

Rauin, Udo

## Die Pädagogik im Bann empirischer Mythen. Wie aus empirischen Vermutungen scheinbare pädagogische Gewissheit wird

*Pädagogische Korrespondenz (2004) 32, S. 39-49*



Empfohlene Zitierung/ Suggested Citation:

Rauin, Udo: Die Pädagogik im Bann empirischer Mythen. Wie aus empirischen Vermutungen scheinbare pädagogische Gewissheit wird - In: Pädagogische Korrespondenz (2004) 32, S. 39-49 -  
URN: urn:nbn:de:0111-opus-80482

<http://nbn-resolving.de/urn:nbn:de:0111-opus-80482>

in Kooperation mit / in cooperation with:



<https://pk.budrich-journals.de>

### 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](mailto:pedocs@dipf.de)  
Internet: [www.pedocs.de](http://www.pedocs.de)

Digitalisiert

Mitglied der

  
Leibniz-Gemeinschaft

## **Das aktuelle Thema**

- 5 *Andreas Gruschka*  
Empirische Bildungsforschung – das muss keineswegs, aber es kann die Erforschung von Bildungsprozessen bedeuten.  
Oder: Was lässt sich zukünftig von der forschenden Pädagogik erwarten?

## **Dokumentation 1**

- 36 Ein Gütesiegel für Spitzenleistungen

## **Dokumentation 2**

- 38 Interview mit Prof. Ernst Mayr

## **Aus Wissenschaft und Forschung**

- 39 *Udo Rauin*  
Die Pädagogik im Bann empirischer Mythen –  
Wie aus empirischen Vermutungen scheinbare pädagogische Gewissheit wird

## **Sinnbilder**

- 50 *Bertram Ritter*  
Ansatz zu einer kunstsoziologischen Werkanalyse des Gemäldes  
»Komposition im Quadrat« (1922) von Piet Mondrian

## **Diskussion 1**

- 69 *Stefan Blankertz*  
Vier Fäuste für ein Halleluja  
*Demokratie oder Schule?*

## **Dokumentation 3**

- 73 Interview mit dem Philosophen Volker Gerhardt

## **Diskussion 2**

- 74 *Ulrich Oevermann*  
Über den Stellenwert der gesetzlichen Schulpflicht –  
Antwort auf meine Kritiker

## **Diskussion 3**

- 85 *Gerold Scholz*  
Bildungsrecht für Kinder

## **Dokumentation 4**

- 91 Ticket recruitment

*Udo Rauin*

## Die Pädagogik im Bann empirischer Mythen – Wie aus empirischen Vermutungen scheinbare pädagogische Gewissheit wird

Die empirische Pädagogik und die pädagogische Psychologie erfahren nach TIMSS, PISA und IGLU eine starke Aufwertung in der öffentlichen Wahrnehmung. Meinungsmacher in der Fachpresse und in der Politik erwarten, dass die so genannte empirische Bildungsforschung sehr rasch Rezepte für ein effektiveres Erziehungssystem entwickeln könnte. Andere Disziplinen der Erziehungswissenschaft, die bislang weniger auf harte Empirie setzten, stehen wie gebannt vor den wild ins Kraut schießenden Zahlenbergen und komplexen Grafiken und hoffen darauf, ihr Erfahrungswissen und ihre Theorien würden vom Glanze der sakrosankten Daten profitieren. Dieser Boom empirischer Bildungsforschung ist weniger dem erreichten Forschungsniveau, den bislang entwickelten und überprüften Theorien und Befunden oder gar dem Beitrag dieser Forschung zur Steigerung der Schul- und Unterrichtsqualität geschuldet. Er ist vielmehr ein recht großzügiger Kredit, der von der Hoffnung auf die technologische Verwertbarkeit einiger respektabler Erkenntnisse lebt. Sicherlich ist es unangemessen im Stile des Positivismusstreits der empirischen Forschung, ihre Existenzberechtigung zu bestreiten. Aber die massive Überschätzung vorläufiger Befunde oder die Häufigkeit waghalsiger Schlussfolgerungen aus den Daten empirischer Forschung, die durch die methodische Unbefangenheit mancher Produzenten und vieler Nutzer empirischer Studien verursacht wird, fordern zu einer kritischen Betrachtung heraus.

Die oft zitierten internationalen empirischen Schulleistungsstudien sind keine hypothesenprüfenden Untersuchungen zur Schul- oder Unterrichtsqualität, wie manchmal irreführend angenommen wird, sondern sie schätzen Parameter (z.B. die durchschnittliche Lesekompetenz) einer Altersgruppe oder einer Klassenstufe im Vergleich zwischen Regionen, Ländern oder auch Schulformen. Derartige Messungen sind sicherlich verdienstvoll und sie liefern wichtige Informationen für die Einschätzung der Situationen in Schulen und in Regionen, aber sie erklären nicht, auf welche Ursachen die beobachtbaren Unterschiede zurückgeführt werden können. Ausgereifte und erprobte Modelle zur Vorhersage von derartigen Parametern müssten Sozialdaten, Schülermerkmale, Unterrichtsbedingungen, Merkmale der einzelnen Schule und der jeweiligen Schulstruktur in schlüssiger Weise aufeinander beziehen können. Solche komplexen Mehrebenenmodelle liegen allenfalls in Ansätzen vor (vgl. z.B. Wang, Haertel und Walberg 1993), sie haben aber weder national noch international

