingungen des Lernens betreffen. Auch »harte« Daten sprechen nicht für sich. Diese Einsicht vermisse ich in der Qualitätsdebatte, in der z. B. unterstellt wird, wir hätten Konsens und Klarheit darüber, was in der Schule zu lernen sei. Nur weil alle zustimmen, dass in der heutigen Gesellschaft

Lesen eine Schlüsselkompetenz darstellt, heißt das noch nicht, dass die Form, in der diese Schlüsselkompetenz bei PISA* oder VerA* erfasst wird, allseits Zustimmung findet.

Mythos 1: Die Ziele von Unterricht und die gesellschaftlichen Erwartungen an Schule sind klar, es gibt damit konsentierte Kriterien für die Evaluation

In der gegenwärtigen Diskussion scheint es so, als ob es nur um die Prüfung der Wirkungen von Unterricht ginge, als ob also die Maßstäbe, an denen diese Wirkungen zu messen sind, nach denen Erfolg zu beurteilen ist, klar wären. Wer in Richtlinien und Lehrplänen oder Bekundungen von Politikern liest, merkt schnell, dass schon die politischen Vorgaben sehr allgemein sind, dass sie auslegbare Kompromisse darstellen bzw. oft als Spannungen zwischen konkurrierenden Ansprüchen formuliert sind, die sich nicht im Sinne klarer Prioritäten auflösen lassen und nur einen Scheinkonsens darstellen.

Nicht anders sieht es aus, wenn man Befragungen von Betrieben oder Ausbilderinnen und Ausbildern auswertet, in denen nach den Erwartungen der »Abnehmer« an die Schulen gefragt wird. Erhebliche Unterschiede zeigen sich nicht nur zwischen verschiedenen Branchen (z. B. gewerblich-technisch vs. kaufmännisch), sondern auch innerhalb der Branchen, nicht nur in den Anforderungen der Betriebe, sondern auch in den Erfahrungen der Lehrlinge. Auch Befragungen von Eltern oder von repräsentativen Stichproben* der Bevölkerung zeigen unterschiedliche Erwartungen an die Allgemeinbildung.

Mit der Verabschiedung der sogenannten »Bildungsstandards« hat die Kultusministerkonferenz versucht, hier Klarheit zu schaffen. Die Standards beschränken sich aber auf fachbezogene Leistungen, werden also dem umfassenden Anspruch von *Bildungs*standards nicht gerecht. Wirksam für den Unterricht werden sie in Form der auf sie bezogenen Kompetenztests – und damit zusätzlich eingeschränkt auf Leistungen, die in diesen Formen messbar sind. Die Entwicklung von Standards in Verbindung mit zentralen Kompetenztests unterstellt zudem eine Plan- und Kontrollierbarkeit von Lernprozessen, die die technologischen Möglichkeiten von Didaktik überschätzt. Schließlich sind die Standards bewusst produktorientiert formuliert und vernachlässigen die Qualität der Lernprozesse, über die Kompetenzen erworben wurden. Dieselbe Leistung entspricht aber nicht ohne Weiteres derselben Kompetenz, je nachdem, auf welchem Weg und unter welchen Bedingungen sie entwickelt wurde. Lesenkönnen, das über einen kleinschrittigen Lehrgang vermittelt wurde, hat eine andere Qualität als Lesenkönnen, das sich durch selbstständige Lektüre selbst gewählter Texte entwickelt hat.

Mythos 2: Alle Schüler können alles lernen, alle Lehrer und alle Schulen können gut werden

Schon bei Comenius hieß der Anspruch: »omnes omnia omnino« (auf Deutsch: »alle alles allumfassend«). Die Erfahrungen mit der kompensatorischen Erziehung haben uns Bescheidenheit gelehrt (auch wenn ihr Scheitern etwas voreilig verkündet worden ist, s. u. S. 99). Insbesondere ist unsere Einsicht gewachsen, dass sich selbst bedeutsame Lernfortschritte nicht automatisch in besseren Noten oder höheren Berechtigungen niederschlagen. Die Selektionsfunktion des Schulwesens führt über die unterstellte Normalverteilung des Bewertungssystems dazu, dass nicht die faktische Bedeutsamkeit von Leistungsunterschieden, sondern ihr Beitrag zur Einstufung von Schüler/inne/n bestimmt, welche Noten und Berechtigungen vergeben werden. Ausschlaggebend ist nicht, ob sieben statt fünf Fehlern auf 100 Worte die Verständlichkeit eines Textes tatsächlich mehr beeinträchtigen, sondern die Anforderung, »gute« von »schlechten« Schüler/inne/n zu unterscheiden. So, wie im Leistungssport die Zunahme der Fähigkeiten auf höheren Wettbewerbsebenen dazu führt, dass nicht mehr Sekunden, sondern Zehntel- oder gar Tausendstelsekunden über Sieg oder Niederlage entscheiden, ist es auch in der Schule nach der Logik der Ziffernnoten nicht möglich, dass alle Schüler/innen »gut« sind – unabhängig von der Qualität des Unterrichts. Und so, wie der Lernzuwachs aller Schüler/innen die Preise für alle verdirbt, indem der Anspruch immer weiter nach oben geschraubt wird, kann auch der eine Lehrer nur auf Kosten seiner Kollegin »gut« sein, solange Erfolgskriterien nicht gruppenunabhängig durch sach- bzw. situationsbezogen bestimmte Anforderungen definiert werden. Davon sind wir aber weit entfernt, da die konjunkturelle Nachfrage in Berufen oder Ausbildungsgängen darüber bestimmt, ob viele oder nur wenige Absolventinnen und Absolventen gebraucht werden.

Das eine Problem ist also, dass nicht alle »gut« sein dürfen. Aber können sie es überhaupt? Die Klagen über unzureichende Leistungen von Schüler/inne/n und Schulen unterstellen ein Ideal, das meines Erachtens utopisch ist. Trotz aller Verklärung der Vergangenheit: Es hat nie ein Schulsystem gegeben, das durchgängig »gut« war, es gibt dieses auch nicht in anderen Ländern, und es gibt auch keinen anderen gesellschaftlichen Bereich, in dem ein System insgesamt zufriedenstellend funktioniert. Das Gesundheitswesen wie das Rechtswesen haben ihre weißen Flecken und kennen ihre schwarzen Schafe.

Gesellschaftspolitik lebt von der Illusion, dass sich erfolgreiche Inseln beliebig erweitern, dass sich Modelle sozusagen nach Blaupausen-Manier verbreiten lassen. Auch wenn Utopien als regulative Ideen wichtig sein mögen, um erwünschte Entwicklungen zu stützen, so ist eine zweite Perspektive nicht minder wichtig: Wir müssen lernen, mit den Unzulänglichkeiten von Systemen zu leben. So leidet das deutsche Schulsystem unter der Einbahnstraße »Lehrerbildung«: Der Abschluss berechtigt nur zu *einer* Tätigkeit, nämlich der Arbeit in der Schule, meist sogar nur in einer bestimmten Schulart. Auch innerhalb der Schule gibt es kaum Differenzierungsmöglichkeiten für unter-

schiedliche Fähigkeitsprofile. Wer etwa Probleme bei der Arbeit mit einer Lerngruppe hat, kann nicht ausweichen in andere Bereiche. Auch Kontrollsysteme greifen nicht, da ein Ausscheiden aus der Laufbahn nur unter extremen Bedingungen durchsetzbar ist. Als Auswege bleiben nur die Flucht in eine Krankheit oder die Verschiebung von Belastungen auf Kolleginnen und Kollegen.

Für die Entwicklung eines Rechenschaftssystems heißt das: Kontrolle bringt wenig, wenn es keine Maßnahmen zur Entwicklung von Qualität gibt. Für Lehrer/innen wie für Schüler/innen gilt: Statt ideale Normen als Anspruch an alle zu definieren, sind individuelle Zielvereinbarungen erforderlich, die konkret die nächsten Schritte bestimmen. Zwar muss deren Richtung durch eine übergreifende Normperspektive bestimmt werden, aber vom jeweiligen Entwicklungsstand aus auch realistische Erfolgsaussichten bieten.

Die Vorgabe von Niveaus zu festen Zeitpunkten (»Regelstandards«) verkennt die Entwicklungsdifferenz von mehreren Schuljahren in fachlichen Voraussetzungen schon am Schulanfang. Diese lässt sich über die Schulzeit hinweg nicht ausgleichen, weil fast alle Schüler/innen in vergleichbarem Maße dazulernen (»Karawaneneffekt«). Schulerfolg erweist sich in Lernzuwächsen, nicht im gleichzeitigen Erreichen gesetzter Lernziele.

Mythos 3: Die Qualität von Unterricht oder von Schule lässt sich durch Vergleiche der Schülerleistungen in Tests erfassen

Schon die Definition von Leistung ist schwierig. Wenn wir über Qualität und Leistung sprechen, sind drei Unterscheidungen wichtig:

- *Fachliche Leistungen der Schüler/innen vs. soziale und andere Kompetenzen*
  Der Bildungsauftrag der Schule beschränkt sich nicht auf Wissen und Können in den Schulfächern. Wenn Schüler/innen ökonomische Arbeitstechniken nutzen, wenn sie Methoden systematischen Denkens beherrschen, wenn sie aufeinander Rücksicht nehmen, dann sind das bedeutsame Leistungen, die sich nicht (direkt) über fachliches Wissen und Können erfassen lassen.
- *Leistung der Schüler/innen vs. Leistung der Schule*
  Die Fokussierung auf Fachtests suggeriert, die wesentliche Leistung der Schule lasse sich an den Leistungen der Schüler/innen messen. Aber wenn es in einer Schule wenig Gewalt gibt, wenn Mädchen und Jungen freundlich miteinander umgehen, wenn Eltern aktiv am Schulleben mitwirken, dann sind das wesentliche Leistungen der Schule, ohne dass sie sich in kurzfristig messbaren Leistungen der Schüler/innen niederschlagen (müssen).
- *Leistung der Schule vs. Qualität der Schule*
  Viele tun so, als ob sich die Qualität von Schule an ihren Leistungen ablesen lasse. Aber das Engagement von Lehrerinnen und Lehrern, die mit außerschulischen Einrichtungen und Personen zusammenarbeiten, die Ausstattung der Räume oder

das Angebot von Arbeitsgruppen sind Merkmale des Milieus und des Klimas einer Schule, die sich ebenfalls nicht direkt als Wirkungen auf Schüler/innen erfassen lassen. Untersuchungen zur Prognosefähigkeit von Noten zeigen, wie rasch sich der statistisch nachweisbare Zusammenhang selbst zwischen fachlichen Leistungen und außerschulischem Erfolg lockert, wenn man längerfristige Wirkungen zum Maßstab macht.

Ich habe verschiedentlich darauf hingewiesen, dass sich pädagogische Qualität nicht allein am Produkt festmachen lässt, sondern dass man auch Prozesse und Rahmenbedingungen berücksichtigen muss (z. B. Lernvoraussetzungen der Schülerinnen und Schüler, Zusammensetzung der Klassen und Ausfall von Unterricht oder Wechsel der Lehrerin bzw. des Lehrers). Wenn überhaupt, dann können nur individuelle Leistungszuwächse der Schule zugerechnet werden, nicht aber ein im Vergleich mit anderen Schulen höheres oder niedrigeres Leistungsniveau, bei dem die unterschiedlichen Lernvoraussetzungen der Schüler/innen unberücksichtigt bleiben.

Insbesondere aber ist die Art der Lehr-/Lernformen bedeutsam für die Bewertung von Unterricht. Wie bereits erläutert ist die Qualität derselben Leistung (z. B. die Lösung einer Rechenaufgabe) unterschiedlich einzustufen, je nachdem, aus welchem Lernprozess sie hervorgegangen ist (Abruf automatisierter Algorithmen, Anwendung vertrauter Verfahren oder sogar Transfer aus anderen).

Langfristige Ziele wie Kommunikations- und Kooperationsfähigkeit lassen sich nicht als Summe von Teilleistungen »erzeugen« und überprüfen. Solche komplexen Kompetenzen und Haltungen entwickeln sich, indem sie als alltägliche Praxis gefordert und in ihrer Entwicklung unterstützt werden.

In einer Demokratie haben auch schulische Aktivitäten die Grundrechte zu respektieren, zu denen der Anspruch auf Selbst- und Mitbestimmung gehört. Noch spezifischer sind die Vorgaben der UN-Kinderrechtskonvention zu beachten. Damit gewinnen Selbstständigkeit, Kritikfähigkeit, aber auch der Respekt für die Rechte anderer und die Fähigkeit zum Kompromiss eine eigenständige Bedeutung, was bei der Erfassung von Unterrichtsqualität zu berücksichtigen ist.

Hinzu kommt ein Problem, das bereits in der Kritik an einer eingeschränkten Produktorientierung kurz angesprochen wurde: die fragliche prognostische Validität von punktuellen Tests für den Schulerfolg. Warum lernen wir nicht aus alten Fehlern? Vor 50 Jahren setzte die Bildungspolitik in den USA auf Projekte »kompensatorischer Förderung« für Kinder aus schulfernen Milieus. Eines davon war das Head-Start-Programm. Bei seiner Evaluation kam es zunächst zu euphorischen Einschätzungen aufgrund der direkt nach Programmende gemessenen IQ-Zuwächse und der Überlegenheit der Fördergruppen gegenüber den Kontrollgruppen in zielbezogenen Leistungstests. Einige Jahre später dann die Ernüchterung, als die Differenzen in den entsprechenden Tests während der Grundschulzeit dahinschmolzen. Die bildungspolitische Folgerung: drastische Kürzung der Mittel. Erst Jahre später stellte man fest, dass die Programme nachhaltige langfristige Effekte auf die geförderten Kinder

hatten): weniger Überweisungen in Sonderschulen, weniger Klassenwiederholungen, einen höheren Anteil an Highschool-Abschlüssen. Auch später noch konnten im beruflichen und im privaten Leben bedeutsame Wirkungen nachgewiesen werden: höhere Einkommen, dreimal so oft Besitz eigener Häuser, halb so oft wegen Gesetzesverstößen inhaftiert, halb so oft arbeitslos, ein Drittel weniger uneheliche Geburten (Schweinhart u. a. 1993).

Zwischenfazit: Evaluation muss schon auf der »Produktebene« ein breites Spektrum an Aspekten und Kriterien erfassen, und sie muss diese Effekte vor dem Hintergrund der jeweiligen Unterrichtsprozesse und ihrer Rahmenbedingungen interpretieren. Auch »harte« Daten sprechen nicht für sich. Vor allem aber: Die Wirkungen einzelner pädagogischer Maßnahmen variieren konzept- und kontextabhängig, und sie lassen sich nicht nur kurzfristig und mit fachbezogenen Leistungstests erfassen. Produktbezogene Standards sind deshalb durch prozessorientierte Standards zu ergänzen, wie sie etwa für die Vergabe des Deutschen Schulpreises genutzt werden (vgl. Fauser u. a. 2007 ff.):

Leistung
Schulen, die gemessen an ihrer Ausgangslage besondere Schülerleistungen in den Kernfächern (Mathematik, Sprachen, Naturwissenschaften), im künstlerischen Bereich (z. B. Theater, Kunst, Musik oder Tanz), im Sport oder in anderen wichtigen Bereichen (z. B. Projektarbeit, Wettbewerbe) erzielen.

Umgang mit Vielfalt
Schulen, die Mittel und Wege gefunden haben, um produktiv mit den unterschiedlichen Bildungsvoraussetzungen, Interessen und Leistungsmöglichkeiten, mit der kulturellen und nationalen Herkunft, dem Bildungshintergrund der Familie und dem Geschlecht ihrer Schülerinnen und Schüler umzugehen,
Schulen, die wirksam zum Ausgleich von Benachteiligungen beitragen, Schulen, die das individuelle Lernen planvoll und kontinuierlich fördern.

Unterrichtsqualität
Schulen, die dafür sorgen, dass die Schüler ihr Lernen selbst in die Hand nehmen,
Schulen, die ein verständnisintensives und praxisorientiertes Lernen auch an außerschulischen Lernorten ermöglichen,
Schulen, die den Unterricht und die Arbeit von Lehrern mithilfe neuer Erkenntnisse kontinuierlich verbessern.

Verantwortung
Schulen, in denen ein achtungsvoller Umgang miteinander, gewaltfreie Konfliktlösungen und der sorgsame Umgang mit Sachen nicht nur postuliert, sondern gemeinsam vertreten und im Alltag verwirklicht wird,
Schulen, die Mitwirkung und demokratisches Engagement, Eigeninitiative und Gemeinsinn im Unterricht, in der Schule und über die Schule hinaus tatsächlich fordern und umsetzen.

Schulklima, Schulleben und außerschulische Partner
Schulen mit einem guten Klima und anregungsreichen Schulleben,
Schulen, in die Schüler, Lehrer und Eltern gern gehen,
Schulen, die pädagogisch fruchtbare Beziehungen zu außerschulischen Personen und Institutionen sowie zur Öffentlichkeit pflegen.

Schule als lernende Institution
Schulen, die neue und ergebnisorientierte Formen der Zusammenarbeit im Kollegium, der Führung und des demokratischen Managements praktizieren und die Motivation und Professionalität ihrer Lehrer planvoll fördern,
Schulen, die in der Bewältigung der Stofffülle, der Verbesserung des Lehrplans, der Organisation und Evaluation des Schulgeschehens eigene Aufgaben für sich erkennen und daran selbstständig und nachhaltig arbeiten.

In standardisierter Form verkürzen Raster der Schulinspektion diese Dimensionen (vgl. dazu die Beobachtungsbögen der Institute für Qualitätssicherung in den verschiedenen Bundesländern). Um auch die Qualität der Prozesse und Bedingungen angemessener zu erfassen, sind die Kriterien in den Standards des Schulverbunds »Blick über den Zaun« bewusst offener formuliert (s. Kap. 5.3).

Trotz solcher Relativierungen einer einseitigen Produktorientierung bleibt ein grundsätzliches Problem: die Unterstellung einer kausalen Beziehung zwischen bestimmten Lehraktivitäten und konkreten Lernfortschritten. Wie auch immer relativiert, ist eine deterministische Vorstellung von Pädagogik und damit verknüpft die Annahme, in einem Feld A entdeckte Beziehungen zwischen Bedingungen und Wirkungen ließen sich als Programm oder Methode in Feld B übertragen, äußerst problematisch.

