Von Fall zu Fall[1]

Plädoyer für einen Miss-Marple-Stil in der Bildungsforschung[2]

von Hans Brügelmann (2005i)[3]

Flächendeckende Lernstandserhebungen sind nicht neu. Beispielsweise wurden im Frankreich der 1920er Jahre am Ende der Grundschulzeit standardisierte Leistungsprüfungen durchgeführt. Siebzig Jahre später sind diese Abschlussprüfungen in zwei Regionen von 1925 wiederholt. 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 fünfzehn Disziplinen waren die Leistungen 1995 gleichwertig, in zwei Disziplinen, darunter im Aufsatz, fielen sie 1995 besser aus. Allerdings hatten 1925 die Schülerinnen und Schüler 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 entscheiden. Die scheinbar so klaren Zahlen verdecken nämlich eine Reihe von Störfaktoren. Ein Gedankenexperiment: Wie hätten wohl Schülerinnen und Schüler von 1925 abgeschnitten, wenn sie heute an einer aktuellern Lernstandserhebung teilnehmen müssten:
- in Fächern wie Geometrie, die es damals in der Grundschule noch gar nicht gab;
- mit Themen, die im Lehrplan vor 50 Jahren noch 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?

Ergebnisse der Bildungsforschung haben ein niedriges Verfallsdatum

Über dieses Gedankenexperiment wird deutlich, wie schwierig ein Leistungsvergleich mit den (insofern nur scheinbar) „gleichen" Aufgaben über einen Zeitraum von siebzig Jahren ist. Es kommt hinzu, dass die gesellschaftliche und motivationale Wertigkeit einzelner Inhalte und Fähigkeiten nicht dieselbe ist: Nicht nur, *was* im

---

[1] **Veröffentlicht am 4.10.2005 im „Forum Kritische Pädagogik" Ł http://forum-kritische-paedagogik.de/start/download.php?view.15**

[2] Ich danke Axel Backhaus, Heinrich Bauersfeld, Erika Brinkmann, Georg Lind, Ludwig Stecher, und Jürgen Zinnecker für hilfreiche Kommentare zu einer ersten Fassung dieses Manuskripts.

[3] Hans Brügelmann ist Professor für Erziehungswissenschaft an der Universität Siegen. Zum Thema dieses Beitrags ist im September 2005 sein Buch „Schule verstehen und gestalten" im Libelle-Verlag, CH-Lengwil, erschienen.

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 Jahrhunderts die Rechtschreibung ein Schlüssel zum sozialen Aufstieg, und „fehlerfreie Schönschrift machte aus Bauernkindern kleine Angestellte", wie Fritz-Vannahme 1996 in einem Kommentar in der ZEIT schrieb. Heute sind andere Fähigkeiten wichtiger.

Zahlen sprechen nicht für sich

In den 1950er Jahren erregte folgender Befund einer empirischen Studie einiges Aufsehen: 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-/Rechtschreibschwierigkeiten durch das Verschenken von Staubsaugern an die Eltern der betroffenen Schüler 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 die Statistik, auf die sie sich bezieht. Zahlen sprechen nicht für sich.

Damit wird deutlich: Nicht nur die Alltagserfahrung von Praktikern ist personengebunden, sondern auch die angeblich so „objektive" Bildungsforschung à la PISA. Pointiert gesagt, sind die OECD-Studien nicht deshalb wertvoll, 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 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 gegeneinander gestellt werden, um die scheinbare Expertenautorität zu relativieren.