irgendeine Bewährung erfahren (vgl. z. B. Ditton 2002). Das Problem der Verbindung mehrerer Ebenen (Schüler, Klasse, Schule, Schulstruktur) ist nicht allein wegen der riesigen Anzahl an Variablen und der nur sehr lückenhaften Datenbasis so schwierig zu lösen, sondern auch wegen der noch weitgehend unbekanntem Wechselwirkung zwischen den Ebenen. Häufiger sind reduzierte Modelle zur Erklärung von Leistungsparametern anzutreffen, die sich lediglich mit der Kontrolle weniger Faktoren einer Ebene beschäftigen. Untersucht werden individuelles Lernverhalten in experimentellen Trainingsstudien, Unterrichtsmerkmale (z. B. das so genannte Klassenklima), die Einstellung von Lehrkräften (fachliche oder pädagogische »beliefs«) und Lehrerverhalten. Die Untersuchung des Vorhandenseins und der Wirkung von Unterrichtsmerkmalen ohne Berücksichtigung komplexer Wechselwirkungen zwischen Individuen, Einzelschule und Schulstruktur ist in der empirischen Unterrichtsforschung durchaus üblich und z. B. für die Entwicklung von fachdidaktischen Theorien sicherlich auch nützlich. Aber selbst die im Vergleich zu Mehrebenenmodellen bescheideneren Ansprüche lokaler Theorien, wie sie z. B. die DFG-Projekte zur Unterrichtsqualität im naturwissenschaftlichen Unterricht oder im Mathematikunterricht (BIQUA) verfolgen, sind noch weit entfernt von irgendeiner gehaltvollen Erklärung und Vorhersage von Schülerleistungen oder gar von der Produktion technologisch verwertbaren Wissens. Das mag angesichts der hochfliegenden Programme und der großen Forschungsmittel, die der empirischen Unterrichtsforschung vor allem in den USA zur Verfügung stehen, verwundern. Wenn man aber einen Blick auf das Instrumentarium dieser Forschung wirft, dann erkennt man einige basale Mängel, die auch auf die europäische Forschung Auswirkungen haben. Ein zentrales Problem der Unterrichtsforschung liegt in der unzureichenden Entwicklung von Theorien des Unterrichts, die den unterschiedlichen Unterrichtskulturen der Fächer und Schulstufen gerecht werden, ein anderes Problem im Fehlen valider Instrumente der Unterrichtsbeobachtung.

An einem besonders oft zitierten Beispiel der Unterrichtsforschung möchte ich das Missverhältnis von empirischem Gehalt und überzogener Rezeption durch die Fachöffentlichkeit diskutieren. Der frühere Direktor des MPI für Psychologie in München Franz E. Weinert und der heute in Landau lehrende Psychologe Andreas Helmke haben unter dem Titel »Entwicklung im Grundschulalter« 1997 einen Band herausgegeben, der eine gemeinsam entwickelte Studie vorstellt. Dieser Band wird in der Grundschulpädagogik breit diskutiert. Wolfgang Einsiedler (1997, S. 309) schreibt in einem Überblicksreferat über die empirische Grundschulforschung: »Das grundschulbezogene SCHOLASTIK-Projekt [...] könnte für die Grundschulpädagogik eine Anregung sein, die dort behandelten Problemstellungen weiter zu bearbeiten.« Edith Glumpler (1998, S. 55) erwartet »wichtige Impulse für die zukünftige bundesdeutsche Grundschulforschung«. Hans Brügelmann (1998, S. 167) empfindet die Lektüre der Beiträge als »Herausforderung, zudem anregend und ertragreich, auch unmittelbar für die Schulpraxis«. Nur Renate Valtin (2000, S. 561) bleibt skeptisch, wenn sie feststellt: »Die Grundschul-ExpertInnen sind zwar höchst angetan von der SCHOLASTIK-Studie, doch ziehen sie keinerlei konkrete und praktische Schlussfolgerungen und lassen uns bei der Rezeption dieser Studie etwas ratlos zurück.«