Mythos 4: Gute Schulen sind die Kehrseite schlechter Schulen – indem wir Regelverstöße von Lehrern verhindern, fördern wir die Qualität ihres Unterrichts oder des Unterrichts generell

Vertrauen ist gut, Kontrolle ist besser – dieses Prinzip Lenins bestimmt auch die gegenwärtigen Vorstellungen über eine Rechenschaftspflicht der Schule. So richtig die Forderung ist, dass Verstöße gegen rechtliche Vorgaben nicht geduldet werden können – als Kern eines Rechenschaftssystems verkennt es die Möglichkeiten des Rechts. Die Kritik schlechter Zustände und Sanktionen mag berechtigt sein. Aber sie führt nicht – sozusagen spiegelbildlich – zur Entwicklung erwünschter Verhaltensweisen. Schon der Behaviorismus lehrt, dass sich unerwünschtes Verhalten durch Bestrafung nur unterdrücken lässt, dass Alternativen aber durch Verstärkung gesondert aufgebaut werden müssen.

Ein elaboriertes Familien-, Ehe- und Scheidungsrecht kann helfen, Konflikte so zu lösen, dass nicht ein Partner vom anderen drangsaliert wird. Ein liebevolles Verhältnis der Partnerinnen und Partner zueinander und zu ihren Kindern kann es aber nicht

sichern. Als Negativkontrolle kann das Recht eine Überschreitung von Grenzen verhindern, positive Impulse gibt es damit noch nicht. Insofern geht die unterstellte Gleichung »Verhinderung von Verstößen = Förderung von Qualität« nicht auf. Gegen Unpünktlichkeit von Lehrkräften oder Vandalismus von Schüler/inne/n vorzugehen ist notwendig, aber dies ist nur eine notwendige, nicht eine hinreichende Voraussetzung für eine gute Schule (möglicherweise sogar eher ihre Folge).

Eine gut funktionierende Schulaufsicht, die auf die Einhaltung der rechtlichen Vorgaben dringt und Verstöße ahndet, ist wichtig, um Minimalstandards zu sichern. Die Qualität von Unterricht zu entwickeln verlangt andere Mechanismen. *Payment by results* etwa hat in den angelsächsischen Ländern eine lange, aber nicht sehr erfolgreiche Geschichte. Wenn das Gesamtbudget fixiert ist, steht von vorneherein fest, dass die Hälfte der Lehrer/innen verlieren muss, gleichgültig, wie sehr sie sich um einen guten Unterricht bemüht. Denn Leistung wird relativ zur Gruppe beurteilt und Erfolg kann man nur haben, wenn andere schlechter sind (s. analog das Problem der Ziffernnoten in Kap. 6.1). Dies aber ist keine gute Voraussetzung für Zusammenarbeit und gegenseitige Unterstützung, zwei Prinzipien, die in der Forschung zur Schulentwicklung als Schlüsselvariablen betrachtet werden. Insofern gilt in Abwandlung des Lenin zugeschriebenen Satzes: Kontrolle verhindert Schlechtes, Vertrauen fördert Gutes.

Zwischenfazit: Neben der Kontrollorientierung brauchen wir im Schulsystem Instanzen und Mechanismen, die die Entwicklung von Qualität fördern. Dazu gehört vor allem die Institutionalisierung von Verfahren der Selbstevaluation – angefangen bei den Schülerinnen und Schülern, über die Lehrer/innen bis hin zum Kollegium und der Schule als ganzer. Damit werden weder eine formalisierte Fremdevaluation noch standardisierte Erhebungen überflüssig. Aber sie werden eingeordnet und auf ihre leistbaren Funktionen in einem differenzierten Verbund spezialisierter Verantwortlichkeiten und Verfahren beschränkt (vgl. die konkreten Hinweise in Kap. 2.7).

5.2 Wahrheit durch VerA?

Im Folgenden geht es um die Rolle von standardisierten Vergleichsarbeiten für die Schul- und Unterrichtsentwicklung. Die in Klasse 3 und Klasse 8 unter dem Stichwort VerA* (= **Verg**leichs**A**rbeiten) durchgeführten Lernstandserhebungen verursachen jährlich neue Aufregung und haben seitens der Lehrerschaft bzw. ihrer Verbände vielfach heftige Kritik ausgelöst. Meine Reaktion ist ambivalent: Als Forscher habe ich Respekt davor, was die VerA-Kollegen unter den gegebenen Rahmenbedingungen geleistet haben. Gleichzeitig frage ich mich, ob es für eine solche Evaluation richtig war, sich auf die politischen und organisatorischen Rahmenbedingungen einzulassen. Funktion, Anlage und Durchführung der Erhebungen haben darunter gelitten. Doch vorweg einige grundsätzliche Anmerkungen.

Gleiche Aufgaben bedeuten nicht gleiche Anforderungen

Flächendeckende Lernstandserhebungen sind nicht neu. Schon im Frankreich der 1920er Jahre wurden am Ende der Grundschulzeit standardisierte Leistungsprüfungen durchgeführt. 70 Jahre später sind die Abschlussprüfungen von 1925 in zwei Regionen wiederholt worden (Fritz-Vannahme 1996). Insgesamt rund 6 000 Kinder, zur einen Hälfte aus demselben Bezirk, zur anderen aus einer Region, die heute als sozialstrukturell vergleichbar gilt, hatten Aufgaben aus 25 (Teil-)Fächern zu bearbeiten. Das Ergebnis: In 15 Disziplinen waren die Leistungen 1995 gleichwertig, in zwei Disziplinen, darunter im Aufsatz, fielen sie 1995 besser aus. Allerdings hatten 1925 die Schüler/innen in acht Disziplinen, darunter Multiplikation und Rechtschreibung, die Nase vorn. Ein Beweis für den oft beschworenen Leistungsverfall? Oder gar kein so schlechtes Ergebnis für die heutige Jugend?

So einfach ist diese Frage nicht zu beantworten. Die scheinbar klaren Zahlen verdecken nämlich eine Reihe von Störfaktoren. Dies verdeutlicht ein Gedankenexperiment: Wie hätten wohl Schüler/innen von 1925 abgeschnitten, wenn sie heute an einer aktuellen Lernstandserhebung teilnehmen müssten:

- in Fachgebieten wie Geometrie oder Englisch, die damals in der Grundschule noch unbekannt waren;
- mit Themen, die im Lehrplan vor 50 Jahren noch gar nicht vorkamen – vor allem im naturwissenschaftlichen Unterricht;
- mit Inhalten, die ihnen damals weder aus dem Alltag noch aus der Schule vertraut waren, und
- mit Aufgaben, deren Form und Sprache ihnen häufig unbekannt oder zumindest fremd sein mussten, z. B. beim Sachrechnen?

An diesem Gedankenexperiment wird deutlich, wie schwierig ein Leistungsvergleich mit den (insofern nur scheinbar) »gleichen« Aufgaben über einen Zeitraum von 70 Jahren ist. Es kommt hinzu, dass die gesellschaftliche und motivationale Wertigkeit einzelner Inhalte und Fähigkeiten nicht dieselbe ist: Nicht nur, was im Unterricht intensiv bearbeitet und geübt wurde, auch welche Rolle diese Inhalte und Fähigkeiten in der Wahrnehmung der Kinder und ihrer Eltern spielen, hat sich nachhaltig verändert. Diese außerschulische Bedeutung ist für den Lernerfolg aber von großer Bedeutung. So war zu Beginn des vorigen Jahrhunderts die Rechtschreibung ein Schlüssel zum sozialen Aufstieg, und »fehlerfreie Schönschrift machte aus Bauernkindern kleine Angestellte« (Fritz-Vannahme 1996). Heute sind andere Fähigkeiten wichtiger. Gleiche Aufgaben bedeuten insofern nicht gleiche Anforderungen.

Ein grundsätzliches Problem teilen Lernstandserhebungen wie VerA* mit PISA*: Die Fokussierung auf den Output lässt andere wichtige Merkmale der Qualität von Schule und Unterricht außer Acht. Evaluation wird auf »measurement« verkürzt (dazu noch auf *ein* Messmodell). Damit gewinnen sehr spezifische teststatistische Anforderungen Vorrang vor inhaltlicher Relevanz: nur wenige Fächer, passende Inhalte und Aufga-

benformate in diesen Fächern, Auswertung nach einseitig bestimmter »Richtigkeit« der Lösung statt nach Produktivität des Lösungswegs.

Es gibt aber auch wesentliche Unterschiede zwischen VerA und PISA: flächendeckende Durchführung statt bloß Stichproben*, jährliche Wiederholung statt Drei-Jahres-Rhythmus – und vor allem Identifizierbarkeit der beteiligten Schulen, Lehrer/innen und Schüler/innen. Diese Merkmale erzeugen zusätzliche Probleme: Die flächendeckende Durchführung ist sehr viel aufwendiger und zugleich fehleranfälliger. Eine jährliche Wiederholung wäre angesichts der Trägheit von Schulsystemen nicht nötig, zieht sie doch finanzielle und personelle Ressourcen von anderen Aufgaben ab. Ferner hat sie sich für alle Beteiligten als sehr belastend erwiesen – zumal die Ergebnisse von vielen nicht als Hilfe erlebt werden, da ihr diagnostischer Ertrag für die Förderplanung gering ist.

Schließlich führt die fehlende Anonymisierung zu all den unerwünschten Nebenwirkungen, die aus angelsächsischen Studien zum »High-Stakes-Testing« bekannt sind. Dabei geht es um zentrale Leistungsüberprüfungen, von deren Ergebnis für die Beteiligten viel abhängt: der Zugang zu Elite(hoch)schulen für Schüler/innen, die Gehaltshöhe für Lehrer/innen, Sonderzuweisungen an Schulbezirke. Beklagt werden insbesondere eine fachliche Verengung des Curriculums vor Ort und eine Ausrichtung des Unterrichts an Testformaten. Zwar verweisen Klieme/Maichle (1990) darauf, dass ein gezieltes Testtraining keine Vorteile verschaffe. Aber ihr Versuch hat sich auf eine nur sechsstündige Einübung in die Testbearbeitung (zudem bei Studierenden) beschränkt, lässt sich also nicht mit den Wirkungen eines insgesamt auf »High-Stakes-Testing« abgestimmten Schulsystems vergleichen. Dort kommt es zu einer unterschiedlichen Auslegung und Durchführung der Aufgaben durch die Lehrer/innen bis hin zu bewussten Täuschungsversuchen.

Insbesondere Studien in den USA belegen, dass die Schulleistungen in Bundesstaaten mit hohen Sanktionen für Institutionen oder Personengruppen in der Regel schlechter sind als in Bundesstaaten mit niedrigen Sanktionen: Über dem nationalen Durchschnitt liegen 60 bis 70 Prozent der Bundesstaaten mit niedrigen Sanktionen, aber nur 10 bis 20 Prozent der Staaten mit hohen Sanktionen (vgl. Sacks 1999, S. 89 f.). Und wo die Leistungen in standardbezogenen Tests zunehmen, fallen sie mehrheitlich in unabhängigen Tests ab. So zeigt eine Studie von Amrein/Berliner (2002), dass der berichtete Leistungsanstieg in der Regel nur für den engen Bereich der im jeweiligen Bundesstaat etablierten Tests galt. In unabhängigen Tests, z.B. für die Zulassung zu den Hochschulen, wurde für zwei Drittel der 28 Staaten eine Abnahme der Testleistungen festgestellt. Zusätzlich wurden wachsende Dropout-Raten* berichtet, das heißt, dass leistungsschwächere Schüler/innen aus dem System ganz herausfielen, was die gemessenen Testleistungen zusätzlich in die Höhe trieb, ohne dass sich der Output tatsächlich verbesserte. Kirkland (1971) und Linn (2000) konnten sogar zeigen, dass die Leistungen nur so lange »stiegen«, wie sich der Test nicht änderte. Bei Einführung einer neuen Version desselben Tests sanken sie wieder auf das Ursprungsniveau. Danach »stiegen« sie erneut – bis die erste Testform wieder eingesetzt wurde und die

Leistungen ein zweites Mal auf ihr Ursprungsniveau »sanken«. Diese Befunde machen deutlich, wie sehr Sanktionen zu einem »Teaching to the Test« verführen, das die angestrebte inhaltliche Verbesserung des Unterrichts überlagern oder sogar gefährden kann. Das belegen auch Studien in Deutschland, die zeigen, wie unterschiedlich Lehrpersonen das Testprogramm wahrnehmen (Böttcher 2013; Demmer 2014; Richter u. a. 2014).

Gefährlich wird es vor allem, wenn Testformate Unterrichtsmuster bestimmen. Zwar können sie auch positive Impulse geben. So ist es ein Verdienst von VerA, dass die Lesefähigkeit in der Grundschule nicht mehr am Vorlesen geübter Texte überprüft wird, sondern durch Verständnisfragen zum Inhalt nach leisem Lesen. Problematisch wird aber auch dieses Format, wenn es im Unterricht nur noch um ein »richtig« oder »falsch« bei standardisiert vorgegebenen Fragen und Antworten geht und damit die persönliche Sicht auf eine Geschichte oder gar ein Gedicht unberücksichtigt bleibt. Schon Anfang der 70er Jahre hat Horst Rumpf (1971) solche Reduktionen an vielen Beispielen eindrücklich kritisiert.

Besondere Probleme ergeben sich aus dem Anspruch, Aussagen über die Qualität des Unterrichts einzelner Lehrer/innen oder über den Lernstand von Schüler/inne/n machen zu können. Damit wird der Kredit von Tests überzogen, wie die folgenden Hinweise zeigen.

Was sollen und was können Testprogramme für die Arbeit in der Schule leisten?

Drei Funktionen von VerA* sind in der Außendarstellung von verschiedenen Beteiligten unterschiedlich stark betont worden und faktisch auch unterschiedlich gut einlösbar:

- Bei der Bestandsaufnahme von grundlegenden Leistungen auf Landesebene (»System-Monitoring«) hat das deutsche Schulsystem tatsächlich einen Nachholbedarf. Diese Funktion können standardisierte Testprogramme wie PISA erfüllen, die sich auf Stichproben* beschränken. Aufwendige flächendeckende Lernstandserhebungen wie VerA sind nicht erforderlich. Es reicht außerdem, solche Untersuchungen alle fünf bis sieben Jahre durchzuführen. Denn selbst bei scheinbar tief greifenden Reformen erweist sich das Bildungswesen als ein vergleichsweise träges System.
- Auch die Rückmeldung der Ergebnisse an einzelne Schulen und Lehrer/innen ist hilfreich. Zum einen bekommen Lehrer/innen Anhaltspunkte, um besser einzuschätzen, in welchen Bereichen ihre Klassen relativ zu anderen Lerngruppen Stärken und Schwächen haben. Allerdings lassen diese nicht direkte Rückschlüsse auf die Qualität des Unterrichts zu – diese kann bedingt sein durch besondere Vorerfahrungen der Kinder, durch unterschiedliche Schwerpunkte im Unterricht, durch die Anlage der verwendeten Schulbücher oder andere Faktoren. Deren jeweilige Bedeutung ist insofern nur vor Ort von den Beteiligten selbst zu klären.

Zum anderen können Lehrer/innen ihre Anforderungen an die Kinder und ihre Maßstäbe bei der Leistungsbeurteilung im Vergleich mit Klassen, die unter ähnlichen Bedingungen arbeiten (»Referenzschulen«), überprüfen. Dies darf aber nicht einfach eine Anpassung an die externen Kriterien bedeuten, sondern verlangt eine Reflexion der externen *und* der eigenen Annahmen.

- Der Anspruch einer Diagnose des Lernstands einzelner Kinder überfordert die Instrumente völlig (ausführlicher Kap. 6). Eine punktuelle Messung muss inhaltlich auf wenige Ausschnitte begrenzt werden. Zudem ist sie immer fehlerbehaftet, das heißt, der festgestellte Wert kann nur als Anhaltspunkt für eine Bandbreite genommen werden, innerhalb derer der »wahre« Wert mit einiger Sicherheit liegt. Je mehr man das Fehlerrisiko minimieren will, desto breiter muss man die Bandbreite möglicher Schwankungen ansetzen – und umso weniger hilfreich ist dann das Ergebnis.
  In Kennwerten für Gruppen, also in Werten für ganze Klassen oder gar ein Bundesland, gleichen sich individuelle Schwankungen weitgehend aus, sodass das Fehlerrisiko der jeweiligen Durchschnittswerte entsprechend gering ist. Einmalige Tests bei einzelnen Personen bieten dagegen nur grobe Annäherungen an die tatsächliche Leistung. Sinnvoll nutzen lassen sich die Ergebnisse trotzdem, wenn man Abweichungen zur eigenen Einschätzung als »Warnlampe« nutzt, also als Anlass, um die Differenzen durch weitergehende Beobachtungen aufzuklären.

Für alle drei Ebenen gilt gleichermaßen: Die Ergebnisse sind als *ein* Element in einem umfassenderen Rechenschaftssystem zu sehen – als ein wichtiges und bisher unterrepräsentiertes Evaluationsinstrument, aber auch als ein in seiner Aussagekraft und Geltung begrenztes. Das größte Problem in der seit TIMSS* und PISA* öffentlich geführten Schuldebatte ist, dass sie auf den Vergleich von Punktwerten schrumpft und die Testautorität überhöht. Politiker/innen und Medien schauen nur noch auf den Output. Berichte der Schulaufsicht oder Forschungsergebnisse zu Lehr-/Lernprozessen und ihren Bedingungen verlieren an Bedeutung. Auch viele Lehrer/innen nehmen die Vergleichsdaten nicht als *eine* nützliche Zusatzinformation über Stärken und Schwächen ihrer Klasse wahr, sondern trauen oft ihrem eigenen Urteil nicht mehr, auch wenn es aus einer längerfristigen Erfahrung erwächst. Testergebnisse einzelner Kinder werden für Eltern (und oft auch für Lehrer/innen) zum »wahren« Wert für ihre Leistung, statt zu einem Indikator, der interpretationsbedürftig ist.