Empirie ohne Theorie ist blind

Wie wäre es beispielsweise, wenn in einer Studie herauskäme, dass ein hoher TV-Konsum mit niedrigen Leseleistungen oder mit hoher Aggressivität korreliert? Neigen wir nicht dazu, hier eine direkte Kausalität zwischen den beiden Variablen zu unterstellen? Nach dem Staubsaugerbeispiel sollten wir vorsichtig sein. Für das Lesen widersprechen schon die empirischen Daten unseren Vorurteilen. Beispielsweise fand sich in der internationalen Lesestudie der IEA von 1991 kein Beleg dafür, dass Kinder aus Ländern mit einer höheren Fernsehdauer schlechter lasen als Kinder aus Ländern mit geringerer Fern-

sehhäufigkeit. Aus dem PISA-Vorzeigeland Finnland wurden sogar deutlich überdurchschnittliche Fernsehzeiten berichtet.

Hätte man die IEA-Studie schon Anfang der 1960er Jahre durchgeführt, wäre vermutlich 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. Eines ist nicht die Wirkung des anderen, sondern beide sind möglicher Weise die Folge eines dritten Faktors, beispielsweise des sozio-ökonomischen Status der Familie. Sagt unser Hintergrundwissen.

Dieses Hintergrundwissen hilft uns auch, die Widersprüche zwischen Daten aus verschiedenen Zeiten aufzulösen. Fernsehen Anfang der 1960er Jahre *bedeutet* etwas anderes als Fernsehen 1991. Befunde humanwissenschaftlicher Forschung haben ein niedriges 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 1970er Jahren. Diese lassen sich nicht unbesehen auf die heutige Situation übertragen, da sich die gesellschaftlichen Normen, aber auch die tatsächlichen Wahrnehmungs- bzw. Verhaltensmuster erheblich verändert haben – auch als Folge der viel diskutierten Befunde solcher Studien.

Unschärferelation in der Bildungsforschung

Denn in den Humanwissenschaften verändert sich der Gegenstand auch durch die Forschung selbst. Immer wieder entzieht sie durch die Veröffentlichung ihrer Ergebnisse diesen selbst den Boden ihrer Geltung. Analog zur Heisenberg'schen Unschärferelation in der Physik, wenn auch in anderer Form gibt es auch in den Humanwissenschaften einen Beobachtereffekt.

Konkret: Als die 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 – bzw. war – 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 der Verarbeitung dieser Erfahrungen war. Mit dem anschließend propagierten Training des *Verhaltens* wurden aber genau diese Erfahrungen nicht vermittelt.

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 durch eigenaktiven Umgang mit Schrift erworben oder durch Übung 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.

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

Aussagen der Bildungsforschung gelten nur kontextbezogen

Die Lesestudie der IEA 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 fern sahen.

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 auch ganz andere Erklärungen denkbar – und nötig, wenn man an das oben zitierte Beispiel Finnland denkt. In den skandinavischen Ländern werden – anders als etwa in Deutschland und in den USA – viele Filme in der Originalsprache und nur mit Untertiteln in der Nationalsprache gezeigt. Es lohnt sich also für Kinder, lesen zu lernen, um die Filme besser zu verstehen.

Zudem bietet ihnen das Fernsehen eine effektive Lerngelegenheit, wie Laborexperimente gezeigt haben: rasches Lesen unter hoch motivierenden Umständen und gestützt durch einen Bildkontext fördert das Lesenlernen junger 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.

Befunde der Bildungsforschung sind also nicht nur zeitgebunden, sondern auch situationsabhängig. Für verschiedene nationale Kulturen mag dies unmittelbar einleuchten. Kontextabhängigkeit gilt aber auch für Subkulturen wie das soziale Milieu der Familie oder einzelner Klassenzimmer. Was bei Lehrerin A gelingt, kann bei Lehrer B schief gehen. Trotz *Super-Nanny* und ihrer „Erfolge": Unterrichtsmethoden und Didaktik, erst recht aber pädagogische

Konzepte sind keine Techniken, die situationsübergreifend funktionieren.