Gabriele Faust wagt in einem Thesenpapier (Faust 2002) sogar einen prophetischen Blick: »Das am Ende der Pflichtschulzeit erreichte Lernniveau baut auf den Lernergebnissen der vorherigen Stufen auf. In verschiedenen Studien, z.B. LOGIK und SCHOLASTIK, haben sich die jeweiligen Vorkenntnisse als erklärungskräftigster Pfad für nachfolgende Leistungen erwiesen. Mathematische Kompetenzen am Beginn der Kindergartenzeit korrelieren relativ hoch mit Mathematikleistungen am Ende von Klasse 2 und diese wiederum mit den Testergebnissen am Ende von Klasse 5 (Korrelationen .63 und .61). Es ist darum berechtigt, besorgt von PISA den Blick auf den deutschen Elementar- und Primarbereich zu richten, auch wenn den Ergebnissen von PIRLS/IGLU im Mai 2003 nicht vorgegriffen werden darf.« Wenn das zuträfe, dann wäre das schlechte Abschneiden deutscher Schüler in PISA in erster Linie ein Versäumnis der Elementar- und Primärerziehung, dann müssten sich Kindergarten und Grundschule mehr um die Propädeutik des fachlichen Lernens kümmern. Dass Frau Faust hier eine Grundregel der empirischen Forschung verletzt – Korrelationen dürfen nicht kausal interpretiert werden, denn es könnte sich z.B. auch um eine Scheinkorrelation handeln, die durch den Einfluss einer dritten Größe zustande kommt – will ich nicht vertiefen. Inzwischen hat IGLU ein unerwartet freundliches Bild der deutschen Grundschulen gezeichnet und Fausts Befürchtungen nicht bestätigt. Aber auch diese neueren Parameterschätzungen geben zu vielen weiteren, wenig begründeten Spekulationen Anlass, auf die ich hier nicht eingehen will. Erstaunlich erscheint mir vielmehr, wie sich die Zunft der Pädagogen kollektiv von Zahlen und Bildern blenden lässt.

Einsiedler (1997) zitiert in ähnlicher Weise zustimmend ein beeindruckendes Modell (Abbildung VII.4 in Weinert/Helmke 1997) zur Erklärung der Schülerleistungen in Mathematik aus einem Projekt der Münchener Forschergruppe. Danach ließe sich ein großer Teil der Leistungsentwicklung durch wenige Variablen, nämlich durch die »Vorkenntnisse« und einige Merkmale der Unterrichtsgestaltung aufklären. Mit diesem Modell will ich mich auseinander setzen, weil es die Mystik der Zahlen gut demonstriert.

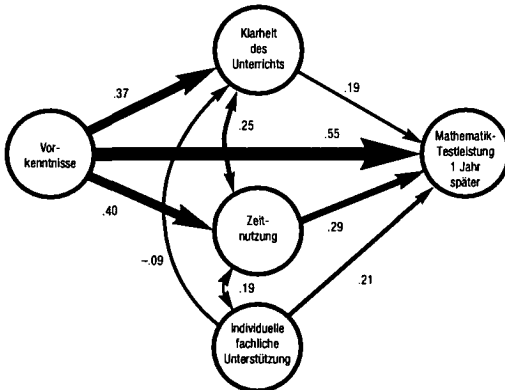


Abb.: Zusammenhänge zwischen Vorkenntnissen der Schüler, Unterrichtsvariablen und Unterrichtsergebnissen (Weinert, Schrader & Helmke, 1990)

Das in der Grafik dargestellte Strukturgleichungsmodell ist in den vorliegenden Veröffentlichungen nicht vollständig dokumentiert. Ich kann deshalb auf statistische Details dieser Modellierung nicht eingehen. Der Leser sollte aber wissen, dass man mit Hilfe von Strukturgleichungsmodellen (z. B. mit den bekannten Programmen EQS oder LISREL oder AMOS) aus dem gleichen Datensatz völlig unterschiedliche Lösungen, ja sogar gegensätzliche Lösungen entwickeln kann. Die hier eingezeichneten kausalen Pfade sind zunächst einmal nur Hypothesen. Man bezeichnet solche ersten Lösungen deshalb auch als explorativ. Bestätigungen einer Theorie sind nur dann anzunehmen, wenn ein solches Modell mit unterschiedlichen Datensätzen immer wieder zu vergleichbaren Lösungen führt. Das haben die Autoren zwar möglicherweise versucht, es ist ihnen aber offenbar nicht gelungen, auch nicht mit weiteren Erhebungen im Rahmen der SCHOLASTIK-Studie, denn sie greifen auf dieses Modell in späteren Studien und in der Veröffentlichung aus dem Jahr 1996 nicht mehr zurück. Man mag das für einen Einzelfall halten, aber jeder Leser kann sich in den einschlägigen Zeitschriften der empirischen Pädagogik selbst ein Bild davon machen. Es gibt keinen Mangel an immer wieder neuen und noch komplexeren Modellen, aber es gibt einen eklatanten Mangel an Replikationsstudien zum gleichen Modell. Explorative Studien sind die Regel, aber hypothesenprüfende, explanative Untersuchungen kommen in der Unterrichtsforschung und in vielen anderen Bereichen der empirischen Pädagogik kaum vor.