Konkrete Anmerkungen zu den Aufgaben und Verfahren von VerA

Analysiert man die Test- und die Begleitunterlagen, unterhält man sich mit Kolleginnen aus der Grundschulpädagogik oder der Fachdidaktik und befragt man Lehrer/innen aus den beteiligten Bundesländern nach ihren Erfahrungen mit VerA*, so stößt man auf sehr unterschiedliche Reaktionen. Manche Kolleginnen, die VerA im oben genannten Sinn als *ein* Element im Repertoire verschiedener Informationsquellen ver-

stehen und nutzen, finden die Tests hilfreich. Auch viele Kinder haben die Aufgabe, zu zeigen, was sie können, als Herausforderung positiv angenommen. Andere, vor allem leistungsschwächere Schüler/innen und Kinder mit Migrationshintergrund, erlebten viele Aufgaben und den Umfang insgesamt als völlige Überforderung. Darüber hinaus gibt es eine Reihe von Problemen, die bei der Weiterentwicklung des Instrumentariums und seinem zukünftigen Einsatz bedacht werden muss.

Zum inhaltlichen Ansatz

VerA-Aufgaben bieten zum Teil durchaus interessante Anregungen für Leistungskontrollen, die Lehrer/innen nutzen sollten, um ihr eigenes Repertoire zu erweitern. Es sei nur daran erinnert, dass in den Schulen jahrzehntelang die Leseleistung an der Fehlerfreiheit des Vorlesens geübter Texte gemessen wurde. Das Textverständnis bei leisem Lesen zu überprüfen, hat sich erst durch die entsprechenden Testaufgaben in der Praxis verbreitet. Die Aufgabenformate der Tests bringen aber – wegen der notwendigen Standardisierung von Durchführung und Auswertung – unvermeidlich Einschränkungen mit sich.

Aufgaben ohne sinnvollen Kontext (der in den Lehr- und Rahmenplänen für Deutsch überall gefordert wird) sind für viele Schüler/innen überraschend. Sie verändern auch das Lösungsverhalten, z. B. wenn in 20 Minuten ein Text zu einem Thema geschrieben werden soll, obwohl den Schüler/inne/n beigebracht worden ist, dass das Schreiben guter Texte ein Brainstorming, eine Textplanung, mehrere Zyklen der Überarbeitung (z. B. in Schreibkonferenzen) und insgesamt eine rege Kommunikation mit anderen voraussetzt. Dagegen untersagen die Instruktionen bei VerA ausdrücklich Fragen an die Lehrerin und Gespräche der Kinder untereinander.

Die Richtigkeitsorientierung, wie sie für eine standardisierte Auswertung erforderlich ist, gerät in Konflikt mit der grundsätzlichen Mehrdeutigkeit von Verhalten, insbesondere von sprachlichen Vorlagen einerseits und individuellen Lösungen andererseits (s. Kap. 1.4). Zudem fördert sie bei Schüler/inne/n die Haltung, nach der gewünschten Lösung zu suchen, statt dem eigenen Denken zu trauen.

Und wenn sie das Kreuz falsch gesetzt haben? Horst Bartnitzky hat mehrfach Aufgaben aus VerA-Deutsch für 3. Klassen analysiert und verdeutlicht, wie unterschiedlich die Gründe für (im Sinne der Tester) unzulängliche Lösungen sein können. An einem Beispiel illustriert (Bartnitzky 2008): Im Text werden den Kindern Informationen zu Schildkröten angeboten. Daraufhin sollen sie Fragen beantworten. Josias kreuzt alle Aussagen an – Nummer 2 und 3 im Sinne des Tests zu Unrecht:

Kreuze alle richtigen Antworten an:
- ☒ Schildkröten können sehr alt werden.
- ☒ Schildkröten werden genau 152 Jahre alt.
- ☒ Schildkröten sind mit 152 Jahren erwachsen.
- ☒ Schildkröten können über 150 Jahre alt werden.
- ☒ Schildkröten können älter werden als Menschen.

Nach der Logik der Auswertung zählen damit auch seine richtigen Kreuze 1, 4 und 5 nicht. Aber sind 2 und 3 überhaupt falsch? Die Autoren meinen den zweiten Satz im Sinne von »Alle Schildkröten werden höchstens 152 Jahre alt«. Und viele Schüler/innen werden ihn auch so verstehen; so etwas wird in Vortests ja geprüft. Weiß man aber, ob Josias ihn nicht gelesen hat als »Es gibt auch Schildkröten, die genau 152 Jahre alt werden« – wie die im Text beschriebene, von deren 152. Geburtstag die Rede ist und von der man im Übrigen nicht weiß, ob sie überhaupt älter geworden ist?

Den dritten Satz haben die Autoren vermutlich im Sinne von »Schildkröten sind erst/nur mit 152 Jahren erwachsen« gemeint. Josias könnte ihn aber gelesen haben im Sinne von »152 Jahre alte Schildkröten sind erwachsen«. Im Text findet sich keine Information, die gegen diese Aussage spricht.

Was Josias wirklich gedacht hat, wissen wir nicht – weder die Testautoren, noch Bartnitzky, noch Sie oder ich. Auch mag mancher die eine oder andere Alternativdeutung für spitzfindig halten. Aber sie sind möglich, zumal Tests plausible Falschlösungen anbieten müssen, sodass es gerade bei diesen sehr darauf ankommt, wie man sie liest. Und wie sich an weiteren Beispielen zeigt, nicht nur hier – und nicht nur in Deutsch.

Werden Hilfsmittel vorenthalten, deren Gebrauch die Schüler/innen nicht nur gewohnt sind, sondern deren Beherrschung auch explizit als Lernziel von den Lehrplänen eingefordert wird (z.B. Wörterbücher zur Kontrolle der Rechtschreibung), so entsteht eine Konkurrenz zu den Prinzipien des Unterrichts und zu den Erfahrungen der Schüler/innen. Auch das dicht konzentrierte Abarbeiten unzusammenhängender Aufgaben unter Zeitdruck (z.B. in Form von Diktaten) steht im Widerspruch zu Leistungssituationen, wie sie neuere Lehrpläne fordern und wie die Kinder sie in vielen Klassen gewohnt sind. Das ganze »Setting« der Testung wird von vielen Grundschulen, die nach den Richtlinien und Lehrplänen der letzten Jahre arbeiten, als Fremdkörper empfunden. Und diejenigen, die gerade die ersten Schritte machen, werden eher entmutigt – oder sogar in ihren alten Vorstellungen bestätigt. So wird die in Tests unterstellte Eindimensionalität der Fähigkeitsniveaus in Arithmetik, Geometrie und Sachrechnen/Größen in Mathematik nicht dem gerecht, was sich als Erkenntnis über »eigene Wege« des mathematischen und schriftsprachlichen Lernens in der Fachdidaktik zunehmend durchsetzt.

Diese kritischen Anmerkungen stellen nicht infrage, dass die Ergebnisse von Tests als wichtige Indikatoren für die angepeilten Leistungen dienen können. Aber die gewonnenen Daten müssen entsprechend interpretiert werden. Die zuständigen Institute der Bundesländer haben dazu Handreichungen veröffentlicht, die zum Teil sehr hilfreiche Hinweise geben. Zu befürchten ist auf der anderen Seite, dass viele Lehrer/innen und Eltern sich mit Summenwerten begnügen und ihre Urteile auf oberflächliche Vergleiche stützen werden. Punktwerte dürfen nicht ohne inhaltliche Prüfung zu Urteilen über Kinder oder Lehrer/innen werden. Damit sind wir bei der Durchführung von VerA und ihren Folgen.

Zu Besonderheiten der Erhebungssituation

Die Schulen gehen sehr unterschiedlich mit den Vorgaben um, sodass sowohl das Durchschnittsniveau als auch die Vergleichbarkeit der Ergebnisse einzelner Klassen ernsthaft infrage zu stellen sind. Zudem haben sich ambivalente bis problematische Auswirkungen auf den Unterricht gezeigt – bis hin zu Diskussionen, ob die Kinder in der vorgesehenen zehnminütigen Pause während der Deutscharbeit die Klasse verlassen und miteinander sprechen dürften oder ob die Pause schweigend im Klassenraum zu verbringen sei, unterbrochen allenfalls vom Gang zur Toilette.

Besonders kritische Punkte:

- In einer Reihe von Schulen wurden Aufgaben(typen) geübt, in einer Reihe von Klassen wurden sie sogar zum Üben vorher mit nach Hause gegeben, sodass die Ergebnisse schon deshalb über Klassen hinweg nicht vergleichbar sind.
- Manche Lehrer/innen haben ihren Klassen oder einzelnen Kindern entgegen der ausdrücklichen Anweisung geholfen, sodass die empirisch gewonnenen Daten nur schwer als Referenzdaten zu etablieren sind.
- Zudem ging man bei der Bewertung von Lösungen unterschiedlich »kulant« vor – zum Schutz einzelner Schüler/innen, aber auch im Interesse des eigenen Unterrichtserfolgs.
- Aufgabenformate im Unterricht orientieren sich zunehmend an Testformaten, obwohl Lernsituationen anderen Prinzipien gehorchen als Leistungskontrollen.
- Es zeichnet sich eine Einengung des Unterrichts auf die in den Tests geforderten Inhalte und Kompetenzen ab, z. B. auf technische »Informationsentnahme« aus Textstücken statt persönlicher Deutung und Auseinandersetzung mit der inhaltlichen Botschaft.
- Es ist absehbar, dass Eltern und die lokale Presse Informationen aus der verpflichtenden schulinternen Diskussion für ein informelles äußeres Ranking nutzen werden, ohne die Kontextbedingungen mit zu bedenken.
- Der Zeitpunkt der Erhebung und die Dreistufigkeit des Kompetenzmodells legen eine Zuordnung zu den Schularten der Sekundarstufe nahe und werden vor allem von Eltern oft so missverstanden.
- Im Vergleich zu den Parallelarbeiten, die NRW in früheren Jahren eingeführt hatte, schränkt der weniger enge Lehrplanbezug den Wert der Ergebnisse ein, etwa wenn bestimmte inhaltliche Konzepte oder Rechenverfahren in einigen Schulen schon vor dem Test eingeführt waren, in anderen aber noch nicht. Ebenso mindert die nicht mehr notwendige Zusammenarbeit im Kollegium, wie sie für die Absprachen bei den Parallelarbeiten noch erforderlich war, den Ertrag für den innerschulischen Entwicklungsprozess.

Zu den organisatorischen Rahmenbedingungen

Viele Schulen klagen über den Zeitdruck, die unzureichende Vorbereitung und den hohen personellen und finanziellen Aufwand bei der Durchführung. Als abträglich

für die alltägliche Arbeit unter ohnehin schon schwierigen Bedingungen werden insbesondere genannt:

- hoher Materialaufwand für das Kopieren bei begrenztem Etat (sofern die Materialien nicht zentral vervielfältigt wurden);
- hoher Zeitaufwand, um einsetzbare Testhefte zusammenzustellen, und erheblicher Korrekturaufwand – beides zieht Zeit von anderen Aktivitäten ab;
- sehr kurzfristige Anforderung der Ergebnisse aus den Referenzschulen und damit immenser Zeitdruck bei der Auswertung;
- die Fehleranfälligkeit bzw. unzulängliche Passung der komplizierten Auswertungsvorgaben und des Computersystems;
- die nicht immer abgestimmten Informationen aus verschiedenen Quellen und an verschiedene Zielgruppen, die zu Unsicherheit und zum Teil auch zu Argwohn geführt haben;
- die als nicht repräsentativ wahrgenommene Auswahl der Beispielaufgaben, die den Schulen zur Vorinformation zugeschickt wurden;
- der geringe Ertrag für die Weiterentwicklung des Unterrichts und die Förderung einzelner Kinder.

Die von der Politik erzeugte Hektik hat die Arbeit für alle Beteiligten erschwert und manchen guten Willen in den Schulen verschenkt. Insbesondere das Versprechen, es werde kein Ranking geben, wird durch die Verpflichtung, allen Eltern nicht nur die Ergebnisse ihres Kindes, sondern ebenso die Durchschnittswerte seiner Klasse und der Schule mitzuteilen, faktisch wertlos. Wer in Ländern wie England und den USA beobachtet hat, welche Konsequenzen die Publikation von globalen Testdaten einzelner Schulen z. B. auf die Immobilienpreise von Stadtteilen hat, wird sich keine Illusionen machen, was den verständigen Umgang mit solchen Daten in der Öffentlichkeit betrifft. Zu hoffen ist, dass aus diesen Schwierigkeiten für weitere Erhebungen gelernt wird.

Fazit

- Für ein regelmäßiges System-Monitoring, mit dem die Entwicklung des Schulsystems insgesamt erfasst werden soll, würde es reichen, alle vier bis sechs Jahre Erhebungen durchzuführen. Außerdem könnte man sich (wie bei PISA* und IGLU*) auf repräsentative Stichproben* beschränken und anderen Schulen die freiwillige Teilnahme ermöglichen. Das würde die Belastungen solcher Testprogramme mindern und dürfte ihre Akzeptanz beträchtlich erhöhen. Zugleich würden damit Mittel frei für andere Evaluationsaktivitäten. Gegen die Notwendigkeit flächendeckender Testprogramme spricht außerdem, dass es erfolgreiche PISA-Länder gibt, die ohne sie auskommen.

- Um Lehrkräften hilfreiche Informationen zur Kalibrierung* ihrer Maßstäbe und für den Vergleich mit anderen Klassen zu geben, wäre nicht nur ein Wechsel auf andere Kompetenzbereiche in Deutsch bzw. Mathematik, sondern auch auf andere Lernbereiche wie Sachunterricht und die musisch-ästhetischen Fächer sinnvoll. Dabei sollten auch weniger standardisierte Formate stärker erprobt werden. Schulen und einzelnen Lehrpersonen würde der freie Zugang zu einer Datenbank mit normierten Testaufgaben ermöglichen, gezielt Aufgabensets zusammenzustellen, um ihren Unterricht passgenau evaluieren zu können.
- Für die Individualbeobachtung müssten Instrumente zur Lernbegleitung entwickelt werden, um zu differenzierteren Einschätzungen, als sie durch eine punktuelle Messung möglich sind, zu gelangen (vgl. die Hilfen in Bartnitzky u. a. 2005 ff.). Diese Aufgabe hat in der nahen Zukunft Priorität, soll das Evaluationssystem nicht Schlagseite bekommen. Es ist zu hoffen, dass auch staatliche Institutionen zumindest einen Teil ihrer Ressourcen in solche Aktivitäten investieren.

Dass punktuelle Erhebungen mit standardisierten Instrumenten wie PISA & Co. inzwischen zum Paradigma der Evaluation von Leistungen geworden sind, stellt aus meiner Sicht das eigentliche Problem dar. Der grundsätzliche forschungsmethodische Fehler: Methoden, die für die Untersuchung von größeren Populationen* angemessen sind, eignen sich nur sehr eingeschränkt für die Evaluation von Einzelfällen wie konkreten Schulen, Lehrer/inne/n oder gar Schüler/inne/n. Tests können auch dann heuristisch hilfreich sein, ihre Ergebnisse haben aber immer den Status von Hypothesen, die mithilfe weiterer Daten überprüft werden müssen.

Auch hier kann die Medizin als Beispiel dienen. Ergebnisse aus der Erprobung von Medikamenten werden als Durchschnittsaussagen berichtet und führen zu allgemeinen Dosierungsanweisungen. Meist wird nur nach Kindern und Erwachsenen unterschieden – ohne Rücksicht auf Differenzen innerhalb dieser Untergruppen. Dabei kann dieselbe Arznei unterschiedlich wirken – bei Dicken und Dünnen, bei mehr oder weniger reizempfindlichen Personen, bei Frauen während oder außerhalb der Menstruation, je nachdem, wie sich jemand ernährt oder welche Medikamente sonst noch genommen werden. Es bedarf der langjährigen Praxiserfahrung eines Arztes und zusätzlich seiner Vertrautheit mit dem Patienten, um die Therapie optimal anzupassen. Eine solche »Einstellung« der Medikamentierung ist aber auch im medizinischen Praxisalltag nicht selbstverständlich, sofern es nicht um so heikle Krankheiten wie Parkinson und ihre Behandlung geht.

Diese Kritik trifft auch die Verfahren der Schulinspektion. Immerhin weiten sie den Blick über den Output auf Prozesse und Bedingungen, sodass ihre Rückmeldungen differenzierter und damit für die Schulen hilfreicher sind. Aber auch dabei wird auf eine Standardisierung der Datenerhebung und -auswertung gesetzt. Denn im Gefolge von PISA & Co. haben nicht nur Tests, sondern auch andere standardisierte Verfahren ein hohes Prestige gewonnen, z. B. Kriterienraster zur Beobachtung von Unterricht. Sie gelten als objektiv, ihre Daten als sicher. Demgegenüber haben die Urteile von Perso-

nen an Kredit verloren. Dazu zählt eben auch die Beurteilung von Lehrer/inne/n durch Beamte der Schulaufsicht.

Diese Präferenz für standardisierte gegenüber ausdrücklich interpretativen Verfahren fokussiert ein Kernproblem der Evaluation: In der Sozialforschung unterscheidet man hoch- und niedrig-inferente, das heißt mehr oder weniger interpretationsoffene Methoden bzw. Instrumente der Untersuchung. Viele der Prozess-Standards, die im Schulverbund »Blick über den Zaun« formuliert wurden (vgl. Groeben u.a. 2005), sind hoch-inferent. Standards wie »Die Lehrenden bedienen individuell verschiedene Lernmöglichkeiten und -wege« oder »Für Lernprobleme finden sie [die Schüler/innen] geeignet Hilfen« sind für die Beobachter in besonderem Maße interpretationsbedürftig, denn sie erfordern Einschätzungen der individuellen Lernmöglichkeiten und der Eignung von Lernhilfen, die unter Pädagogen und Psychologen Glaubenskriege auslösen können.

Dagegen gelten Standards wie »Die Schülerinnen und Schüler werden täglich begrüßt und verabschiedet, wenn sie in die Schule kommen bzw. die Schule verlassen« oder »Jede Gruppe wird von zwei Erwachsenen betreut« als niedrig-inferente Kriterien, das heißt als interpretationsarm, weil sie sich an Oberflächenmerkmalen festmachen lassen, bei denen verschiedene Beobachter in der Regel zu gleichen Klassifikationen kommen. Methodisch-technisch kontrollierbar sind die Datenerhebungen also nur im zweiten Fall – und damit objektiv interpretierbar, behaupten zumindest die Verfechter des Standardisierungsparadigmas. Aber sind sie das wirklich?