Zurzeit werden Kindergärten und Grundschulen mit Programmen zum Training der sog. „phonologischen Bewusstheit" überschwemmt. Die Kinder werden darin trainiert, nicht nur auf die Bedeutung von Wörtern, sondern auch auf ihre Lautform zu achten. Reimen, Zerlegen von Wörtern in Einzellaute und deren anschließende Synthese sollen das Lesen- und Schreibenlernen fördern. Verwiesen wird auf Studien zur Vorhersagekraft schlechter phonologischer Testleistungen für spätere Lese-/ Rechtschreibschwierigkeiten und auf positive Evaluationen von Trainingsprogrammen.

Im Durchschnitt decken diese Untersuchungen die Versprechungen der Autoren. Aber der Teufel steckt auch hier im Detail. Viele Studien wurden in den USA durchgeführt. Dort wird 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.

Aber auch die in Deutschland durchgeführte Studien haben nur eine begrenze Aussagekraft. Verschiedene Lehrer arbeiten didaktisch-methodisch sehr unterschiedlich. Immer mehr von ihnen lassen die Kinder eigene Wörter Laut für Laut mit Hilfe einer Anlauttabelle verschriften. Kinder, die nach dieser Methode schreiben und lesen lernen, brauchen kein vorgängiges Phonologie-Training – sie erhalten es sozusagen *on the job*. „Lernen im Gebrauch" nennt die Berliner Professorin Barbara Kochan dieses Phänomen. Was als Vorbereitung eines – heute verpönten – naiv ganzheitlichen Unterrichts durchaus Sinn gemacht hat, ist unter diesen Umständen Zeitvergeudung.

Statistische „Signifikanz" verspricht nicht inhaltliche Bedeutsamkeit

Damit sind wir bei einem zentralen Problem der Bildungsforschung – jedenfalls wie sie von der OECD, von den Max-Planck-Instituten für Psychologie und für Bildungsforschung und von der Deutschen Forschungsgemeinschaft zurzeit favorisiert wird. *Generalisierbarkeit* gilt als das Markenzeichen von Großuntersuchungen wie TIMSS, PISA und IGLU. Aber dieses Konzept ist genau so problematisch wie die Scheinobjektivität der 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. 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 gewonnen werden können. Dabei ist die Verdichtung von Informationen aus einer Stichprobe ist noch relativ einfach: Man erhält Mittelwerte und Korrelationen, und man kann Ergebnisse nach Untergruppen differenzieren. Deren Nutzen für diejenigen, die entscheiden und handeln müssen, hängt aber von zwei anderen Werten ab: der Streuung um den Mittelwert und der Höhe der gewonnenen Korrelationen – nicht nur von ihrer statistischen Signifikanz. Diese zeigt nur, 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 nichts aus.

Selbst in der viel zitierten Schulstudie SCHOLASTIK des Max-Planck-Instituts für Psychologie in München werden Korrelationen von .60 oder gar .30 als *inhaltlich* bedeutsam interpretiert. Dabei signalisieren solche Werte eine außerordentlich lockere Beziehung zwischen zwei Variablen. Sie sind nur wegen der großen Stichprobe „signifikant" - im Sinne von *verlässlich*. Bei Fallanalysen einzelner Lehrer fanden die Forscher denn auch „bizarre Merkmalsprofile" der untersuchten Aspekte 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 9.6.97.

Auch wenn die statistischen Trendwerte – wie in diesem Fall zugunsten einer straffen Unterrichtsführung – recht gering ausfallen: Ihr Sog scheint stärker zu sein als die Zentrifugalkräfte der individuellen Streuungen um diese Mittelwerte herum. Wen wundert es, wenn angesichts solcher Kunstfehler schon in der Forschung auch Medien und Politiker den Kredit statistischer Aussagen oft weit überziehen. Seriöse Statistiker kennen die Grenzen ihrer Kennwerte. Und mit Hilfe von komplexeren Verfahren wie Mehrebenen-Analysen kommen sie zu differenzierteren Ergebnissen. Aber das Kernproblem bleibt.