Einsiedler zitiert diese explorative Modellierung von Schrader und Helmke so, als handele es sich um eine bestätigte Theorie. Das geschieht übrigens in der Erziehungswissenschaft häufiger und belegt die mangelnde Kenntnis empirischer Forschungsmethoden in weiten Teilen dieser Zunft. Aber das ist nicht mein zentrales Argument. Im Folgenden möchte ich danach fragen, ob die in diesem Modell verarbeiteten Beobachtungsdaten den Unterricht in Grundschulen überhaupt zutreffend beschreiben können. Nach meiner Auffassung sind die meisten Instrumente der Unterrichtsbeobachtung nicht ausreichend entwickelt und erprobt worden. Sie kamen zum Einsatz, bevor man sie auf ihre Leistungsfähigkeit überprüfte. Wenn die Instrumente der Unterrichtsbeobachtung aber wenig taugen, dann wären die sich darauf stützenden Befunde Artefakte und Zahlenspielerien.

Seit dem berühmten Aufsatz von Brophy und Good (1986) im »Handbook of Research on Teaching« sucht die empirische Unterrichtsforschung nach einem Beleg für die These, dass man mit Hilfe von einigen beobachtbaren Prozessmerkmalen des Lehrerhandelns den Lernfortschritt von Schülern gut vorhersagen könne. Die meisten Studien zum Beleg dieser These aus den Bereichen der »Prozess-Produkt-Forschung« und »Lehrereffektivitätsforschung« wurden in Elementarschulen durchgeführt. Bei den überprüften Lerneffekten handelte es sich meistens um basale Kenntnisse und Fertigkeiten. Ein Ziel der Forschungen war die Identifikation von Mustern effektiven Lehrerverhaltens oder so genannter »teaching functions« (vgl. Rosenshine/Stevens 1986). Beispiele für als effektiv ermitteltes Lehrerverhalten sind: Klarheit des Ausdrucks, methodisch-didaktische Variabilität, Enthusiasmus und Kritik, die Verwendung indirekter Lehrstrategien, das Schaffen von Lerngelegenheiten, die

Strukturierung des Unterrichts und die Variation des Anspruchsniveaus bei Fragen und Darbietungen (vgl. Rosenshine/Furst 1973, S. 155f.). Beispiele für teaching functions sind Hausaufgabenkontrolle, strukturierte Präsentation, gezielte Korrekturen und gezieltes Feedback. Die komplexe und nicht immer eindeutige Forschungslage wurde in praxisorientierter Auswertung auf eine überschaubare Anzahl von Ratschlägen reduziert: »Beginne eine Stunde mit einer kurzen Wiederholung und einer kurzen Zielangabe. Präsentiere Neues in kleinen Schritten und führe zu jedem Schritt Übungen durch«. Helmke knüpft in seinem vor kurzem erschienenen Lehrbuch »Unterrichtsqualität« noch einmal an diese Tradition an, weil die Ideen der 80er Jahre nach seiner Auffassung noch immer »state of the art« seien (Helmke 2003, S. 62). So anregend manche Ratschläge für Studierende in den Anfangssemestern des Lehramtsstudiums auch sein mögen, die meisten Botschaften bewegen sich auf einem recht trivialen Niveau (z.B. »Giving information: gesichert ist, dass Einzel- und Stillarbeit erst auf einer inhaltlich soliden Basis erfolgen sollen.« (Helmke 2003, S. 63)) und plädieren für eine methodische Monokultur des lehrerzentrierten Unterrichts.

Interessanterweise decken sich diese Ratschläge weitgehend mit jener Methode, die Bereiter und Engelmann (1966) zur Förderung lernschwacher Schüler in Vorschulen beschrieben hatten und die unter dem Etikett »Direct Instruction« in den USA zu einer regelrechten pädagogischen Bewegung führten. Tatsächlich legen Vergleichsstudien zwischen unterschiedlichen Unterrichtsverfahren die Überlegenheit der Direkten Instruktion im Hinblick auf einige Leistungsbereiche nahe (vgl. Lüders/Rauin 2004). Möglicherweise hat dies die Forscher des MPI dazu bewegt, in ihrer Studie auf Konstrukte aus dieser amerikanischen Tradition des »effektiven Lehrens« zurückzugreifen. Sie unterstellen dabei, dass die Unterrichtspraxis an bayerischen Grundschulen in vielen Merkmalen mit der Praxis amerikanischer Elementarschulen übereinstimmt, die nach dem Modell der Direkten Instruktion unterrichten. Ihre Anlehnung an die amerikanische Forschung geht sogar soweit, dass sie nicht nur die implizite Annahme der Überlegenheit des lehrergesteuerten Unterrichts, sondern auch die auf diesen Unterricht zugeschnittenen Beobachtungsinventare übernehmen (vgl. Weinert/Helmke 1997, S. 510ff, die in ihrer Studie die Instrumente von Anderson, Ryan und Shapiro adaptieren).

Bei der Studie von Weinert und Helmke handelt sich um eine quantitative Beobachtungsstudie (Lehrerhandeln), die mit einer Veränderungsmessung (Schülerleistung) verknüpft werden soll. Jede systematische Beobachtung muss das Problem der Selektion, der Abstraktion und der Klassifikation lösen (vgl. Bortz/Döring 2002).