Ob die Schüler/innen einzeln begrüßt werden, lässt sich leicht feststellen. Aber pädagogisch gesehen ist entscheidend, *wie* die Schüler/innen begrüßt werden. Denn in der Formulierung der Standards wird ein Oberflächenverhalten als Indikator für eine Haltung genommen, die es nicht notwendig repräsentiert. Äußerungen wie »Guten Morgen, Herr Lehrer« im Chor oder der Handschlag des Lehrers an der Tür mit Prüfung der Schmutzränder und der Abkaulänge der Fingernägel des Schülers fallen in die Kategorie »tägliche Begrüßung«, bedeuten aber etwas anderes, als in den Standards gemeint ist. Daraus folgt: Auch mit standardisierten Instrumenten erhobene Daten sind hoch-inferent. Das Prädikat »evidenzbasiert« ist deshalb auf halb- oder nicht-standardisierte Formen systematischer Datenerhebung zu erweitern, wie dies z.B. Helmke u.a. (2014) zu Recht für ihre Instrumente zur Unterrichtsbeobachtung (EMU) beanspruchen.

Hoch-inferent müssen Evaluationsverfahren in der Pädagogik grundsätzlich sein, wenn sie dem Gegenstand gerecht werden wollen. »Eine Schule ist kein Auto«, »Schulen sind Kunstwerke«, sagt Seydel (2005, S. 11; 2009), der sie aufgrund seiner Evaluationserfahrung als jeweils einmalig, komplex, flüchtig und nicht objektiv bewertbar einschätzt. Insofern ist Pädagogik zumindest *auch* Kunst (→ »Die fünf Welten des Curriculum« 1980). Wie kann man dann ihre Qualität angemessen mit standardisierten Verfahren untersuchen und beurteilen? Auch wenn Kunst nicht als Beliebigkeit missverstanden werden darf, verlangt sie doch handwerkliches Geschick und technische Genauigkeit.

5.3 Schulevaluation durch »kritische Freunde«

Wer Schule und Unterricht verbessern will, hat zwei Optionen: Er kann auf der zentralen Ebene ansetzen und versuchen, die Rahmenbedingungen zu verändern: andere Schulstruktur, neue Lehrpläne, mehr Ressourcen. Wenn man die Klagen von Lehrkräften hört »Wir können nicht, weil ...«, kann man diesem Weg viel abgewinnen. Andererseits fällt auf, wie unterschiedlich der Alltag innerhalb derselben Strukturen aussieht. Entsprechend lauten die Klagen von Politik und Verwaltung: »Aber die Lehrer/innen ...«. Diesen Schuldzuweisungen und schärferen Kontrollsystemen werden die Schulen nur entgehen können, wenn sie selbst wirksame Formen der Evaluation und Rechenschaft entwickeln. Dafür gibt es überzeugende Beispiele (→ »Scharfe Brillen, wache Augen und ein einfühlsamer Blick« 2007).

Unterrichts- und Schulentwicklung beginnen mit Selbstevaluation, mit dem täglichen Nachdenken über die Qualität der eigenen Arbeit, angeregt durch informelle Beobachtungen, was mehr oder besser gelungen ist, systematisiert über gezielte Erhebungen, z. B. mit Tests oder mit Fragebögen, die inzwischen in erprobten Formen vorliegen, und mit wechselseitigen Hospitationen. Interne Evaluation hat den Vorteil, dass man die Situation vor Ort besser kennt und sich nicht einer Bedrohung von außen ausgesetzt sieht. Andererseits ist der Blick von außen wichtig, um scheinbare Selbstverständlichkeiten infrage zu stellen. Zudem braucht jeder Distanz zum Alltag, um sich den erkannten Schwächen zu stellen. Diese Einsichten stehen hinter den Lernstandserhebungen und der Einrichtung von Schulinspektionen. Sie machen die Forderung nach externer Evaluation stark. Aber die Frage ist, in welcher Rolle die Externen kommen: als Experten, als Autoritäten oder als Partner? »Von außen« heißt nicht zwingend »von oben«.

Externe Evaluation ist notwendig. Für die Schulentwicklung wird sie aber produktiver, wenn sie von Kontrolle abgekoppelt ist. Die Schulaufsicht sollte deshalb auf fachlicher Ebene durch ein kollegiales Peer-Review ergänzt werden, das den Unterrichtsbetrieb nicht rechtlich kontrolliert, sondern inhaltlich berät. Dies gelingt nur selten mit denselben Personen, wie auch die Beratung von Lehramtsanwärtern durch die sie bewertenden Fachleiter zeigt. Die Schule kann den Fremdblick durch eine interne Bestandsaufnahme vorbereiten: Was sind unsere Ziele, wo liegen unsere Stärken, welche Probleme haben wir? Und sie sollte die Berichte der externen Beobachter anschließend öffentlich kommentieren: Welche Einschätzungen teilen wir, was wollen wir unternehmen, um unsere Arbeit weiter zu verbessern? Auch ohne formelle Sanktionsbefugnisse kann eine solche Inspektion durch ihre fachliche Kompetenz und durch den sozialen Druck wirken, den die Berichterstattung bedeutet.

Den Versuch, externe Sicht und notwendige Vertrautheit auszutarieren, unternimmt die Initiative »Blick über den Zaun«. Dieser Verbund von inzwischen rund 140 reformpädagogisch orientierten Schulen besteht seit 1989. In ihm schließen sich jeweils sechs bis zehn Schulen (bewusst aus verschiedenen Schulformen, Regionen und Traditionen) zu einem Arbeitskreis zusammen. Gemeinsam sind den Schulen Stan-

dards, die sich vor allem auf die Qualität der pädagogischen Prozesse beziehen. Zweimal im Jahr wird eine Schule von jeweils zwei Vertretern der anderen Schulen des gleichen Arbeitskreises, also insgesamt etwa 15 Personen besucht, die zwei bis drei Tage in der Schule verbringen (vgl. zur Gestaltung solcher Besuche Backhaus u.a. 2009). Die gastgebende Schule kann einen spezifischen Beobachtungsauftrag formulieren, doch unabhängig davon sind die Gäste frei, das zu beobachten, was ihnen wichtig erscheint. Sie nehmen am Unterricht teil, unterhalten sich mit Kollegen, Schülern und Elternvertretern. In einer Schlussrunde spiegeln die Besucher dem Kollegium jeweils ihre individuellen und damit unterschiedlichen Eindrücke zurück. Die Erträge dieses kollegialen Austauschs sind vielfältig – sowohl für die besuchte Schule als auch für die Besucher, die ihre Beobachtungen als Anregung in die eigene Schule zurückbringen (vgl. Seydel 2007, S. 5 ff.).

Wenn man den Unterricht einer Klasse auf diese Weise in den Blick nimmt, können für die Gespräche mit den Kollegen Fragen wie die folgenden hilfreich sein:

- »Was sind *Ihre* wichtigsten Ziele und Prinzipien?« Nachfrage: »Wie stehen Sie zu folgenden Vorgaben der Richtlinien/Lehrpläne, zu folgenden Standards oder Positionen der schulpädagogischen und (fach-)didaktischen Diskussion?«
- »Wo steht Ihre Lerngruppe, wo stehen einzelne Kinder in den zentralen Entwicklungsdimensionen?« Nachfrage: »Ist Ihnen aufgefallen, dass Marc ...?«
- »Wo sehen Sie die Stärken und die Schwächen Ihrer Arbeit, also Ihrer Versuche, die eigenen Ansprüche und die vorgegebenen Anforderungen umzusetzen?« Nachfrage: »Mir ist bei der Beobachtung Ihres Unterrichts aufgefallen, dass ...«
- »Welche Umstände erschweren es Ihnen, Ihre Ansprüche im Alltag umzusetzen?« Nachfrage: »Könnte es auch daran liegen, dass ...?«
- »Welche nächsten Schritte haben Sie sich zur Entwicklung Ihrer Arbeit vorgenommen?« Nachfrage: »Haben Sie auch an folgende Möglichkeiten gedacht ...?«
- »Welche Unterstützung und welche Rahmenbedingungen wären nötig?« Nachfrage: »Würde es Ihnen helfen, wenn ...?«

Beratung statt Kontrolle bedeutet, mit einer anderen Sicht konfrontiert zu werden, ohne dass sich diese als Autorität versteht. Die Grundidee: Die normative Diskussion über die Kriterien für »guten Unterricht« ist zu trennen von der empirischen Frage nach der Qualität seiner tatsächlichen Umsetzung. Es macht wenig Sinn, den tatsächlichen Unterricht mit Ansprüchen zu bewerten, die die Bewerteten gar nicht teilen. Das heißt nicht, die selbst gesetzten Ziele einer Schule oder Lehrperson unhinterfragt zu übernehmen. Aber normative Differenzen sind *separat* zu klären – beispielsweise in einer Diskussion über das Schulprogramm, seine Stärken und Schwächen. Dabei kann zugleich eine Verständigung über die Kriterien erzielt werden, mit deren Hilfe der beobachtete Unterricht sinnvoll zu beurteilen ist. Testleistungen der Schüler/innen können eine Informationsquelle sein, um Schwächen auf die Spur zu kommen. Aber noch ertragreicher sind die Beobachtungen und das in langjähriger Erfahrung fundierte Urteil der Kolleginnen und Kollegen aus anderen Schulen.

Auch dieses Verfahren hat seine Schwierigkeiten – und das Instrumentarium ist entwicklungsfähig. Entscheidend ist der Ansatz: Evaluation von außen, nicht von oben, seien es die wissenschaftlichen Experten der Lernstandserhebungen, seien es die Vertreter der Verwaltung bei der Schulinspektion. Begegnung auf Augenhöhe ist die Grundlage für Offenheit und damit für die Bereitschaft, sich den eigenen Schwächen zu stellen und ernsthaft an ihnen zu arbeiten. Dieser Anspruch schließt ein, Schüler/innen nicht nur als Informationsquelle, sondern als Experten für Kriterien guter Schule und ihre Umsetzung mit in den Prozess der Evaluation einzubeziehen. Dafür gibt es bereits positive Beispiele (s. dazu → »Was ist eine gute Schule? Unsere Standards für Kinder« 2010, und Brügelmann/Backhaus 2011).

Evaluation dialogisch auszugestalten ist auch Anspruch an Formen der Leistungsbewertung bei Schülern, wie ich in Abschnitt 6.3 zeigen werde.

6. Die Not mit den Noten

Die Wahl geeigneter Beurteilungsformen wirft sehr unterschiedliche Probleme auf. Drei Entscheidungsebenen mit je besonderen Problemen sind zu unterscheiden:

- die Wahl der Verfahren zur *Feststellung* des Lernerfolgs: informelle vs. standardisierte Aufgaben, offene vs. strukturierte Beobachtung;
- die Wahl der Bezugsnorm zur *Bewertung* des Lernerfolgs: nach Annäherung an das Lernziel oder individuellem Lernfortschritt oder relativer Leistungsposition in einer Gruppe;
- die Wahl der *Darstellungsform* in der Rückmeldung: Beschreibung vs. Bewertung; freie Formulierung vs. Ziffern.

Die genannten Alternativen werden im Schul- und Berufsalltag in verschiedenen Kombinationen realisiert. Dadurch wird eine sachliche Diskussion über den Sinn von Noten erschwert.

6.1 Differenzierung der Notenkritik

Noten werden unter den Betroffenen, aber auch in Fachkreisen immer wieder kontrovers diskutiert. Ihr Nutzen war und ist heftig umstritten. Dabei liegen seit 50 Jahren zentrale empirische Befunde zur Problematik von Noten vor, zum Teil noch länger. Seit den 70er Jahren sind diese Studien im deutschen Sprachraum in systematisierenden Überblicken publiziert worden (so schon bei Ingenkamp 1971). Ihre Kritik ist in der Zwischenzeit durch weitere Studien bestätigt und erneut mehrfach zusammengefasst worden (Arbeitsgruppe Primarstufe 2006, Kap. 1; Faber/Billmann-Mahecha 2010; Maaz u.a. 2011). Dennoch hat sich in der Wahrnehmung von Schüler/inne/n, Eltern und Lehrer/inne/n, aber auch in der Öffentlichkeit und erst recht im Schulalltag nur wenig verändert. Soweit es in den letzten Jahren Veränderungen gab, waren sogar teilweise – vor allem im Grundschulbereich – wieder Tendenzen zu beobachten, das Rad der Entwicklung zurückzudrehen, ja, die Noten über den Leistungsbereich hinaus auszuweiten auf das Sozialverhalten. Andernorts dagegen, z. B. in Schleswig-Holstein, wurden Ziffernnoten in der Grundschule zur begründungspflichtigen Ausnahme erklärt. Vor diesem Hintergrund lohnt es, die empirische Forschung zu Noten und alternativen Formen der Leistungsbeurteilung zu sichten und im Kontext der aktuellen Diskussion zu bewerten. Im Folgenden werden vor allem die für unser Thema relevanten Befunde zusammenfassend dargestellt (→ »Sind Noten nützlich – und nötig?« 2007).

Die Fundiertheit von Beurteilungen hängt zunächst einmal von ihrer Datengrundlage ab. Forschungsmethodisch wird deren Qualität in der Tradition quantitativ-standardisierter Forschung über drei Gütekriterien bestimmt:

- *Validität:* Erfasst das Instrument tatsächlich das, was es zu messen vorgibt? Leistungen sind beobachtbare Verhaltensweisen. Ihre Beurteilung zielt aber nicht nur auf das sichtbare Verhalten (Performanz), sondern auch auf die zugrunde liegenden Fähigkeiten (Kompetenz). Ob diese treffend erfasst werden, ist schwer einzuschätzen, da die Außenkriterien selbst umstritten sind.
- *Objektivität:* Sind die Ergebnisse der Messung unabhängig von der Person, die sie durchführt und auswertet?
- *Reliabilität:* Sind die Ergebnisse verlässlich, also situationsunabhängig, das heißt verändern sie sich nicht unter anderen Bedingungen, z.B. bei Wiederholung einer Prüfung?

Die Antworten auf diese Fragen fallen durchweg negativ aus. Denn wie die bereits erwähnten Studien zeigen, sind Noten entgegen den ihnen unterstellten Vorzügen

- nicht valide: In eine Beurteilung gehen – oft ungewollt – immer fachfremde Merkmale mit ein, z.B. Schrift und Rechtschreibung bei der Beurteilung von Aufsätzen (Arbeitsgruppe Primarstufe 2006, Kap. 1.1);
- nicht objektiv: Dieselbe Leistung wird von verschiedenen Lehrpersonen sehr unterschiedlich beurteilt, z.B. weil sie unterschiedliche Schwerpunkte setzen oder eigene Maßstäbe anlegen (ebd., Kap. 1.2);
- nicht verlässlich: Dieselbe Leistung wird von denselben Lehrer/inne/n zu verschiedenen Zeitpunkten unterschiedlich bewertet (ebd., Kap. 1.3);
- nicht vergleichbar: Für dieselbe Leistung werden in verschiedenen Klassen je nach Niveau unterschiedliche Noten vergeben (ebd., Kap. 2);
- nicht informativ: Differenzierte Leistungsprofile schrumpfen auf eine pauschale Ziffer, was z.B. eine gezielte Förderung erschwert (ebd., Kap. 3.1);
- nicht dauerhaft motivierend: Die Leistungen in Klassen oder Schulsystemen, die auf Noten verzichten, sind nicht schlechter als dort, wo benotet wird (ebd., Kap. 2.3, 3.2.3 und 3.2.4).

Die oft beschworene Erwartung, durch Noten wisse man besser als durch verbale Urteile, wo genau ein Kind stehe, erweist sich also als eine Chimäre. Obwohl die Kultusministerkonferenz schon vor fast 50 Jahren eine lernzielbezogene Beurteilung von Leistungen gefordert hat (KMK 1968), orientieren sich Noten üblicherweise an der sozialen Bezugsnorm – mit der jeweiligen Schulklasse als dominierendem Maßstab. Je nachdem, wie leistungsstark die Klasse ist, fällt die Bewertung derselben Leistung also unterschiedlich aus.

Erst recht gilt das für die Prognose der weiteren Schullaufbahn. Leistungen – als beobachtbare Verhaltensweisen – werden zwar rückblickend beurteilt. Die Beurteilung beansprucht aber, zugrunde liegende Fähigkeiten zu erfassen und damit auch Aufschluss zu geben über künftig zu erwartende Leistungen. Wie weit solche Vorhersagen eintreffen, ist dabei das einzige Außenkriterium, um die Aussagekraft von Noten einzuschätzen. Diese hat sich aber als sehr begrenzt erwiesen. So wurde die Vorher-

sagekraft von Beurteilungen, insbesondere von Noten, für verschiedene Phasen der Bildungslaufbahn untersucht. Über die Gesamtgruppe hinweg gibt es Korrelationen* mit späteren Leistungen, für einzelne Schüler/innen aber lässt sich nur mit hohem Fehlerrisiko vorhersagen, in welcher Leistungsgruppe sie später landen werden (Arbeitsgruppe Primarstufe 2006, Kap. 1.1.3). Eine besondere Bedeutung haben solche individuellen Klassifikationen an den Übergängen, z. B. von der Grundschule in die weiterführenden Schulen oder aus der Schule ins Berufsleben.

Im Blick auf die Alternative einer stärkeren Standardisierung verdient vor allem das Verhältnis von Noten und Tests eine genauere Betrachtung. Die Validität von Noten wird in den letzten Jahren vor allem mit dem Hinweis kritisiert, dass Noten nicht gut mit den Ergebnissen von Leistungstests in den entsprechenden Fächern übereinstimmten. Andererseits wird die inhaltliche Gültigkeit von Tests in der Regel damit begründet, dass ihre Ergebnisse in der Normierungsphase »gut« mit den Lehrerurteilen übereinstimmen. Damit entstehen Kreisschlüsse. So stimmen nach Ditton (2010, S. 69) zumindest für die Grundschule Noten und Testergebnisse vergleichsweise gut überein. Er schließt daraus, dass das Lehrerurteil so schlecht doch nicht sein kann, wie dem häufigen Vorwurf einer »diagnostischen Inkompetenz« von Lehrer/inne/n zu entnehmen sei. Aber auch für seine positive Einschätzung gilt: Da Tests in der Regel durch den Vergleich mit den Fachnoten der Schüler/innen geeicht werden, verwundert es nicht, dass hinten herauskommt, was vorne hineingesteckt wurde.

Soweit es Differenzen gibt, wird aber eher den Lehrer/inne/n der schwarze Peter zugeschoben. Problematisch an der Diskussion nach PISA ist, dass Tests dabei fast selbstverständlich als Maßstab für die »wahre« Leistung von Schüler/inne/n gesetzt werden. Vergleiche mit weiteren Kriterien zeigen aber, dass Lehrerurteil und Tests unterschiedliche Aspekte fachlicher Leistungen erfassen. Es macht deshalb keinen Sinn, die Qualität des einen Verfahrens allein durch den Grad der Übereinstimmung mit den Ergebnissen des anderen zu bestimmen.