Mittelwerte helfen wenig bei Fallentscheidungen

Statistische Aussagen 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 im Einzelfall zutreffen, kann man vorher nicht sagen.

Die einzelne Lehrerin aber hat es mit einem konkreten Kind zu tun und sie muss entscheiden, ob ihre Jule bzw. ihr Ben ein typisches Kind ist oder nicht. Unter welchen Bedingungen eine Allgemeinaussage auf einen konkreten Schüler zutrifft – z. B. bei der Übergangsentscheidung Ende 4. Klasse – das sagen uns Studien wie PISA und SCHOLASTIK nicht. Aber gerade für Entscheidungen dieser Art erwarten Pädagogen Hilfe von der Forschung.

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 vorher sieben macht, können sich hinter dieser Verbesserung aber ganz unterschiedliche 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 Problem, wenn sie einzelne Kinder gezielt fördern wollen.

Zu Recht wird kritisiert, wer glaubt, aus Einzelbeobachtungen generalisierbares Wissen ableiten zu können. Dazu braucht man viele Fälle. Diese muss man zudem unter bestimmten Bedingungen (Zufallsstichprobe) untersuchen, um zu Verallgemeinerungen zu kommen. Aber weniger beachtet wird der umgekehrte, mindestens gleich gewichtige Irrtum: Aus statistischem Wissen könne man Fallwissen für praktisches Handeln im Alltag ableiten.

Denn für Pädagogen besteht das eigentliche Problem darin, allgemeine Aussagen – seien es empirische Befunde der Forschung oder normative Vorgaben wie Lehrpläne - auf ihre je 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 Politiker zu wenig Aufmerksamkeit.

Juristen 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 unsere deutschen Gerichte argumentieren mit den Präzendenzfällen 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 Alexander Luria und Oliver Sacks erinnert. Eine Forschung, in der Menschen noch als Personen erkennbar sind. Die Analysen verdeutlichen, dass ver-

schiedene Menschen bei gleichem neurologischen Befund mit ihrem Handicap ganz unterschiedlich umgehen. Einzelne Merkmale determinieren nicht die Entwicklung und das Verhalten einen Person. Menschen sind nicht berechenbar.

Analoges statt induktiv-deduktivem Denken

In Einzelfallstudien kann mehr von dem erhalten bleiben, was man wissen muss, um einen neuen Fall zu verstehen, als im Modelldenken der abstrahierten Variablen. Diese Argumentationsform ist nicht induktiv-deduktiv, sondern beruht auf Analogie, auf einem Denken von „Fall zu Fall". Damit kommt sie den Anforderungen der pädagogischen Praxis, aber auch dem Denken von Praktikern entgegen. Fallberichte sind hilfreich, weil sie analoges Denken ermöglichen: die Deutung einer neuen Situation im Lichte vergangener Fälle.

Ich habe das vor Jahren einmal als das Miss-Marple-Paradigma bezeichnet. Bei Agatha Christie löst Miss Marple die Kriminalfälle im Rückgriff auf ihre *persönliche* Alltagserfahrung mit analogen *Einzel*fällen: „So war das damals auch bei Onkel John, als er...". Forschung unterscheidet sich von Alltagserfahrung, dass sie ihren Annahmen systematisch prüft. Aber die Qualität einer Theorie kann nicht 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.

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 aber wird rasch zur *Diagnose*, was allenfalls ein Mosaikstein bei der *Deutung* von Verhalten sein kann.