»Selektion« meint die Auswahl bestimmter Beobachtungsgegenstände bzw. das Herausfiltern bestimmter Reize aus der Vielzahl gleichzeitig wahrnehmbarer Reize. Dies geschieht – wenn auch nicht sehr systematisch – anhand eines Konstrukts aus der amerikanischen Unterrichtsforschung. Der Unterricht der bayerischen Lehrkräfte wird also daraufhin beobachtet, ob er Merkmale zeigt, die sich in der Forschung zur Direkten Instruktion (als beim Lehren einfacher Regeln oder von begrifflichem Wissen) als erfolgreich erwiesen haben. Eine solche Selektion der Modellvariablen ist

keineswegs ohne Alternative. Besonders dann, wenn man indirekte Unterrichtsstrategien (z.B. Problemlösen) beobachten wollte, hätten sich andere Selektionsentscheidungen aufgedrängt. Jede Selektionsentscheidung schließt notwendigerweise andere – nicht selektierte Unterrichtsaktivitäten – aus dem Beobachtungsprozess aus. Typische Merkmale eines effektiven Unterrichts sind nach der Lesart der Direkten Instruktion z.B. Zeitnutzung, Strukturiertheit, Variabilität der Unterrichtsformen, Klarheit und »engagement in the learning process« (bei Helmke/Weinert übersetzt mit »Motivierungsqualität«). Diese Merkmale sollen im Unterricht beobachtet und quantifiziert werden und schließlich will man damit Leistungssteigerungen von der



3. zur 4. Klassenstufe in Mathematik und Rechtschreiben erklären. Nicht beobachtet werden dagegen z.B. die Förderung von Schülerdiskussionen, die Auswahl des Lehrmaterials oder die Art der gestellten Aufgaben.

Das Abstrahieren in quantitativen Beobachtungsstudien besteht darin, ein Ereignis aus seinem jeweiligen konkreten Umfeld herauszulösen. Das Ereignis wird auf seine wesentliche Bedeutung für die Untersuchungsfrage reduziert. Dieser Vorgang wird sichtbar, wenn man die Operationalisierungen hinsichtlich des Beobachtungsinstrumentes nachvollzieht. In der Regel werden Indikatoren gebildet, mit denen das Vorhandensein eines Merkmals verknüpft werden kann. Die Lösung dieser Probleme in der SCHOLASTIK-Studie ist überraschend schlicht. Ich will das am Beispiel der Variablen Strukturierung zeigen. Die Definition von Strukturierung (Weinert/Helmke 1997, S. 512) ist sogar im Vergleich zur amerikanischen Diskussion sehr eng: »Prägnanz der Ausdrucksweise des Lehrers; Strukturierung des Lernstoffs durch aufmerksamkeitsregulierende Hinweise und andere Hinweise.« Für dieses Konstrukt werden zwei Items entwickelt:

(pg1) Prägnanz: »Kurze, direkte und prägnante Ausdrucksweise.«

(hv1) Hervorhebungen (Cues): »Betonung der Wichtigkeit bestimmter Ziel- oder Stoffaspekte; aufmerksamkeitsregulierende Bemerkungen; explizite Herstellung von Zusammenhängen zwischen verschiedenen Teilen des Stoffs und von Bezügen zu Zielen.« Prägnanz ist ein Merkmal der Sprache und war bei Good & Brophy (1994) oder anderen Autoren deshalb mit der Skala Klarheit der Sprache und nicht mit Strukturierung verknüpft. Es bleibt deshalb eigentlich nur ein Item (Hinweise), um das gesamte Spektrum von didaktischen Möglichkeiten der Strukturierung zu erfassen.

Damit wird das dritte Problem jeder systematischen Beobachtungsstudie sichtbar: Die Einzelbeobachtungen müssen sich eindeutig klassifizieren lassen. Als Klassifikation bezeichnet man die Bildung von Ereignis- oder Merkmalsklassen und den Vorgang der Zuordnung von Zeichen und Symbolen zu diesen Klassen. In unserem Beispiel sind Handlungsweisen der Lehrkräfte zu bestimmen, die das Merkmal »Strukturierung« bezeichnen, klassifizieren oder sogar graduieren sollen. Wenn man verschiedene Arten von Strukturierung erfassen wollte, dann wären unterschiedliche Aspekte von Strukturierung zu klassifizieren (vgl. dazu Good 1996). Rosenshine und Meister (1992) haben »Vorher-« und »Nachher-Strukturierungen«, »Hervorhebungen (Cues)« und »Signal-Strukturierungen« unterschieden. Ein typisches Beispiel für eine Vorherstrukturierung sind »Advance Organizer« (Ausubel), also Hinweise, die erklären, was in der nächsten Unterrichtsphase wichtig ist und gelernt werden soll. Indikatoren für Nachher-Strukturierungen sind z.B. die Verwendung von Merksätzen, Zusammenfassungen oder ähnlichem. In welcher Unterrichtsphase und in welchem Umfang Strukturierungen hilfreich für den Lernprozess sind, lässt sich aber kaum kontextfrei bestimmen. Noch komplexer wird die Fragestellung, wenn es nicht nur um eine Rangordnung der Beobachtungen (»Der Unterricht ist mehr oder weniger strukturiert!«), sondern sogar um intervallskalierte Ausprägungsgrade geht (»Der Unterricht ist 2 Punkte mehr strukturiert!«). Man könnte – wie bei Good & Brophy