Aber selbst wenn man die Testergebnisse als Maßstab nimmt, sind für die beobachteten Abweichungen immer noch zwei verschiedene Erklärungen denkbar: Lehrer/innen erkennen schlechter als Tests, wo Kinder in ihrer Entwicklung stehen, welche Schwierigkeiten sie bei der Auseinandersetzung mit dem jeweiligen Gegenstand haben (fehlende diagnostische Kompetenz). Oder: Lehrer/innen verschiedener Klassen ordnen den richtig erkannten Leistungsstand unterschiedlichen Notenstufen zu (abweichende Bewertungsmaßstäbe).

Die vorliegenden Studien sprechen eher für die zweite Sicht. So korrelieren Noten und Testwerte innerhalb von Klassen sehr viel höher miteinander als über verschiedene Klassen hinweg. Lehrer/innen differenzieren unterschiedliche Lernstände also auch im Sinne von Leistungstests einigermaßen zutreffend, aber sie setzen anschließend den Bezugspunkt für die Bewertung durch Noten unterschiedlich an. Bestätigt wird diese Deutung durch noch höhere Korrelationen*, wenn man von den Lehrkräften qualitative Urteile über die voraussichtliche Entwicklung ihrer Schüler/innen erfragt und diese mit Tests zur kognitiven Leistungsfähigkeit abgleicht.

Abschließend ist festzuhalten: Ziffernnoten erfüllen weder die ihnen oft zuerkannten Qualitäten wie Objektivität und Vergleichbarkeit noch pädagogische Ansprüche wie Förderung von Selbstständigkeit oder dauerhafte Motivation.

6.2 Tests oder Texte als Alternative zu Noten

Es bieten sich zwei Auswege aus den Problemen an, die Noten aufwerfen: eine methodisch-technische Perfektionierung der Erhebung und Auswertung von Leistungsdaten, z. B. durch standardisierte Tests, oder eine Offenlegung der Subjektivität von Wahrnehmungen und Urteilen, z. B. in Verbalgutachten, und ihre stärkere soziale Kontrolle, z. B. durch die Kombination mehrerer Perspektiven. Bei der Entscheidung zwischen diesen Optionen ist zu beachten, dass Beschreibungen und Bewertungen für die Betroffenen nicht nur kognitiv nachvollziehbar, sondern auch sozial annehmbar sein müssen. Damit werden Standards wie Glaubwürdigkeit, Fairness, Nützlichkeit und Verständlichkeit bedeutsam, die in der Diskussion über Leistungsbeurteilung noch zu wenig Beachtung finden.

Nutzen und Grenzen von standardisierten Tests

Die folgenden vier Vorteile zeichnen Tests als Instrument aus:
- die Fokussierung der Datenerhebung auf vorweg begründete Leistungsaspekte;
- damit: die Transparenz der Anforderungen und der Bewertung;
- die (klassenübergreifende) Kalibrierung* der Maßstäbe durch Bezug auf Normstichproben;
- das Aufdecken blinder Flecken in Unterricht und Beurteilung (in einzelnen Fällen auch als Modell für Lern- oder Prüfformate, vgl. die Ablösung des lauten Vorlesens durch das Beantworten von inhaltlichen Fragen zum Text).

Insofern sind standardisierte Tests – stärker als früher üblich – in das Repertoire schulischer Leistungsbeurteilung einzubeziehen. Allerdings ist ihr Status zu relativieren. Sinnvoll sind sie eher als heuristisches Instrument (»Thermometer«, »Warnlampe«). Denn eine Stärke wie die klare Fokussierung auf bestimmte inhaltliche Anforderungen bedeutet im Umkehrschluss, dass schon aus ökonomischen Gründen immer nur kleine Ausschnitte aus dem Spektrum einer fachlichen Kompetenz erfasst werden können. Die Diskussion über die Bildungsstandards zeigt aber ein hohes Maß an Uneinigkeit darüber, was als Mindest- oder Regelleistung eingefordert werden kann. Die Kritik an den Aufgaben der landesweiten Lernstandserhebungen und der internationalen Leistungsvergleiche hat offengelegt, wie umstritten die Annahmen zu den angeblich erfassten Fähigkeiten sind. Daraus erklärt sich, dass Schüler/innen in verschiedenen Mathematik- oder Lesetests unterschiedliche Ergebnisse erzielen. Auch so folgen-

schwere Entscheidungen wie die Einstufung von Kindern als »Legastheniker« führen zu unterschiedlichen Ergebnissen, je nachdem, mit welchem Intelligenztest einerseits und welchem Lese- oder Rechtschreibtest andererseits die Diskrepanz zwischen beiden Leistungen bestimmt wurde. Zudem wird dieselbe Fähigkeit (z. B. Lesekompetenz) über unterschiedliche Aufgaben operationalisiert. Entsprechend unterschiedlich fällt die Einschätzung der Lesefähigkeit deutscher Schüler/innen etwa bei PISA* und DESI* aus.

Besonders problematisch ist die gängige Verfahrensweise, die Validität eines neuen Tests durch die Korrelation* seiner Ergebnisse mit anderen Tests zu belegen – oft älteren, die er eigentlich ersetzen soll. Wie Rackwitz (2010) für Lese- und Rechtschreibtests gezeigt hat, entstehen so regelrecht »Validierungsketten«, in denen sich jeder Test auf seine (wenn auch meist begrenzte) Korrelation* mit einem anderen Test beruft – bis dahin, dass sich Tests wechselseitig als Beleg für ihre inhaltliche Aussagekraft heranziehen. Dass Tests auch durch ihre Korrelation* mit den angeblich so unzuverlässigen Zeugnisnoten »validiert« werden, wurde bereits erwähnt.

Ein weiteres Problem: Aus der Entscheidung für eine Standardisierung folgt, dass nur bestimmte Aufgabentypen genutzt werden können – vorzugsweise Multiple-Choice-Aufgaben. Und selbst bei offenen Fragen muss die Klassifikation und Bewertung von Lösungen vorweg festgelegt werden. Menschliches Verhalten ist aber – wie schon mehrfach betont – kontextabhängig und mehrdeutig. Das gilt auch für das Lösen von Testaufgaben und ihre Auswertung. Ohne kognitive und emotionale Empathie können Leistungen weder erklärt noch angemessen gewürdigt werden. Wie Prüflinge eine Aufgabe gedeutet und wie sie ihre Antworten gemeint haben, ist durch die bewusste Ausblendung persönlicher Interaktionen nicht mehr verhandelbar. Damit wird nicht Objektivität gesichert, sondern die Subjektivität der Test-Entwickler und -Auswerter über die der beurteilten Personen gestellt. Standardisierung erzeugt somit zwei neue Probleme: In Tests und in Fragebögen ist die konkrete Bedeutung von Fragen und Antworten zwischen den Beteiligten nicht mehr verhandelbar, und die Kategorien müssen notgedrungen sehr allgemein gewählt werden.

So mag es für Politiker/innen interessant sein, dass Buchbesitz in der Familie stärker mit der Leseleistung von Schüler/inne/n korreliert als andere Merkmale. Lassen wir beiseite, dass »Buchbesitz« vielleicht nur ein Indikator für das umfassendere kulturelle Kapital oder dass das damit korrelierende »häufige Vorlesen« die eigentliche Ursache ist. Für den Einsatz öffentlicher Mittel kann diese Korrelation* ein wichtiges Argument sein, z. B. in einem Programm, um Eltern bei den Vorsorgeuntersuchungen ihrer Kinder jeweils ein Bilder- oder Kinderbuch zu schenken. Bei der Erklärung von individuellen Lernschwierigkeiten kann das Merkmal »Buchbesitz« dagegen völlig in die Irre führen. Hier gewinnen Faktoren an Bedeutung, die man mit kurzen Fragen in standardisierter Form nicht erfassen kann, z. B. wie mit diesen Büchern in der Familie umgegangen wird.

Ebenso kann die Bildung von Untergruppen mit besonderem Risiko, etwa Jungen, Migranten- oder Unterschichtkinder, durchaus helfen, den bildungspolitischen Blick

zu differenzieren. Aber auf der Schul- und Unterrichtsebene verschließen solche Kategorien den Blick auf den Einzelfall und seine Besonderheit: das Mädchen mit den Leseproblemen, das unmotivierte Oberschichtkind, den gut lesenden Migranten. Es kommt nicht nur zu normativen Etikettierungen, sondern auch zu so irritierenden Vorschlägen wie der Forderung, die gerade mühsam errungene Koedukation von Mädchen und Jungen aufzuheben. Irritierend, denn die Streuung der Leistungen, der Interessen und der Lernstile innerhalb solcher Gruppen ist um ein Vielfaches größer als die Differenzen zwischen ihnen.

Mit PISA eingeführte Begriffe wie »Migrationshintergrund« erfassen den relevanten Sachverhalt zwar besser als »Staatsangehörigkeit«. Doch in standardisierter Form führt auch diese Klassifikation in die Irre. So kann man den – im Vergleich zu Kanada – niedrigeren Bildungserfolg von Immigranten in Deutschland nicht unbesehen den Schulen anlasten. Nach Deutschland migrierte Familien unterscheiden sich beispielsweise in ihrer sozialen Schichtzugehörigkeit deutlich von kanadischen Einwanderern, sodass in Deutschland dieser Faktor zusätzlich ins Spiel kommt: Nicht der Migrationshintergrund oder die Zweisprachigkeit an sich sind das Problem, sondern die Lebensbedingungen, die familiäre Unterstützung und andere durch das sozioökonomische Milieu bedingte Einflüsse. Die Differenzierung lässt sich fortsetzen. So gibt es unter den deutschen Immigranten je nach nationaler Herkunft große Unterschiede. Migranten-Jungen wiederum haben – nicht anders als unter den »eingeborenen« Deutschen – mehr Probleme als Mädchen. Schließlich landet man beim Individuum.

Ein Problem bei Einzelfalldiagnosen ist zudem die breite Schwankung einer punktuell erfassten Testleistung um den »wahren Wert« der eigentlich angezielten Fähigkeit (= hoher Messfehler bei Individualdaten). Bei Aussagen über größere Gruppen, wie sie für bildungspolitische Entscheidungen genutzt werden, stellt sich dieses Problem in geringerem Umfang, weil sich individuelle Schwankungen in den Kennwerten für die Stichprobe* insgesamt ausgleichen. Insofern liefern Studien wie PISA*, IGLU* und VerA* einigermaßen aussagekräftige Daten für eine schulübergreifende Systemevaluation. Ihre Ergebnisse sind aber nicht zureichend verlässlich für die Erfassung und Bewertung individueller Leistungen von Schüler/inne/n (oder auch Lehrer/inne/n). Insbesondere wenn es um die Vorhersage individueller Entwicklungen geht, ist nicht nur das Lehrerurteil, sondern sind auch Tests in hohem Maße unsicher (→ »Das Prognoserisiko von Risikoprognosen« 2005).

Diese Hinweise unterstreichen das bereits erläuterte grundsätzliche Dilemma des Forschungsparadigmas, dem PISA & Co. verpflichtet sind: Die Kategorien, mit denen sich Daten einer derart großen Stichprobe statistisch verarbeiten lassen, sind zu grob, um die Vielfalt relevanter Bedingungen zu erfassen. Dies schränkt selbst die Möglichkeit bildungspolitischer Folgerungen ein: Für welche (Teil-)Gruppen genau ist ein Förderprogramm zu entwickeln, wenn man die knappen Ressourcen konzentrieren muss?

Wenig ergiebig sind die Befunde aber besonders für diejenigen, die vor Ort mit einzelnen Schülern zu tun haben. Sie helfen nicht beim Umgang mit einem muslimischen

Mädchen aus einer türkischen Mittelschichtfamilie. Solche Schubladen mögen nützlich sein, um die Fülle der erhobenen Variablen etwas zu ordnen. Auf der Unterrichtsebene verstellen sie die Wahrnehmung für die je besondere Lernbiografie einzelner Schüler. Das Sprachproblem von Migranten ist eben nicht nur ein sprachtechnisches, das sich durch entsprechende Förderkurse ausgleichen ließe, sondern ein sprachkulturelles, wie es auch deutsche Unterschichtkinder in die Schule mitbringen. Dieses aber ist nur mit genauen qualitativen Beobachtungen zu erfassen. Der in den letzten 20 bis 30 Jahren in der pädagogischen Praxis mühsam errungene Blick auf das Individuum droht im Nach-PISA-Diskurs, der von Grobkategorien geprägt ist, wieder verloren zu gehen – übrigens entgegen der erklärten Absicht der Autoren: eine der ungewollten, aber absehbaren Nebenwirkungen. Von PISA & Co. profitiert vielleicht die Politik, nicht aber die pädagogische Praxis.

In der Pädagogik muss man deshalb in einem ähnlichen Sinn »experimentell« vorgehen wie in der Medizin: mit einer erfahrungsoffenen Haltung, nicht als bloß technische Anwendung von Erkenntnissen. Als Grobfilter (»Screening«) können Tests hilfreich sein, wenn sie eine unkomplizierte Durchführung und Auswertung erlauben. Ihre Ergebnisse dienen dann als Ausgangshypothese, um für einzelne Kinder(gruppen) geeignete Aufgaben zu suchen und sie mit ihnen auszuprobieren. Insofern werden diese – wie ein verordnetes Medikament – selbst zum Diagnoseinstrument. Erst die Reaktion der Schülerin oder des Schülers zeigt, ob die Aufgabe passt oder nicht. Im gemeinsam reflektierten Probieren wird versucht, die Passung zu verbessern. Tests haben also eine heuristische Funktion – nicht mehr, aber auch nicht weniger. Insofern ist das anschließende Gespräch mit den Schüler/inne/n wichtig, vor allem bei Fehlern: Wie hast du das gerechnet? Warum hast du das so geschrieben? Kannst du deine Lösung erläutern? Das Motto lautet: Dialog (auch der Schüler/innen untereinander) statt Urteil von oben.

Fazit: Tests sollten genutzt werden, um Lernbeobachtungen und Lehrerurteile zu ergänzen. Sie können diese aber nicht ersetzen, denn

- sie müssen sich auf bestimmte Aufgabentypen und Auswertungsformen beschränken (»Standardisierung«);
- sie konzentrieren sich inhaltlich auf wenige Ausschnitte der Leistungsprofile (Messbarkeit; Ökonomie);
- ihre Ergebnisse sind bei punktueller Erhebung nicht verlässlich, das heißt auf Individualebene ist der Messfehler, der sich in großen Stichproben ausgleicht, zu groß, um biografische Entscheidungen abzusichern;
- ihre prognostische Validität ist ähnlich unsicher wie die der Noten.

Schwächen und Stärken von Verbalbeurteilungen

Texte als Alternative zu Tests haben ähnliche Schwächen wie Noten, sind sie doch genauso abhängig vom persönlichen Urteil der Lehrer/innen. Insofern gelten auch einige der oben zu Noten genannten Einschränkungen wie:

- beschränkte Validität
- fehlende Objektivität
- mangelnde Verlässlichkeit
- eingeschränkte Vergleichbarkeit.

Immerhin macht die sprachliche Fassung (statt einer eindeutig erscheinenden Ziffer) die Personenabhängigkeit der Beurteilung sichtbar und damit auch diskutierbar. Allerdings beklagen Eltern häufig die Unverständlichkeit der Formulierungen – zumal wenn die Noten lediglich formal durch Textbausteine ersetzt werden.

Mit dieser Scheinform werden die Chancen nicht genutzt, die verbale Beurteilungen bieten und die ihr besonderes Potenzial ausmachen: Kompetenzen können inhaltlich beschrieben werden (differenziertes Leistungsprofil statt bloßer Ziffer). Bei der Erklärung von Leistungsbesonderheiten können äußere Bedingungen berücksichtigt werden (systemische Betrachtung statt einseitiger Personalisierung von Schwierigkeiten). Leistung kann als individuelle Entwicklung hin auf inhaltlich bestimmte Lernziele bewertet werden (persönlicher Lernzuwachs statt Vergleich mit einer Bezugsgruppe). Die Rückmeldung beschränkt sich nicht auf den Ausweis erbrachter Leistungen, sondern kann Folgerungen für die Förderung ziehen (nächste Schritte statt Lernstand). Kompetenzraster können zur Fokussierung hilfreich sein, aber nicht als additive Liste von Teilleistungen. Die beschriebenen Kompetenzen müssen sich einerseits an den angestrebten Zielen orientieren, andererseits entwicklungsbezogen Schritte, insbesondere »kritische Stellen« auf dem Weg dorthin beschreiben (→ »Aufgaben zur Beobachtung und Förderung« 2012). Denn Teilkompetenzen in Komponenten-Modellen lassen sich nicht mit Entwicklungsschritten für den Erwerb der übergreifenden Fähigkeit gleichsetzen.

Damit die Aussagekraft von Beurteilungen generell verbessert werden kann, sind folgende Anforderungen zu berücksichtigen:

- Für die Beurteilung sind Kriterien zu entwickeln, die sich auf die Ziele des Unterrichts und nicht nur auf spezifische Aufgaben beziehen. So kann Lehrkräften geholfen werden, ein tieferes Verständnis der Unterrichtziele zu gewinnen und die Beurteilung besser auf diese abzustimmen.
- Lehrer/innen brauchen mehr Aus- und Fortbildung, die sie für die Risiken der Leistungsbewertung sensibilisiert und auf ihre unterschiedlichen Funktionen vorbereitet. Erfolgreich sind solche Maßnahmen, wenn sie sich auf konkrete Beispiele einer *good practice* beziehen können.
- Eine wechselseitige Abstimmung der Kriterien – z. B. über den Austausch von Schülerarbeiten – hilft Lehrkräften, Klarheit über die Unterrichtziele und die darauf bezogenen Beurteilungskriterien zu gewinnen.

Zugleich wird deutlich: Ein Wechsel von Noten oder Tests zu Texten bringt nichts, wenn nicht auch der Unterricht Raum für die Unterschiede zwischen den Schüler/inne/n gibt und die individuelle Förderung in den Vordergrund tritt (→ »Öffnung des Unterrichts« 2008). Je stärker die Selektion ein Schulsystem bestimmt, desto schwerer wird es, Leistungen kriterien- und entwicklungsorientiert zu bewerten. Entscheidend ist außerdem eine Veränderung des sozialen Kontexts, innerhalb dessen Leistungen wahrgenommen und bewertet werden: von der Hierarchie zur Kooperation und zum Dialog.