Erkenntnis bewährt sich in der Anwendung auf den Einzelfall

Statistische Aussagen sind lediglich *Ausgangspunkte* der Forschung – oder wie es der englische Bildungsforscher Lawrence Stenhouse schon vor über 30 Jahren formuliert hat: Nicht die Fallstudien erzeugen Hypothesen für Großstudien, sondern umgekehrt liefern die Repräsentativerhebungen Folien für die Analyse von Einzelfällen. Sie untersuchen *wenige* Merkmale an *vielen* Fällen als Einstieg in die sorgfältige Untersuchung *vieler* Merkmale des *Einzelfalls*. Ergebnisse der Forschung über die „Wirksamkeit" von pädagogischen Maßnahmen könne nur deren *Potenzial* bzw. ihre *Risiken* und wahrscheinliche Bedingungen für ihre Realisierung benennen. Allgemeinaussagen zu ihrer Qualität stehen immer unter kontextabhängigen Vorbehalten. Am deutschsprachigen PISA-Siegerland Südtirol im Verlierersystem Italien (vgl.

Leitzgen 2005; Meraner 2005) lässt sich der Einfluss spezifischer sozio-kultureller und pädagogischer Bedingungen eindrucksvoll studieren – auch im Vergleich mit den deutschen Bundesländern.

Halten wir fest:

- Befunde in den Humanwissenschaften sind in hohem Maße situationsabhängig. Die Bedeutung eines Faktors verändert sich von Fall zu Fall durch den Einfluss anderer Faktoren (Interaktionseffekte).
- Auch Untersuchungen mit standardisierten Verfahren sind angewiesen auf die Interpretationsfähigkeiten ihrer Autoren. Zahlen sprechen nicht für sich.
- Die Verallgemeinerung von Befunden sichert keine Übertragbarkeit auf neue Situationen. Nicht das arithmetische Mittel, sondern die Streuung der Einzelwerte um den Durchschnitt ist das praxisrelevante Datum.

Damit ist deutlich, dass die „Anwendung" von Allgemeinaussagen auf neue Fälle kein technischer Vorgang ist. Sie bedarf eigener Forschungskompetenz, nämlich der Fähigkeit zur Erkundung relevanter Bedingungen für die Übertragbarkeit von Befunden und Modellen.

Was aber tut die Bildungsforschung, was tut die Bildungspolitik, um diesen zentralen Schritt der Erkenntnisgewinnung zu unterstützen? *Hier* müssen Ressourcen konzentriert werden, wenn sich die Qualität von Unterricht verbessern soll.

Studien wie PISA sind hilfreich zur Evaluation des Bildungssystems. Dafür aber reichen Stichproben. Aufwändige flächendeckende Lernstandserhebungen wie VERA sind dafür 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. So haben sich in den USA trotz einer Vielfalt von Interventionen die Testergebnisse des National Assessment of Educational Progress über mehr als 30 Jahre hinweg kaum verändert.

Einen Nachholbedarf aber haben Forschung und Ausbildung im Bereich kontextbezogener Fallanalysen, die mehrere Perspektiven in die Deutung der Daten einbeziehen.

Fehlende Offenheit für Ereignisse „wider Erwarten"

In der ZEIT v. 5.1.2005 hat Harro Albrecht die medizinische Forschung mit 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 vom Geschlecht, vom Alter, vom Gewicht, vom Blutdruck, von der Ernäh-

rung, vom aktuellen Stress? Nur die alten Landärzte und einige wenige Hausärzte, die über viel Erfahrung verfügen, die sich Zeit nehmen und die ihre Patienten persönlich kennen, 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.

„Reine Normen" haben auch in anderen Bereichen unserer Gesellschaft Konjunktur, zurzeit eben besonders im Schulwesen. „Regelstandards", die von allen Schülern dieselbe Leistung zu demselben Termin fordern, und Kompetenztests, die ihr Erreichen landesweit abprüfen, 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 „nach PISA" eher wieder zunehmen wird.?

Dass Bildung, dass Gesundheit, dass Gerechtigkeit etwas mit individuellen Besonderheiten, mit persönlichen Beziehungen zu tun hat, gerät leicht in Vergessenheit. Standardisierte Behandlung des vermessenen Menschen – eine Sackgasse nicht nur in der Medizin. Früher lernten gute Praktiker aus der sorgfältigen Analyse von kumulierten Einzelfällen. Heute gilt vielen als Rechtfertigung für eine Maßnahme, wenn sie 75% der „Bedürftigen" hilft. Was sie für die anderen 25% 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.