(1994) demonstriert – das Vorhandensein erfolgreicher Strukturierungen in einer bestimmten Zeiteinheit auswerten. Das würde aber einen erheblichen Beobachtungsaufwand mit sich bringen und ist in Live-Beobachtungen kaum realisierbar. Das soll genügen, um zu zeigen, wie man sich in der Unterrichtsforschung dem Problem der Abstraktion und Klassifikation von Strukturierung genähert hat und welche Probleme dabei zu beachten sind.

Die Münchener Studie löst auch dieses Problem überraschend unbefangen. Anstelle von inhaltlich bestimmten Kategorien der Klassifikation von Unterrichtsmerkmalen wird für alle Items eine einheitliche fünfstufige Rating-Skala gebildet.

»Antwortkategorien: trifft vollkommen zu / trifft ziemlich zu / trifft mittelmäßig zu / trifft weniger zu / trifft gar nicht zu«

Am Ende jeder Stunde werden die Beobachter aufgefordert, summarisch über jeden Einzelaspekt ein Urteil zu fällen. Die Beobachter erhalten keine weitere Hilfe, auf welche Weise sie die fünf Antwortkategorien zu deuten haben. Die Antwortvorgaben passen häufig nicht zu den Items. Wann trifft z.B. »Hervorhebung« vollkommen oder wann nur mittelmäßig zu?

Schlichte Beobachterratings dieser Art führen zwar zu vielen Daten, aber notwendigerweise auch zu sehr vagen Einschätzungen und werden deshalb als hoch interpretativ oder hoch inferent bezeichnet. Die Messgenauigkeit der Daten ist notwendigerweise gering. Ein entsprechendes Maß für die Reliabilität – etwa die Beobachterübereinstimmung – wurde in der Studie noch nicht einmal publiziert. Die unterschiedlichen Konstrukte (Skalen) für den Mathematikunterricht korrelieren überraschenderweise recht hoch miteinander. Man könnte das positiv lesen und daraus folgern, dass »gute« Lehrkräfte in allen Dimensionen »gut« waren und »schlechte« durchgängig schlecht. Wer gut strukturieren kann, der nutzt auch die Unterrichtszeit optimal. In einer kritischen Lesart deutet der hohe Zusammenhang aber darauf hin, dass nicht differentielle Merkmale der Unterrichtsqualität, sondern ein globaler Eindruck der Beobachter in den Daten sichtbar wird.

Das gesamte Vorgehen bei der Entwicklung des Instruments ist unbefriedigend und das trifft beinahe auf alle Skalen der Münchener Studie zu, die das Lehrerverhalten beschreiben. Klare Indikatoren für Unterrichtsmerkmale und für die Graduierung der Merkmale liegen nicht vor. Solche Instrumente müssten mit den im Untersuchungsfeld üblichen Handlungsrouninen abgeglichen werden. Man hätte erheben können, wie Lehrer Neues einführen, Arbeitsschritte gliedern, Interpunktieren, die Bearbeitung von Aufgaben zeitlich sequenzieren. Aber darauf wurde verzichtet, weil so etwas im Konzept der adaptierten Instrumente nicht vorkommt. Man hätte bereits bei der Entwicklung des Instruments Fachleute aus der Grundschule oder aus dem Bereich der Ausbildung zu Rate ziehen können, um sicher zu stellen, dass für das Handlungsfeld Relevantes abgebildet wird. Das ist zeitaufwändig und mühsam, aber unverzichtbar. Viele Nachfolgerstudien (z.B. die zurzeit in Mode gekommenen Videostudien) haben aus diesen Fehlern gelernt. Den Nachweis, ob sie näher an die

Erfassung von Unterrichtsqualität herankommen, müssen sie aber noch leisten. Bislang liegen auch hier nur explorative Studien vor.

Weil in der Münchener Grundschulstudie auf eine sorgfältige Entwicklung der Instrumente zur Erfassung der Unterrichtsqualität verzichtet wurde, sind die Daten ziemlich wertlos. Die Qualität der Daten ist für den unbefangenen Leser auch daran erkennbar, dass sie ein ziemlich inkonsistentes Bild ergeben. Auch hier soll ein Beispiel genügen. Zur Interpretation der Daten wird auf ein Paradigma aus den 60er Jahren zurückgegriffen. Gage (1967) hatte in seinem berühmten Beitrag zum »Handbuch der Unterrichtsforschung« angenommen, Merkmale des Lehrerhandelns, die mit Effektmaßen (z.B. mit der fachlichen Lernleistung) hoch korrelierten, wären im Umkehrschluss geeignete Kriterien zur Beurteilung der Qualität von Unterricht. Die Münchener Studie sucht in gleicher Weise nach hohen Korrelationen zwischen ihren Skalen und Lernleistungen. (Tabelle VII.9 in Weinert/Helmke 1997)