6.3 Dialogische Verbindung von Fremd- und Selbstbeurteilungen in einer pädagogischen Leistungskultur

Welche Formen der Leistungsbeurteilung praktiziert werden, ist eng verknüpft mit dem Bild, das Lehrer/innen von Kindern, von ihrer Lernfähigkeit und von der eigenen Rolle haben. In den bisherigen Überlegungen ist Leistungsbewertung – wie traditionell in der Schule üblich – überwiegend als Urteil von Lehrenden über Lernende verstanden worden. Diese einseitige Abhängigkeit ist nicht mit der begrenzten Aussagekraft und hohen Fehleranfälligkeit von Fremdurteilen zu vereinbaren. Sie widerspricht zudem dem Anspruch einer demokratischen Schule, junge Menschen in der Entwicklung ihrer Selbstständigkeit zu unterstützen. Dies setzt die Bereitschaft und die Fähigkeit voraus, sich selbst und insbesondere konkrete Leistungen kritisch einzuschätzen. Beurteilung sollte deshalb in der Schule – wie zum Teil auch schon im Berufsleben üblich – dialogisch gestaltet werden. Die Kompetenz der Selbsteinschätzung zu fördern müsste ein zentrales Ziel von Unterricht sein. Empirische Studien zeigen außerdem, dass kontroll-orientierte Rückmeldungen nicht nur die Motivation, sondern auch die Leistung mindern. Selbst gute Noten wirken auf Dauer kontraproduktiv, weil sie die intrinsische Motivation beeinträchtigen.

Verbalbeurteilungen wird zu Recht ihre Subjektivität vorgehalten. Zugleich ist dies aber auch ihre Stärke, da sie – anders als Ziffernnoten – keine Objektivität vorgaukeln. Verbale Beurteilungen geben zurzeit zwar eher Auskunft über die Beurteilungspraxis der Lehrkraft; sie machen damit aber das Dilemma dieser Praxis offenkundig und zwingen zur Suche nach Verfahren, um die unvermeidliche Subjektivität zu kontrollieren. Leistungsbeurteilung ist ein kommunikativer Akt. Seine Mehrdeutigkeit kann nicht technisch-methodisch ausgeschaltet, sondern nur sozial kontrolliert werden.

Je nachhaltiger die Folgen einer Beurteilung, desto wichtiger die Beteiligung verschiedener Personen – vor allem der Betroffenen selbst. Fremd- und Selbstbeurteilungen sind miteinander auszuhandeln. An die Stelle des Zeugnisses »von oben« tritt das Abschlussgespräch zum (Halb-)Jahresende, in dem Schüler/in, Lehrkraft und Eltern ihre Wahrnehmungen der Entwicklung in den vorausgegangenen Monaten austauschen und gemeinsam Ziele für die kommende Zeit absprechen. Besondere Bedeutung kommt dabei der begleitenden Lernbeobachtung zu. Statt nur abzuprüfen, ob Ziele

erreicht sind, sollen Lernfortschritte und Schwierigkeiten im Lernprozess erfasst und diagnostisch gedeutet werden.

Dass die üblichen Klassenarbeiten und ihre Benotung diese Funktion nicht erfüllen können, ist seit vielen Jahrzehnten bekannt; standardisierte Tests, die als Alternative angeboten werden, haben – wie oben gezeigt – ihre eigenen Probleme. Es braucht Aufgaben, in denen nicht nur die Kinder etwas über die Sache lernen, sondern auch die Lehrer/innen etwas über die Vorstellungen, Strategien und vor allem Schwierigkeiten der Kinder (→ »Gute Aufgaben und Kompetenzorientierung« 2011). Der Grundschulverband hat deshalb sein Konzept »Pädagogische Leistungskultur« entwickelt (vgl. Bartnitzky u.a. 2005 ff.) und darin folgende Kriterien für Formen der Lernbeobachtung formuliert:

Sie sollten nach Möglichkeit
- der Lehrperson Informationen erbringen,
  - nicht nur über aktuelle Einzelleistungen,
  - sondern auch über zugrunde liegende Strategien (»Tiefenstrukturen«)
  - und über deren Entwicklung (»Lerngeschichte«);
- für die Schüler/innen auch inhaltlich eine produktive Lernsituation darstellen;
- dialogisch angelegt sein als wechselseitige Verständigung über Lernziele, Bewertungskriterien und tatsächliche Leistungen und damit die Fähigkeit der Kinder zur Selbsteinschätzung eigener Arbeiten entwickeln helfen.

Das diagnostische Repertoire von Lehrkräften kann durch heuristisch eingesetzte Tests, durch Beobachtungsbögen und verschiedene Dokumentationsformen (wie Portfolios) differenziert werden. Das Konzept einer pädagogischen Leistungskultur fordert daneben aber verschiedene »Institutionen« im Unterrichtsalltag, die den Schüler/inne/n helfen, ihre eigene Arbeit kritisch-konstruktiv zu überprüfen und an den Arbeiten anderer ihre Maßstäbe zu schärfen. Damit stellen sich besondere Anforderungen an die Ausbildung von Lehrkräften (→ »Lernbiografien im schulischen und außerschulischen Kontext« 2012). Diagnostische Kompetenz verlangt mehr, als standardisierte Tests einzusetzen und deren Ergebnisse interpretieren zu können. Insofern ist eine begleitende Lernbeobachtung umfassender anzulegen. Einige Beispiele aus dem Grundschulbereich:
- Schreibkonferenzen, in denen nach bestimmten Regeln Textentwürfe vorgestellt, kommentiert und dann mithilfe anderer Kinder überarbeitet werden;
- Rechendiskussionen in der Klasse, z.B. zum »harten Brocken des Tages«, sodass die Kinder Schwierigkeiten, Hypothesen und Lösungsstrategien austauschen und damit voneinander lernen können;
- im Sachunterricht, in Kunst und Musik Meta-Gespräche über Stärken und Schwächen von Präsentationen vor der Klasse, über Arbeitsergebnisse von Gruppen oder Einzelnen bis hin zu deren Bewertung durch das Plenum nach gemeinsam verabredeten Kriterien.

Ausgangspunkt ist dabei das Können der Kinder. Für die Begleitung des Schriftspracherwerbs hat Dehn (1994, S. 21) drei Fragen formuliert, die auf die weitere Entwicklung der vorhandenen Kompetenzen zielen – statt sie nur an vorgegebenen Zielen zu messen:
(1) Was kann das Kind schon?
(2) Was muss es noch lernen?
(3) Was kann es als nächstes lernen?

Der Lesedidaktiker Schmalohr (1997, S. 42 f.) hat diese Perspektive in seiner Arbeit mit jugendlichen und erwachsenen Analphabeten erweitert. Auch er hat drei einfache Fragen gestellt, um die Lernenden in die Mitverantwortung zu nehmen, um sie zum Nachdenken über ihre Probleme zu bringen und mit ihnen in ein Gespräch über sinnvolle Lernwege zu kommen:
(1) Wie lese ich, wo habe ich Schwierigkeiten?
(2) Woran könnte das liegen?
(3) Was kann ich tun?

Bosse u.a. (2015) haben in ihrer Diskussion über verschiedene Verfahren der Leistungsbeurteilung deutlich gemacht, dass jedes Instrument seine Risiken hat. Das gilt auch für Kompetenzraster und für Verbalbeurteilungen. Insofern gewinnt der soziale Kontext ihrer Umsetzung an Bedeutung. Dann lassen sich auch die Schwächen der Notenurteile ein Stück weit kompensieren, z. B. wenn ihre Bedeutung durch Kommentare differenziert und in einem Gespräch zwischen Schüler/in, Eltern und Lehrperson gemeinsam durchdacht wird.

In diesen Ansätzen zu einer »dialogischen Diagnostik« zeigt sich derselbe Geist wie im Peer-Review auf Schulebene: Herausforderung und Beratung statt Beurteilung und Kontrolle. Ein solches Verständnis von Evaluation respektiert und nutzt die Kompetenz der Betroffenen, ihre Probleme selbst zu erkennen, Ursachen für diese zu finden und Ideen zu ihrer Überwindung zu entwickeln – ohne sich ganz auf sie zu verlassen oder sie mit ihren Problemen allein zu lassen.

Meine abschließende Kernbotschaft: Bildungsforschung ist weder zu Vorschriften berechtigt noch in der Lage, Entscheidungen zu legitimieren. Ihre Aufgabe ist Anregung, Herausforderung und Kritik. Dialog und Unterstützung statt Autoritätsanspruch und Vorschrift – diese Prinzipien sollten auch die Evaluation im Bildungswesen generell bestimmen: im Verhältnis von Lehrern zu Schülern, aber auch von Wissenschaft zu Praxis. Dies folgt aus den Grenzen der jeweiligen Kompetenz, wie sie in diesem Buch vielfältig beschrieben worden sind. Es folgt aber auch aus den normativen Ansprüchen eines demokratischen Verständnisses von Schule und ihrer Evaluation.

7. Zehn Thesen zur Diskussion

1 *Fieber zu messen, ist noch keine Diagnose – und Fieber zu senken, keine Therapie.* Denn gleiche Testwerte bedeuten nicht gleiche Kompetenzen und ein Anstieg richtiger Lösungen nicht immer eine Verbesserung der Leistungen.

2 *Menschliche Verhaltensweisen – auch in standardisierten Tests – sind immer mehrdeutig:* Dieselben Aufgaben, aber auch deren Lösungen, werden von anderen Beteiligten je nach ihren Erfahrungen und Theorien unterschiedlich gelesen.

3 *Die Wirkungen pädagogischer Maßnahmen sind konzept- und kontextabhängig.* Menschen und ihr Lernen lassen sich nicht vermessen wie Maschinen und technische Prozesse – und noch vermessener wäre es zu glauben, sie ließen sich vergleichbar steuern und planen.

4 *Erfolge und Misserfolge von Interventionen streuen innerhalb eines jeden Ansatzes je nach Kontext breit.* Deshalb kann die Forschung zwar jeweils besondere Potenziale und Risiken benennen, aber nur in dieser vorsichtigen Form generelle Qualitätsurteile über konkurrierende Konzepte fällen.

5 *Zahlen sprechen nicht für sich.* Je komplexer die statistischen Verfahren, desto begründungsbedürftiger ist ihre Wahl und desto interpretationsbedürftiger sind ihre Ergebnisse. Sobald es zwei oder mehr Studien gibt, haben wir oft – wie in der Alltagserfahrung – widersprüchliche Ergebnisse; und damit denselben Interpretationsbedarf, der ebenfalls personenabhängig gedeckt wird.

6 *Die Stärken von Großstudien mit standardisierten Instrumenten sind zugleich ihre Schwächen:* Sie können nur wenige Faktoren erfassen, die Kategorien der Untersuchungen sind notwendig abstrakt und individuelle Bedeutungen gehen verloren.

7 *Wahrscheinlichkeitsaussagen, die aus der Untersuchung großer Populationen gewonnen werden, eignen sich nur sehr eingeschränkt für die Evaluation einzelner Schulen, Lehrer/innen oder Schüler/innen.* Die Forschung muss sich auf unterschiedliche Zielgruppen einstellen: Politik und Verwaltung sind primär an Häufigkeiten und Durchschnitten interessiert, Lehrer/innen und Eltern an Besonderheiten und Streuungen.

8 *Anders als im Verhältnis von Physik und Bautechnik lassen sich Allgemeinaussagen in der Pädagogik nicht einfach »anwenden«.* Sie müssen kontextbezogen

interpretiert werden. Insofern kann Evidenzbasierung Subjektivität nicht ausschalten.

9 *Die sozialen Kontexte und die politischen Implikationen von Evaluation werden nicht zureichend reflektiert.* Technisch-methodisch wird in der heutigen Bildungsforschung auf einem sehr viel höheren Niveau gearbeitet und diskutiert.

10 *Evaluation als Blick von außen ist wichtig, um scheinbare Selbstverständlichkeiten in der Praxis infrage zu stellen – »von außen« heißt aber nicht zwingend »von oben«.* Dialog und Unterstützung statt Autoritätsanspruch und Vorschrift – nur so wird Evaluation im Bildungswesen produktiv. Die besondere Kompetenz von Praktiker/inne/n anzuerkennen, bedeutet, dass man von ihnen keine Fortbildung über empirische Methoden fordert, wenn man nicht auch von den Wissenschaftler/inne/n verlangt zu lernen, was es heißt, unter besonderen Bedingungen und unter Zeitdruck Entscheidungen zu fällen.

Fazit: Weder ist Bildungsforschung zu Vorschriften berechtigt noch ist sie in der Lage, Handlungsentscheidungen aus Daten abzuleiten. Ihre Aufgabe ist Anregung, Herausforderung und kritische Befragung.

Hinweise für eine vertiefende Lektüre

Neben meinen eigenen (im Text durch → für den Download zugänglich gemachten) Ergänzungsbeiträgen empfehle ich zur Vertiefung der Diskussion folgende Bücher zu den einzelnen Kapiteln:

Kap. 1
Der erste deutschsprachige Sammelband zum Thema »Evaluation« (Wulf 1972) gibt auch heute noch einen guten Einblick in die Grundfragen, vor denen eine jede Begleitforschung steht. Besondere Anforderungen und Probleme von »Leistungsmessungen in Schulen« werden – von der Schüler- bis zur Systemebene – in dem gleichnamigen Sammelband von Weinert (2001) differenziert diskutiert.

Kap. 2
Einen Überblick über die wesentlichen Elemente einer output-orientierten Steuerung bzw. Evaluation von Einrichtungen und. Maßnahmen im Bildungssystem geben Oelkers/Reusser (2008) in ihrer BMBF-Expertise »Qualität entwickeln – Standards sichern – mit Differenz umgehen«. Ein besseres Verständnis grundlegender Verfahren und Fallen statistischen Denkens vermittelt – auf recht unterhaltsame Weise – das Buch »Der Hund, der Eier legt« von Beck-Bornholdt/Dubben (2001).

Kap. 3
Der Ansatz sogenannter Meta-Analysen und ganz konkret die viel zitierte Meta-Meta-Studie von Hattie (2013) werden aus verschiedenen Perspektiven verständlich kommentiert in den Beiträgen zu dem Sammelband »Die Hattie-Studie in der Diskussion«, herausgegeben von Terhart (2014). Inhaltliche Erträge und methodische Schwierigkeiten von empirischer Forschung im Bildungsbereich habe ich an alltäglichen Problemen von Unterricht und bildungspolitischen Entscheidungsfragen erläutert in meinem Buch »Schule verstehen und gestalten« (2005).

Kap. 4
Ein erstes, auch selbstkritisches Resümee ihrer Arbeit im PISA-Programm ziehen Klieme u. a. (2010) in »PISA 2009. Bilanz nach einem Jahrzehnt«. Grundsätzlich – und mit Beispielen aus den angelsächsischen Ländern illustriert – stellt Herzog (2013) den Ansatz einer output-orientierten Steuerung und Evaluation in seinem Buch »Bildungsstandards« infrage. PISA selbst wird – auch unter forschungstechnischen Gesichtspunkten – kritisch diskutiert in den Beiträgen zu »PISA zufolge PISA« von Hopmann u. a. (2007).

Kap. 5

Evaluation beginnt mit Selbstbeobachtung und -kritik, wie der Sammelband »Schulinterne Evaluation an Reformschulen« von Tillman/Wischer (1998) anhand unterschiedlicher Konzepte und Praxisbeispiele zeigt. In den Publikationen des Schulverbunds »Blick über den Zaun« finden sich vielfältige Anregungen für Fremdevaluation als »Außenblick auf Augenhöhe«: Als Alternative zu den output-orientierten Bildungsstandards der KMK haben Groeben u.a. (2005) mit »Unsere Standards« Kriterien für Prozesse und Bedingungen »guten« Unterrichts vorgelegt; ergänzend beschreiben Backhaus u.a. (2010) das Verfahren eines Peer-Review über wechselseitige Schulbesuche.

Kap. 6

Die breite Forschung zu Ziffernnoten und zu alternativen Formen der Leistungsbeurteilung ist übersichtlich aufgearbeitet in der Expertise unserer Arbeitsgruppe Primarstufe (2006/2014) »Sind Noten nützlich und nötig?« Konkrete Hilfen für notenfreie Formen der Lernbeobachtung finden sich für alle Fächer und Jahrgangsstufen der Grundschule in den Heften »Pädagogische Leistungskultur« von Bartnitzky u.a. (2005 ff.). Für die Sekundarstufe verweise ich auf den Band »Leistungsbewertung« von Winter (2004/2012).

Literatur

In dieser Liste werden nur die explizit zitierten Quellen nachgewiesen. Eine vollständige Liste der verwendeten Veröffentlichungen (mit ihren Bezügen zur Argumentation im Text) kann angefordert werden über hans.bruegelmann@grundschulverband.de.

AERA (2003): Standards and tests: Keeping them aligned. In: Research Points (American Educational Research Association), Vol. 1, No. 1, 1-4.

Albrecht, H. (2005): Kritik der reinen Norm. Klinische Forschung hilft vor allem Standardpatienten. In: DIE ZEIT, Nr. 2, v. 5.1.2005, S. 25.

Albrecht, H. (2008): Jeder Patient zählt. Was klinische Großstudien übersehen. in: Die Zeit. Nr. 22 v. 21.5.2008, 33.

Altrichter, H. (2010). Schul- und Unterrichtsentwicklung durch Datenrückmeldung. In: Altrichter/Maag (2010, 219-254).

Altrichter, H./Maag Merki, K. (Hrsg.) (2010): Handbuch neue Steuerung im Schulsystem. VS Verlag für Sozialwissenschaften: Wiesbaden.

Arbeitsgruppe Primarstufe (2006): Sind Noten nützlich und nötig? Zifferzensuren und ihre Alternativen im empirischen Vergleich. Eine wissenschaftliche Expertise des Grundschulverbandes, erstellt von der Arbeitsgruppe Primarstufe an der Universität Siegen (Hans Brügelmann mit Axel Backhaus u. a.). Grundschulverband e.V.: Frankfurt (2. erg. Aufl. 2014).

Argyris, C./Schön, D. (1978): Organization learning: A theory of Action perspective. Addison Wesley: Reading (MA).