Soziale Kontrolle durch Mehrperspektivität statt Standardisierung durch methodische Präzision

Nun bilden Großstudien wie PISA bei der Auswertung ihrer Daten durchaus Untergruppen. Leistungen werden beispielsweise nach sozialer Schicht und dem sog. „Migrationshintergrund" unterschieden Aber diese Differenzierungen werden angebunden an Oberflächenmerkmale, die „objektiv" erfasst werden können, beispielsweise an den Schulabschluss und die berufliche Position der Eltern. Vernachlässigt wird die immer noch immense Streuung innerhalb auch dieser Untergruppen. Selbst die gern zitierte Zahl der Bücher und die Häufigkeiten des Vorlesens im Vorschulalter erfassen nicht die *Qualität* solcher Aktivitäten: Intensivere Fallstudien zeigen, dass es auf die *Art* des Vorlesens ankommt. Deren Wirkung wiederum hängt aber 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.

Vor 25 Jahren hat das Max-Planck-Institut für Bildungsforschung eine Studie über den Zustand der Grundschulen in Deutschland veröffentlicht. 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 politische Entscheidungen. Zur Verbesserung des pädagogischen Alltags tragen sie wenig bei.

Allerdings sind viele besorgt, dass interpretative Fallstudien einen Rückfall in die Beliebigkeit subjektiver Deutungen bedeuten. Das Problem besteht in der Tat. Durch 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. Ausweisung getrennter Rollen im Verfahren, Besetzung des Gerichts mit mehreren Richtern, Einrichtung von Instanzenzügen sind Beispiele für solche *checks and balances*: Wo Objektivität nicht möglich ist, können die Auswirkungen von Subjektivität kontrolliert oder zumindest transparent gemacht werden. Forschung ist auch ein politisches Geschäft.

Der US-amerikanische Bildungsforscher David C. Berliner bezeichnet Bildungsforschung als *„the hardest science of all"*. Bei der Übersetzung angelsächsischer Forschungsansätze ins Deutsche ist die tiefere Bedeutung dieses Satzes wohl leider verloren gegangen. Wir brauchen *auch* Studien, die dem Standardisierungs-Paradigma 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 sog. „quantitativen" Paradigmas.

Repräsentative Stichprobenerhebungen und Fallanalysen, standardisierte Methoden und interpretative Verfahren können und müssen sich ergänzen. Bildungspolitik und Forschungsförderung bevorzugen zurzeit allerdings einseitig *ein* Paradigma. Damit geht die Komplementarität verschiedener Zugänge verloren. Deren Notwendigkeit wurde bereits früher, zuletzt in den 1970er Jahren sehr 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 des Menschen nicht.

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. Rororo 7338: Frankfurt.
Berliner, D. C. (2002): Educational research: The hardest science of all. In: Educational Researcher, Vol. 31, No. 8, 18-20.
Brügelmann, H. (2005): Schule verstehen und gestalten – Perspektiven der Forschung auf Probleme von Erziehung und Unterricht. Libelle: CH-Lengwil.
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.
Hopf, D., u. a. (1980): Aktuelle Probleme der Grundschule. In: Baumert u. a. (1980, Bd. 2, 1113-1176).
Leitzgen, A. (2005): Neues aus PISA. In: Family & Co, H. 10/2005 v. 15.9.2005.
Meraner, R. (2005): Spitze bei PISA. Die Ergebnisse und erste Überlegungen. In: Info (Informationsschrift für Kindergarten und Schule in Südtirol), H. 1 (Jänner)/2005, 12-16.
Sacks, O. (1991): Einführung zu A.R. Lurija "Der Mann, dessen Welt in Scherben ging". Rowohlt: Reinbek (S. 7-20).
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, 447-452).