Tabelle VII.9: Korrelationen zwischen Merkmalen des Unterrichts und Leistungszuwächsen

	Leistungszuwachs	
	(Residualisierte Nachtestwerte)	
<i>Unterrichtsmerkmale</i>	<b>Mathematik</b>	<b>Rechtschreiben</b>
Klassenführung <sup>a</sup>	.36**	.26
Strukturiertheit <sup>a</sup>	.28*	.17
Individ. fachl. Unterstützung <sup>a</sup>	.32*	.16
Förderungsorientierung <sup>a</sup>	.17	-.02
Variabilität d. Unt.formen <sup>a</sup>	.28	-.04
Soziales Klima <sup>a</sup>	.18	.02
Klarheit <sup>b</sup>	.34*	.17
Motivierungsqualität <sup>c</sup>	.35*	.27*

Anmerkung: \*  $p < .05$ , \*\*  $p < .01$ .

<sup>a</sup>Hochinferente Ratings geschulter Unterrichtsbeobachter, <sup>b</sup>Schülerperzeptionen, auf Individualebene erhoben und anschließend auf Klassenebene aggregiert; <sup>c</sup>niedriginferente Beobachtungen des Schülerverhaltens während des regulären Unterrichts, ebenfalls auf Klassenebene aggregiert.

Selbst dann, wenn wir unterstellten, es handele sich um relevante Befunde, ließe sich mit den Korrelationen wenig anfangen. Die Beziehungen zwischen Lernfortschritten und Unterrichtsmerkmalen sind im Fach Deutsch (Rechtschreibung) schwach und im Fach Mathematik eher mäßig ausgeprägt. Es gibt keinen klaren Trend, der sich in bei-

den Fächern widerspiegelt. Förderungsorientierung ist in beiden Fächern nicht mit Lernleistungen verknüpft. Während Variabilität des Unterrichts in Mathematik einen Zusammenhang mit der Lernentwicklung zeigt, scheint sie im Fach Deutsch völlig irrelevant. Wie würde die Unterrichtsforschung den Befund in eine einfache Botschaft verwandeln? Förderungsorientierung bringt nichts und variable Unterrichtsgestaltung ist für die Rechtschreibung eher hinderlich? Sollte man nun dafür plädieren, mehr in die Motivierungsqualität zu investieren und das soziale Klima zu vernachlässigen? Mit einigem Recht wird der geschulte Leser darauf verweisen, dass solche kausalen Interpretationen von einzelnen korrelativen Zusammenhängen nicht zulässig sind. Aber solche methodischen Vorsichtsregeln werden selbst in der Zukunft der Unterrichtsforscher nicht immer befolgt: »Dagegen erweist es sich für den durchschnittlichen Leistungsfortschritt der Klasse als vergleichsweise irrelevant, ob Lehrer ihren Schwerpunkt auf die besondere Förderung leistungsschwacher Schüler legen.« (Helmke/Weinert 1997, S. 249)

Am Anfang habe ich schon darauf hingewiesen, dass das Motiv für meinen Beitrag nicht in einer grundsätzlichen Ablehnung empirischer Unterrichtsforschung liegt. Ganz im Gegenteil plädiere ich dafür, auch diese Forschungsbemühungen noch zu verstärken. Allerdings sollte die nicht empirisch arbeitende Erziehungswissenschaft sich davor hüten, den »Fachleuten« aus den großen Instituten und den streng empirisch ausgerichteten Disziplinen einen zu großen Vertrauensvorschuss zu gewähren. Auch hier gilt: Vertrauen ist gut, methodische Kontrolle ist besser. Dass die Konjunktur wissenschaftlicher Paradigmen nicht notwendigerweise die Überlegenheit bestimmter Theorien, sondern eher die Marktcompetenz ihrer Vertreter dokumentiert, ist seit Kuhn unter Wissenschaftshistorikern bekannt. Die empirischen Methoden und Theorien der Erziehungswissenschaft sind in Deutschland häufig unterschätzt und manchmal auch überschätzt worden. Zurzeit erleben wir einen starken Bedeutungsgewinn dieser Forschung, der Erwartungen weckt, die vermutlich nicht erfüllt werden können. Es ist an der Zeit, ein wenig bescheidener das gut bestätigte empirische Wissen vom spekulativen Glauben, Hoffen und Meinen zu unterscheiden, um nicht an den selbstverschuldeten übersteigerten Hoffnungen gemessen und für zu leicht befunden zu werden. Aber vielleicht kommt eine solche Bescheidenheit bereits zu spät. Auch der empirischen Pädagogik droht eine starke Konkurrenz, vielleicht sogar eine Wachablösung an den Beratertischen der Politik. Im Süden Deutschlands wird schon an einer neuen wissenschaftlichen Revolution des Lernens und Lehrens gearbeitet, gegen die selbst die kühnsten Spekulationen und Modelle der empirischen Pädagogik nur noch als nebensächlich erscheinen. Am Ulmer »Transferzentrum für Neurowissenschaften und Lernen« entsteht mit großer Unterstützung durch das Land Baden-Württemberg ein Institut, in dem mit Hilfe der Hirnforschung nahezu alle Probleme der Pädagogik in naher Zukunft gelöst werden sollen. »Die Forschung wird so geplant, dass sich ihre Ergebnisse auf Fragen der optimalen Gestaltung von Lernbedingungen anwenden lassen. Die Ergebnisse der Grundlagenforschung werden in Programme umgesetzt. Diese wiederum werden an Schulen auf ihre Anwendbarkeit getestet. Ziel dieser Anwendungsforschung ist die rasche Übertragung neurowissenschaftlicher Erkenntnisse in die Unterrichtspraxis.«