Backhaus, A., u. a. (2009): »Blick über den Zaun«: Schulen lernen von Schulen. Vorschläge zur Planung und organisatorischen Ausgestaltung von Peer-Reviews durch kritische Freunde. Reformpädagogische Arbeitsstelle »Blick über den Zaun« an der Universität: Siegen. www.blickueberdenzaun.de/index.php/textarchiv/167-schulen-lernen-von-schulen-handreichung-fuer-schulbesuche

Balhorn, H./Brügelmann, H. (Hrsg.) (1995): Rätsel des Schriftspracherwerbs. Neue Sichtweisen der Forschung. Auswahlband Theorie der DGLS-Jahrbücher 1-5. Libelle: CH-Lengwil.

Bartnitzky, H. (2008): VerA Deutsch 2008: Wie Kindern durch Vergleichsarbeiten Unrecht geschieht. In: Grundschule aktuell, H. 103 4-8.

Bartnitzky, H., u. a. (Hrsg.) (2005&2006&2007): Pädagogische Leistungskultur: Materialien für Klasse 1/2 und Klasse 3/4. Beiträge zur Reform der Grundschule, Bd. 119 & 121 & 123. Grundschulverband: Frankfurt.

Bauersfeld, H. (1972): Einige Bemerkungen zum Frankfurter Projekt und zum »alef«-Programm. In: Schwartz u. a. (1972, 237-246).

Baumert, J. (2014): Im Fokus der Lehrerfortbildung: Optimierung des Lernens. Jürgen Baumert im Gespräch mit Botho Priebe. In: Lernende Schule, H. 68, 22-24.

Baumert, J., u. a. (1980): Bildung in der Bundesrepublik Deutschland. Bd. 2: Gegenwärtige Probleme. Hrsg. von der Projektgruppe Bildungsbericht am Max-Planck-Institut für Bildungsforschung. Frankfurt.

Beck-Bornholdt, H.-P./ Dubben, H.-H. (2001): Der Hund, der Eier legt. Erkennen von Fehlinformationen durch Querdenken. Rororo 60359: Reinbek (überarb. Neuausgabe; 1. Aufl. 1997)

Beywl, W./Zierer, K. (2013): Lernen sichtbar machen. Zur deutschsprachigen Ausgabe von »Visible Learning«. In: Hattie (2013, VI-XXVI).

Bellmann, J./Müller, T. (Hrsg.) (2011): Wissen, was wirkt. Kritik evidenzbasierter Pädagogik. VS Verlag für Sozialwissenschaften: Wiesbaden.

Berliner, D. C. (2002): Educational research: The hardest science of all. In: Educational Researcher, Vol. 31, No. 8, 18-20.

Biesta, G. (2011): Warum »What works« nicht funktioniert: Evidenzbasierte pädagogische Praxis und das Demokratiedefizit der Bildungsforschung. In: Bellmann/Müller (2011, 95-121).

BMBF (Hrsg.) (2008): Rahmenprogramm zur Förderung der empirischen Bildungsforschung. Bildungsforschung Band 22. Bundesministerium für Bildung und Forschung: Berlin.

Bos, W., u. a. (Hrsg.) (2007a): IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich. Waxmann: Münster.

Bos, W., u. a. (2007b): Zusammenfassung wichtiger Ergebnisse zu Kompetenzen und Einstellungen von Hamburger Schülerinnen und Schülern. In: Bos/Pietsch (2007, 1-8)

Bos, W./Pietsch, M. (Hrsg.) (2007): KESS 4 – Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen. Waxmann: Münster.

Bosse, U., u. a. (2015): Noten, Raster und Berichte. Nach der Bielefelder Herbsttagung: Hans Brügelmann und Ulrich Hecker im Gespräch. In: Grundschule aktuell, Nr. 129, 3-5.

Böttcher, W. (2013): Das Monitoring-Paradigma – Eine Kritik der deutschen Schulreform. In: Empirische Pädagogik, 27. Jg., H. 4, 5-21.

Böttcher, W. (2013): Das Monitoring-Paradigma – Eine Kritik der deutschen Schulreform. In: Empirische Pädagogik, 27. Jg., H. 4, 5-21.

Böttcher, W., u. a. (Hrsg.) (2010): Evaluation, Bildung und Gesellschaft. Steuerungsinstrumente zwischen Anspruch und Wirklichkeit. Waxmann: Münster.

Brinkmann, E., u. a. (2008ff.): ABC-Lernlandschaft. vpm/Klett: Stuttgart.

Bromme, R./Prenzel, M. (Hrsg.) (2014): Von der Forschung zur evidenzbasierten Entscheidung. Die Darstellung und das öffentliche Verständnis der empirischen Bildungsforschung. Zeitschrift für Erziehungswissenschaft, Sonderheft, Bd. 17. VS Verlag für Sozialwissenschaften: Wiesbaden.

Bromme, R., u. a. (2014): Empirische Bildungsforschung und evidenzbasierte Bildungspolitik. Eine Analyse von Anforderungen an die Darstellung, Interpretation und Rezeption empirischer Befunde. In: Bromme/Prenzel (2014, 3-54).

Brügelmann, H. (2004): Leseleistungen von Lehrer/innen und Lehramtsstudierenden im Stolperwörter-Lesetest. Erste Befunde und ihre Deutung. Projekt LUST/FB 2 der Universität: Siegen. Download: http://www2.agprim.uni-siegen.de/lust/stolper_auswertung_lehramt.pdf

Brügelmann, H. (2005): Schule verstehen und gestalten – Perspektiven der Forschung auf Probleme von Erziehung und Unterricht. Libelle: CH-Lengwil (bis 2008 fortlaufend aktualisiert unter: http://www2.agprim.uni-siegen.de/schuleverstehen).

Brügelmann, H. (2007): Scharfe Brillen, wache Augen und ein einfühlsamer Blick. Wie Reformschulen den fremden Blick kritischer Freunde am besten nutzen können: Zur Bedeutung von technischer Präzision und sozialer Kontrolle bei der Evaluation pädagogischer Standards. In: Schulverbund »Blick über den Zaun« (2007, 8-52).

Brügelmann, H., u. a. (1994): »Schreibvergleich BRDDR« 1990/91. In: Brügelmann/Richter (1994, 129 134).

Brügelmann, H./Backhaus, A. (2011): Kinder als Kinderschützer. In: Buchholz u. a. (2011, 309-318).

Brügelmann, H./Richter, S. (Hrsg.) (1994): Wie wir recht schreiben lernen. Zehn Jahre Kinder auf dem Weg zur Schrift. Libelle Verlag: CH-Lengwil (2. Aufl. 1996).

Buchholz, T., u. a. (Hrsg.) (2011): Kinderschutz in gemeinsamer Verantwortung von Jugendhilfe und Schule. VS Verlag für Sozialwissenschaften: Wiesbaden.

Dehn, M. (1994): Schlüsselszenen zum Schriftspracherwerb. Arbeitsbuch zum Lese- und Schreibunterricht in der Grundschule. Beltz: Weinheim.

Demmer, M. (2014): VerA (Vergleichsarbeiten): Praktiken, Einstellungen und Erfahrungen der GEW-Mitglieder in Grundschulen. Pilotstudie auf der Basis einer GEW-Mitgliederbefragung.

Download: www.gew.de/Binaries/Binary112804/MD_VerA-Befragung_GEW_2014-Auswertung_und_KommentierungDECKBLATT.pdf
Dilthey, W. (1957): Gesammelte Schriften. Bd. 5. Teubner: Stuttgart.
DiSalvo, D. (2014): Was Ihr Gehirn glücklich macht ... und warum Sie genau das Gegenteil tun sollten. Springer Spektrum: Berlin/Heidelberg.
Ditton, H. (2010): Wie viel Ungleichheit durch Bildung verträgt eine Demokratie? In: Zeitschrift für Pädagogik, 56. Jg., H. 1, 53-68.
Dohmen, G. (2001): Das informelle Lernen. Die internationale Erschließung einer bisher vernachlässigten Grundform menschlichen Lernens für das lebenslange Lernen aller. Bundesministerium für Bildung und Forschung: Bonn. Download: www.bmbf.de/pub/das_informelle_lernen.pdf.
Dolin, J. (2007): PISA – an example of the use and misuse of large-scale comparative tests. In: Hopmann u. a. (2007, 93-126).
Elliott, J./Kushner, S. (2007): The need for a manifesto for educational programme evaluation. In: Cambridge Journal of Education, Vol. 37, No. 3, 321-336.
Faber, G./Billmann-Mahecha, E. (2010): Notengebung im Spiegel wissenschaftlicher Untersuchungen. Probleme, Erfordernisse und Möglichkeiten aus pädagogisch-psychologischer Sicht. In: Lernchancen, 13. Jg., H. 74, 30-33.
Fauser, P., u. a. (Hrsg.) (2007): Was für Schulen! Portraits der Preisträgerschulen und der nominierten Schulen des Wettbewerbs 2006. Klett Kallmeyer: Stuttgart.
Fay, J. (2010): Die Entwicklung der Rechtschreibkompetenz beim Textschreiben. Eine empirische Untersuchung von Klasse 1 bis 4. Peter Lang: Frankfurt.
Fend, H. (2008): Schule gestalten. Systemsteuerung, Schulentwicklung und Unterrichtsqualität. VS Verlag für Sozialwissenschaften: Wiesbaden.
Fischer, D. (Hrsg.) (1982): Fallstudien in der Pädagogik. Ekkehard Faude: Konstanz.
Freud, S. (1954): Zur Psychopathologie des Alltagslebens. Fischer-Taschenbuch 68: Frankfurt (Erstauflage 1941).
Fritz-Vannahme, J. (1996): Waren damals alle besser? In Frankreich traten Schüler von 1995 gegen die von 1925 an. In: Die Zeit, Nr. 17 v. 19.4.1996, 41.
Gaeth, F. (2009): PISA: Zukunft der Bildung oder Joulupukki? In: Politik-Poker v. 1.10.2009. Download: www.politik-poker.de/pisa-zukunft-der-bildung-oder-joulupukki.php
Gigerenzer, G. (2007): Bauchentscheidungen. Die Intelligenz des Unbewussten und die Macht der Intuition. Bertelsmann: München.
Groeben, A. v. d., u. a. (2005): Unsere Standards. Ein Diskussionsentwurf, vorgelegt von »Blick über den Zaun« – Bündnis reformpädagogisch engagierter Schulen. In: Neue Sammlung, 45. Jg., H. 2, 253-297 (Download über www.blickueberdenzaun.de).
Grotlüschen, A./Riekmann, W. (Hrsg.) (2012): Funktionaler Analphabetismus in Deutschland. Ergebnisse der ersten leo.-Level-One Studie. Waxmann: Münster.
Hattie, J.A.C. (2013): Lernen sichtbar machen. Überarbeitete deutschsprachige Ausgabe von »Visible Learning« besorgt von Wolfgang Beywl und Klaus Zierer. Schneider Hohengehren: Baltmannsweiler (engl. 2009).
Helmke, A., u. a. (2014): EMU – Unterrichtsdiagnostik. Studienbrief Version 4.0. Universität: Koblenz-Landau
Henrich, J., et al. (2010): The weirdest people in the world. In: Behavioral and Brain Sciences, Vol. 33, No. 2-3, 61-83.
Herzog, W. (2013): Bildungsstandards. Eine kritische Einführung. Kohlhammer: Stuttgart.
Heymann, H.W. (Hrsg.) (1997): Allgemeinbildung und Fachunterricht. Bergmann + Helbig: Hamburg.
Hopmann, S. T., u. a. (Hrsg.) (2007): PISA zufolge PISA/PISA According to PISA. Lit-Verlag: Wien/Berlin.
House, E. R. (1980): Evaluating with validity. Sage: Beverly Hills/London.

Ingenkamp, K. (Hrsg.) (1971): Die Fragwürdigkeit der Zensurengebung. Beltz: Weinheim/Basel (9. Aufl. 1995).
Jahnke, T./Meyerhöfer, W. (Hrsg.) (2006): Pisa & Co. Kritik eines Programms. 2. überarb. Aufl. Franzbecker: Hildesheim (2. überarb. Aufl. 2007).
Kirkland, M. C. (1971). The effects of tests on students and schools. In: Review of Educational Research, Vol. 41, No. 4, 303-350.
Klawans, H. (2005): Die Höhlenfrau, die Sprache und wir. 13 merkwürdige Geschichten über das menschliche Gehirn. Klett-Cotta: Stuttgart (engl. 2000).
Klieme, E., u. a. (2010b): PISA 2000–2009: Bilanz der Veränderungen im Schulsystem. In: Klieme u. a. (2010a, 277-300).
Klieme, E., u. a. (Hrsg.) (2010a): PISA 2009. Bilanz nach einem Jahrzehnt. Waxmann: Münster.
Klieme, E./Maichle, U. (1990): Ergebnisse eines Trainings zum Textverstehen und zum Problemlösen in Naturwissenschaften und Medizin. In: Trost (1990, 258-307).
KMK (1968): Erläuterung der Notenstufen bei Schulzeugnissen und Einzelergebnissen in staatlichen Prüfungszeugnissen. Vereinbarung vom 3.10.1968. Kultusministerkonferenz: Bonn.
Knighton, T./Bussière, P. (2006): Educational outcomes at age 19 assoiciated with reading ability at age 15. Statistics Canada: Ottawa.
Kochan, B. (1995): Von der Untersuchung des »Lernens durch Instruktion« zur Untersuchung des »Lernens durch Gebrauch«. In: Balhorn/Brügelmann (1995, 26-28).
Linn, R. L. (2000): Assessments and accountability. In: Educational Researcher, Vol. 29, No. 2, 4-15.
Lurija, A. R. (1991): Der Mann, dessen Welt in Scherben ging. [Und: Kleines Porträt eines großen Gedächtnisses]. Rowohlt: Reinbek (russ. 1968 bzw. 1971, engl. 1968 und 1972).
Maaz, K., u. a. (2011): Herkunft zensiert? Leistungsdiagnostik und soziale Ungleichheiten in der Schule. Vodafone Stiftung Deutschland: Düsseldorf.
MacDonald, B. (1976): Evaluation and the control of education. In: Tawney (1976, 125-136; Nachdruck aus: MacDonald/Walker 1974).
MacDonald, B./Walker, R. (eds.) (1974): Innovation, Evaluation, Research and the Problem of Control. Some interim papers. SAFARI project/Centre for Applied Research in Education/U.E.A.: Norwich.
MacDonald, B./Walker, R. (1976): Changing the curriculum. Open Books: London.
Makel, M. C./Plucker, J. A. (2014): Facts are more important than novelty. Replication in the education sciences. In: Educational Researcher, Vol. 43, No. 6, 304–316.
Mühlhausen, U. (1994): Überraschungen im Unterricht. Situative Unterrichtsplanung. Beltz Praxis: Weinheim/Basel.
NCES (2010): The Nation s Report Card: Reading 2009. National Center for Education Statistics: Washington.
Neuweg, G. H. (2015): Das Schweigen der Könner. Gesammelte Schriften zum impliziten Wissen. Waxmann: Münster/New York.
Oelkers, J./Reusser, K. (2008): Qualität entwickeln – Standards sichern – mit Differenz umgehen. Eine Expertise. Bundesministerium für Bildung und Forschung: Berlin.
Piaget, J. (1980): Das Weltbild des Kindes. Klett-Cotta im Ullstein-Taschenbuch 39001: Frankfurt u. a. (frz. 1926).
Pirsig, R. (1978): Zen und die Kunst, ein Motorrad zu warten. Fischer-Taschenbuch: Frankfurt (engl. 1974).
Polanyi, M. (1958): Personal knowledge: Towards a post-critical philosophy. Routledge and Kegan Paul: London.
Polanyi, M. (1966): The tacit dimension. Doubleday: New York.
Popper, K. R. (1982): Logik der Forschung. Mohr Siebeck: Tübingen (7. erw. Aufl.; 1. Aufl. 1935).
Prenzel, M. (2007): »Wir haben einen Sprung nach vorn gemacht«. Was Experten jetzt von Schule und Politik erwarten. Pisa-Leiter Manfred Prenzel erklärt die Lernerfolge deutscher Schüler –

und verteidigt seine Studie. Interview von T. Warnecke in: Tagesspiegel v. 5.12.2007.
Rackwitz, R.-P. (2010): Tests unter der Lupe: Können sie halten, was sie versprechen? In: Die Grundschulzeitschrift, 24. Jg., H. 234, 4 – 9..
Ratzka, N. (2003): Mathematische Fähigkeiten und Fertigkeiten am Ende der Grundschulzeit – Empirische Studien im Anschluss an TIMSS (Phil. Diss. FB 2 der Universität Siegen). Franzbecker: Hildesheim/Berlin.
Reichen, J. (1982): Lesen durch Schreiben. Leselehrgang, Schülermaterial und Lehrerkommentar. Sabe: Zürich (Neufassung 2001 »Hannah hat Kino im Kopf«: Die Reichen Methode Lesen durch Schreiben und ihre Hintergründe für Lehrerinnen und Lehrer, Studierende und Eltern. Heinevetter-Verlag: Hamburg).
Retzl, M. (2010): Schulqualität entwickeln durch nationale Standards? Grundlegungen für einen alternativen Ansatz der Qualitätsentwicklung. In: Böttcher u. a. (2010, 355-365).
Richter, D., u. a. (2014): Überzeugungen von Lehrkräften zu den Funktionen von Vergleichsarbeiten: Zusammenhänge zu Veränderungen im Unterricht und den Kompetenzen von Schülerinnen und Schülern. In: Zeitschrift für Pädagogik, 60. Jg., H. 2, 225-244.
Richter, S./Brügelmann, H. (Hrsg.) (1994): Mädchen lernen ANDERS lernen Jungen. Geschlechtsspezifische Unterschiede beim Schriftspracherwerb. DGLS Reihe »Lesen und Schreiben«. Libelle: CH Lengwil (2. Auf. 1996).
Rolff, H.-G. (2015): Sind schulische Strukturfaktoren wirklich unwichtig? In: b&w, Nr. 3/2015, 26-29.
Rolff, H.-G. (2013): Die Hattie –Studie: Ein Rorschach-Test. In: Pädagogik, 65. Jg. H. 3, 46-49 (auch in Terhart 2014).
Roth, H. (1962): Die realistische Wende in der pädagogischen Forschung. In: Neue Sammlung, 2. Jg., 481-490.
Rumpf, H. (1971): Scheinklarheiten. Sondierungen von Schule und Unterrichtsforschung. Westermann: Braunschweig.
Sackett, D. L., et al. (2000): Evidence-based medicine: How to practice and teach EBM (2nd ed.). Churchill Livingstone: Edinburgh (2nd ed.1st ed. 1997).
Sacks, O. (1987): Der Mann, der seine Frau mit einem Hut verwechselte. Rowohlt: Reinbek (engl. 1985, 1987).
Schmalohr, E. (1997): Das Erlebnis des Lesens. Grundlagen einer erzählenden Lesepsychologie. Klett-Cotta: Stuttgart.
Schmidt, M. (2014): Ameisen ohne Job. In China steigt die Zahl der Hochschulabsolventen stetig an. Dabei sind ihre Berufsaussichten schlecht. In: Die Zeit, Nr. 44 v. 23.10.2014, 81.
Schön, D. (1983): The reflective practitioner. Temple Smith: London.
Schründer-Lenzen, A. (Hrsg.) (2006): Risikofaktoren kindlicher Entwicklung. Migration, Leistungsangst und Schulübergang, VS Verlag für Sozialwissenschaften: Wiesbaden.
Schründer-Lenzen, A./Merkens, H. (2006): Differenzen schriftsprachlicher Kompetenzentwicklung bei Kindern mit und ohne Migrationshintergrund. In: Schründer-Lenzen (2006, 15-44).
Schulverbund »Blick über den Zaun« (Hrsg.) (2007): Beobachten, bewerten, beraten. Verfahren und Werkzeuge für eine andere Evaluation. Bezug über: www.blickueberdenzaun.de/index.php/textarchiv/168-evaluation-beobachten-bewerten-beraten
Schwartz, E., u. a. (Hrsg.) (1972): Materialien zum Mathematikunterricht in der Grundschule. Beiträge zur Reform der Grundschule Bd. 13. Arbeitskreis Grundschule: Frankfurt.
Schweinhart, L. J., u. a. (1993): Significant benefits: The High/Scope Perry Preschool Study through age 27. Monographs of the High/Scope Educational Research Foundation, Nr. 10. High/Scope Press: Ypsilanti.
Seydel, O. (2005): »Hilfe! Der Inspektor kommt« Oder: Sind Schulen Kunstwerke? In: Pädagogik, 57. Jg., H. 9, 10-15.
Seydel, O. (2007): »Blick über den Zaun ». Wie Schulen voneinander lernen können. In: Schulverbund »Blick über den Zaun » (2007, 2-7).

Seydel, O. (2009): »Eine Schule ist kein Auto«. In: Grundschulzeitschrift, 23. Jg., H. 223, 4-7.
Stake, R. E. (1972): Verschiedene Aspekte pädagogischer Evaluation. In: Wulf (1972, 92-112).
Steinig, W., u.a. (2009): Schreiben von Kindern im diachronen Vergleich. Texte von Viertklässlern aus den Jahren 1972 und 2002. Waxmann: Münster.
Steinig, W./Betzel, D. (2014.): Schreiben Grundschüler heute wirklich schlechter als vor 40 Jahren? Texte von Viertklässlern aus den Jahren 1972, 2002 und 2012. In: Plewnig/Witt (2014, 353-371).
Stenhouse, L. (1975): An introduction to curriculum research and development. Heinemann Educational Books: London et al. (teilweise dt. Zusammenfassung in: Zeitschrift für Pädagogik, 19. Jg., H. 3, S. 447-452).
Stenhouse, L. (1982): Pädagogische Fallstudien: Methodische Traditionen und Untersuchungsalltag. In: Fischer (1982, 24-61; engl. 1981).
Taleb, A. N. (2014): Antifragilität. Anleitung für eine Welt, die wir nicht verstehen. btb Random House: München (1. Aufl. Albrecht Knaus: München; engl. 2012).
Tawney, D. (ed.) (1975): Curriculum evaluation today: Trends and implications. Macmillan: London.
Terhart, E. (Hrsg.) (2014): Die Hattie-Studie in der Diskussion. Probleme sichtbar machen. Klett Kallmeyer: Seelze.
Tillmann, K.-J. (2008): Leistungsvergleichsstudien und Systemsteuerung – oder: Was hat PISA zu bildungspolitischen Entscheidungen beigetragen? Ms. des Vortrags auf dem Symposion zum Gedenken an Hermann Lange am 1.10.2008 in der Universität Hamburg. Download: www.gew.de/Binaries/Binary38502/TillmannVortragHamburg1008.pdf
Tillmann, K.-J./Wischer, B. (Hrsg.) (1998): Schulinterne Evaluation an Reformschulen. Positionen, Konzepte, Praxisbeispiele. Impuls30. Laborschule an der Universität: Bielefeld.
Trost, G. (Hrsg.) (1990): Test für medizinische Studiengänge (TMS). 14. Arbeitsbericht. Institut für Test- und Begabungsforschung: Bonn.
Wagener, A. L. (Hrsg.) (2012): Bill Gates, Kesha und Tom im Siegerland. Kinder und ihre Lernbiografien – Beiträge aus dem Projekt LISA&KO. Arbeitsgruppe Primarstufe in der Fakultät II. Universi Verlag: Siegen.
Weinert, F. E. (Hrsg.) (2001): Leistungsmessungen in Schulen. KMK: Bonn/ Beltz: Weinheim.
Weinert, F.E./Helmke, A. (Hrsg.) (1997): Entwicklung im Grundschulalter. Beltz Psychologie Verlags Union: Weinheim.
Winter, F. (2004): Leistungsbewertung. Eine neue Lernkultur braucht einen anderen Umgang mit Schülerleistungen. Schneider Hohengehren: Baltmannsweiler (5. überarb. und erw. Neuauflage 2012).
Wulf, C. (Hrsg.): (1972). Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen. München.
Wuttke, J. (2006): Fehler, Verzerrungen, Unsicherheiten in der PISA-Auswertung. In: Jahnke/Meyerhöfer (2006, 101-154).
Wuttke, J. (2007): Uncertainties and bias in PISA In: Hopmann u.a. (2007, 241-264).

Quellennachweise

Für diesen Band wurden Auszüge aus folgenden Texten früherer Veröffentlichungen überarbeitet:

»BABEEK und die Kunst, ein Curriculum zu evaluieren«, Manuskript zu einem Vortrag am 20.2.1978 auf der Tagung »Curriculum-Entwicklung – Erfahrungen und neue Perspektiven« am Institut für die Pädagogik der Naturwissenschaften in Kiel → Kap. 0

»Scharfe Brillen, wache Augen und ein freundlicher Blick. Wie Reformschulen den fremden Blick kritischer Freunde am besten nutzen können: Zur Bedeutung von technischer Präzision und sozialer Kontrolle bei der Evaluation pädagogischer Standards«, in: Schulverbund »Blick über den Zaun« (Hrsg.) (2007): Beobachten, bewerten, beraten. Verfahren und Werkzeuge für eine andere Evaluation → Kap. 0 und 5.2

»Qualität und die Kunst, den Erfolg von Unterricht zu messen – oder: Sieben Mythen der aktuellen Diskussion über Evaluation und Rechenschaft«, in: Burkard, C./Haenisch, H. (Red.) (2000): Was ist guter Fachunterricht? Landesinstitut für Schule und Weiterbildung: Soest/Kettler: Bönen, S. 9-41 → Kap. 0 und 5.1

»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, S. 221-234 → Kap. 1

»Kontroversen um die Schulleistungsmessung in Deutschland. Eine fingierte Diskussion über Positionen und Perspektiven in verteilten Rollen«, in: Weinert, F. E. (Hrsg.) (2001): Leistungsmessungen in Schulen. Beltz: Weinheim, S. 33-44, → Kap. 2.1

»Was leisten unsere Schulen?«, in: Bartnitzky, H., u. a. (1999): Zur Qualität der Leistung – 5 Thesen zu Evaluation und Rechenschaft der Grundschularbeit. Grundschulverband e.V.: Frankfurt a. M. → Kap. 2.7

»Metaanalysen der Unterrichtsforschung. Hat Hatties ›Visible Learning‹ einen Nutzen für die pädagogische Arbeit vor Ort?«, in: Lehren und Lernen, 39. Jg., H. 7, S. 20-22 → Kap. 3.1

»Über die Erkenntnisgrenzen von PISA und notwendige Ergänzungen. Ein Gespräch zwischen Imbke Behnken., Hans Brügelmann und Hans Werner Heymann mit ›extrakte‹«, in: Wagener, A. L. (Hrsg.) (2012): Bill Gates, Kesha und Tom im Siegerland. Kinder und ihre Lernbiografien – Beiträge aus dem Projekt LISA&KO. Universi Verlag: Siegen, S. 241-249 → Kap. 3.2

»PISA & Co: Funktion und Grenzen von Leistungsvergleichen – Wie Tests durch eigene Ansprüche und öffentliche Erwartungen leicht überfordert werden« (2011) http://dl.dropbox.com/u/21116330/Lit-Homepage/brue.pisa.enzyklop%C3%A4die_online.dippelhofer.MS_layout_EEO.110813.pdf → Kap. 4

»Thesen zu den Bildungsstandards«, in: Grundschule aktuell, Nr. 90 (2005), S. 8 → Kap. 5.1

»Wahrheit durch VerA? Anmerkungen zum ersten Durchgang der landesweiten Leistungstests in sieben Bundesländern«, in: Grundschule aktuell, Nr. 89 (2005), S. 7-9 → 5.2

»VerA 2010: Für die Schulen auch dieses Mal wenig Ertrag – bei hohem Aufwand und oft unerfreulichen Nebenwirkungen«, in: Grundschule aktuell, H. 111 (2010), S. 20 → Kap. 5.2

»Sind Noten nützlich und nötig? Zifferzensuren und ihre Alternativen im empirischen Vergleich«, Eine wissenschaftliche Expertise des Grundschulverbandes, erstellt von der Arbeitsgruppe Primarstufe an der Universität Siegen. Grundschulverband e.V.: Frankfurt a.M. (2006; 2014) → Kap. 6

Glossar

Die nationale Repräsentativstudie **DESI** konzentrierte sich auf die Schülerleistungen in Deutsch und Englisch zu Beginn und zu Ende des 9. Jahrgangs sowie auf die Leitvorstellungen der Lehrer/innen und ihren Unterricht.

Längsschnittstudien* sind nur aussagekräftig, wenn über die Zeit hinweg nicht zu viele Teilnehmer/innen verloren gehen. Größere Zahlen an solchen **Dropouts** können die Aussagekraft der Studien inhaltlich beeinträchtigen, weil es oft systematische Gründe sind, die zum Aussteigen von Personen(gruppen) führen.

Die **Effektstärke** ist ein Maß für die Stärke eines Zusammenhangs zwischen zwei Faktoren (z. B. Geschlecht und Leseleistung). Sie ist umso größer, je weiter die Mittelwerte von zwei Gruppen (z. B. Jungen und Mädchen) in einem Test (z. B. zum Lesen) auseinanderliegen und je weniger sich die beiden Gruppen unter diesem Aspekt überlappen. Die Effektstärke wird meist zusätzlich berechnet, da auch praktisch kaum bedeutsame Unterschiede statistisch signifikant* werden, wenn die Stichproben nur groß genug sind.

Die internationale **IEA**-Lesestudie wurde 1991 in dritten und achten Klassen auch in Deutschland durchgeführt – kurz nach der Wende getrennt nach Ost- und Westdeutschland.

Kalibrierung bezeichnet die Anpassung individueller Maßstäbe an übergreifenden Kriterien. Da die persönlichen Maßstäbe für die Bewertung von Leistungen stark von individuellen Ansprüchen und situativen Erfahrungen beeinflusst werden, ist es wichtig, sie von Zeit zu Zeit durch Vergleiche der eigenen Lerngruppe mit repräsentativen Stichproben zu überprüfen.

Als **Korrelation** bezeichnet man in der Statistik den Zusammenhang zwischen zwei Faktoren, z. B.: Je höher die Intelligenz von Schüler/inne/n ist, desto wahrscheinlicher ist es, dass sie im Schulsystem erfolgreich sind. Diese Parallelität darf aber nicht ohne Weiteres kausal gedeutet werden: Es könnte auch sein, dass erfolgreiches Lernen die Intelligenz steigert oder dass ein dritter Faktor wie das familiäre Milieu auf beide Merkmale wirkt.

In **Längsschnittstudien** nehmen dieselben Personen über einen längeren Zeitraum an Erhebungen zu ihrer Entwicklung teil. Veränderungen ihrer fachlichen Leistungen, z. B. von der ersten bis zur vierten Klasse, im Längsschnitt zu untersuchen, ist überzeugender, als nur zwei Querschnitte* zu vergleichen, stellt aber hohe forschungslogistische Anforderungen, z. B. um den Anteil der Dropouts* gering zu halten.

Die »Lernausgangslagen-Untersuchung« (**LAU**) ist eine Längsschnittstudie als Vollerhebung in Hamburg von der 5. bis zur 11. Klasse, in der die Leistungsentwicklung einzelner Schüler/innen erhoben und (anonymisiert) analysiert wurde. Analog angelegt sind weitere Studien wie KESS in Hamburg und ELEMENT in Berlin.

Der **Mittelwert** einer Gruppe wird in der Regel als arithmetisches Mittel (= Durchschnitt) berechnet. Bei einer Normalverteilung der Werte (»Glockenkurve«) ist er zugleich der häufigste Wert (z. B. IQ = 100).

PIRLS/IGLU untersuchte 2001, 2006 und 2011 Fachleistungen am Ende der 4. Klassen und ist für den Bereich Lesen eine Nachfolgestudie zur IEA-Lesestudie von 1991.

In den **PISA**-Studien erhebt die OECD die Leistungen 15-Jähriger (und damit in verschiedenen Teilnehmerländern unterschiedliche Klassenstufen!) im Lesen, in Mathematik und in den Naturwissenschaften. PISA wird seit dem Jahr 2000 im dreijährigen Turnus durchgeführt, in Deutschland (als PISA-E) ergänzt um größere Stichproben einzelner Bundesländer.

Population oder **Grundgesamtheit** bezeichnet alle Personen, über die eine Aussage gemacht werden soll (z. B. alle Viertklässler/innen an deutschen Grundschulen). In der Regel werden die Informationen aus ökonomischen Gründen an einer Stichprobe* gewonnen und aus ihr verallgemeinert.

Anders als in Längsschnitten* wird in **Querschnittstudien** aus dem Vergleich von zwei verschiedenen Gruppen zu demselben Zeitpunkt (z. B. Fünft- und Zehntklässlern im Jahre 2015) auf die Entwicklung einer Gruppe geschlossen, z. B. darauf, wie Leseleistungen in einer Gruppe über die Zeit hinweg zunehmen.

Rohwerte sind die erste Stufe der Datenauswertung (z. B. richtige Antworten oder Fehler in einem Test). Um sie über verschiedene Erhebungen hinweg vergleichbar zu machen, werden sie oft in **Standardwerte** umgerechnet, z. B. bei PISA so, dass sie einen Mittelwert* von 500 und eine Standardabweichung* von 100 haben.

Unterschiede zwischen zwei Gruppen können zufällig (z. B. durch eine unbewusst verzerrte Auswahl der Stichprobe) oder durch eine bestimmte Bedingung oder ein bestimmtes Merkmal verursacht sein (z. B. die soziale Herkunft). Mit **Signifikanz**-Tests prüfen Statistiker, wie hoch das Risiko ist, Unterschiede zu Unrecht auf bestimmte Merkmale oder Bedingungen zurückzuführen. Anders formuliert: Wie sicher ist es, dass eine unterstellte Abhängigkeit tatsächlich besteht? Dasselbe gilt für Zusammenhänge zwischen zwei Faktoren (z. B. dem Einsatz einer bestimmten Lehrmethode und den gemessenen Lernerfolgen).

Die **Standardabweichung** ist ein Maß für die Streuung der Einzelwerte um den Mittelwert*. Bei einer Normalverteilung (Glockenkurve) liegen rund zwei Drittel der Werte innerhalb einer Standardabweichung oberhalb und unterhalb des Mittelwerts.

Stichproben sind Teilgruppen aus einer Population* bzw. Grundgesamtheit, die möglichst nach Zufall gezogen werden. Unter dieser Voraussetzung lässt sich statistisch berechnen, mit welcher Wahrscheinlichkeit und mit welchem Fehlerrisiko Befunde aus der Stichprobe auch in der Grundgesamtheit gelten.

Die »Trends in International Mathematics and Science Study« (**TIMSS**) wurde von der *International Association for the Evaluation of Educational Achievement* 1995 am Ende der Grundschulzeit und am Ende der Sekundarstufe I bzw. II durchgeführt, in anderen Ländern inzwischen auch wiederholt. In der Grundschule hat Deutschland erst 2007 im Rahmen von PIRLS/IGLU teilgenommen.

Seit 2004 werden in den 3. und 8. Klassen aller Bundesländer jährlich flächendeckende Vergleichsarbeiten (**VerA**) in zentralen Fächern geschrieben, deren Ergebnisse an die beteiligten Schulen – mit Daten zu den einzelnen Kindern – zurückgemeldet werden.

In den 50er Jahren fand eine Studie heraus: Kinder aus Familien, die einen Staubsauger besitzen, lesen besser als Kinder aus Familien ohne Staubsauger. Nun käme niemand auf die Idee, Leseschwierigkeiten bei Kindern mit dem Kauf eines Staubsaugers zu behandeln. Ein Fernsehverbot halten diesbezüglich aber viele für vernünftig.

Dieses Buch greift eine Reihe ähnlicher Kurzschlüsse der aktuellen Bildungsforschung auf. Verständlich (»Zahlen sprechen nicht für sich«, »Menschliches Verhalten ist immer mehrdeutig«), anschaulich (»Das Zoom-Dilemma in der Bildungsforschung«) und auf den Punkt (»Statistische Signifikanz heißt nicht inhaltliche Bedeutsamkeit«) erläutert Hans Brügelmann Chancen und Grenzen von Großuntersuchungen mit standardisierten Tests. Dabei warnt er eindringlich vor den Nebenwirkungen einer Evaluation von oben. Seine Forderung: Dialog und Unterstützung statt Autorität und Vorschrift.

Prof. Dr. Hans Brügelmann
lehrte als Erziehungswissenschaftler bis 2012 an der Universität Siegen. Er zählt zu den profiliertesten Kritikern einer einseitig testorientierten Bildungsforschung.

PÄDAGOGIK

www.beltz.de

ISBN 978-3-407-25729-1

9 783407 257291

BELTZ