(vgl. <http://www.uni-ulm.de/klinik/psychiatrie3/transferzentrum.html>)

LITERATUR

- Bortz, Jürgen und Döring, Nicola : Forschungsmethoden und Evaluation, 3. Aufl., Berlin 2002.
- Brügelmann, Hans: Rezension von Weinert/Helmke 1997, in: Jahrbuch Grundschulforschung 1998, S. 165–168.
- Brophy, J.E. and Good, T.L.: Teacher-effects results, in: Wittrock, Merlin C. (ed.): Handbook of research on teaching, New York 1986, S. 328–376.
- Ditton, Hartmut: Evaluation und Qualitätssicherung, in: Tippelt, Rudolf (Hrsg.), Handbuch Bildungsforschung, Opladen 2002, S. 775–790.
- Einsiedler, Wolfgang: Literaturüberblick, in: Weinert/Helmke (Hrsg.), Entwicklung im Grundschulalter, München 1997, S. 225 ff.
- Faust, Gabriele: Interventionsperspektiven aus vorschul- und grundschul-pädagogischer Sicht – Sieben Thesen. PISA in Deutschland: Können wir die Schiefelage noch korrigieren? Dokumentation der Tagung der Hanns-Seidel-Stiftung, München, 11.7.2002; im Internet auf ihrer neuen Homepage abrufbar unter:  
[http://www.uni-bamberg.de/ppp/grundschulpaed/mitarbeiter\\_faust.htm](http://www.uni-bamberg.de/ppp/grundschulpaed/mitarbeiter_faust.htm).
- Gage, Nathaniel L. (Hrsg.): Handbook of Research on Teaching, Chicago 1967.
- Good, Thomas L. and Brophy, Jere E.: Looking in classrooms, 6<sup>th</sup> ed., New York 1994.
- Good Thomas: Teaching effects and teacher evaluation, in: J. Sikula (ed.), Handbook of research on teacher education, New York 1996, S. 617–665.
- Gumpler, Edith: Grundschulforschung, in: Grundschule, Heft 7/8, 1998, S. 53–55.
- Helmke, Andreas: Unterrichtsqualität: Erfassen, Bewerten, Verbessern, Seelze 2003.
- Helmke, Andreas/Weinert, Franz E.: Unterrichtsqualität und Leistungsentwicklung. Ergebnisse aus dem SCHOLASTIK-Projekt, in: Entwicklung im Grundschulalter, Weinheim 1997, 241–251.
- Lüders, Manfred/Rauin, Udo: Unterrichts- und Lehr-Lern-Forschung, in: Helsper, Werner/ Böhme, Jeanette (Hrsg.): Handbuch der Schulforschung, Opladen 2004.
- Rosenshine, B. and Furst, N.: The Use of Direct Observation to Study Teaching, in: Travers, R.M.W. (ed.): Second Handbook of Research on Teaching, Chicago 1973, S. 122–183.
- Rosenshine, B., & Meister, C.: The use of scaffolds for teaching higher-order cognitive strategies. Educational Leadership, 49 (7), 1992, S. 26–33.
- [http://www.asu.edu/educ/eps/EPRU/documents/EPRU\\_2002-101/Chapter\\_09-Rosenshine-Final.htm](http://www.asu.edu/educ/eps/EPRU/documents/EPRU_2002-101/Chapter_09-Rosenshine-Final.htm)  
–\_ednref20
- Rosenshine, B. and Stevens, R.: Teaching functions, in: Wittrock, M.C. (ed.): Handbook of Research on Teaching. New York 1986, S. 376–391.
- Valtin, Renate: Grundschulpädagogik als empirische Forschungsdisziplin, Zeitschrift für Pädagogik, Heft 4, 2000, S. 555–570.
- Wang, M.C., Haertel, G.D. and Walberg, H.J.: Toward a Knowledge Base for School Learning, in: Review of Educational Research 63, 1993, S. 249–294.
- Weinert, Franz E./Helmke, Andreas (Hrsg.): Entwicklung im Grundschulalter, Weinheim 1997.
- Weinert, Franz E., Schrader, F.-W. & Helmke, Andreas: Educational expertise – Closing the gap between educational research and classroom practice, School Psychology International 11, 1990, S. 163–180.