Studien zur International
und Interkulturell Vergleichenden
Erziehungswissenschaft
2

Isabell van Ackeren

Evaluation, Rückmeldung und Schulentwicklung

Erfahrungen mit zentralen Tests, Prüfungen und Inspektionen in England, Frankreich und den Niederlanden

WAXMANN

Evaluation, Rückmeldung und Schulentwicklung

Studien zur International und Interkulturell Vergleichenden Erziehungswissenschaft

herausgegeben von

Wilfried Bos, Hamburg
Marianne Krüger-Potratz, Münster
Jürgen Henze, Berlin
Sabine Hornberg, Bochum
Botho von Kopp, Frankfurt (Main)
Knut Schwippert, Hamburg
Dietmar Waterkamp, Dresden
Peter J. Weber, Halle (Saale)

Band 2

Waxmann Münster / New York
München / Berlin

Isabell van Ackeren

Evaluation, Rückmeldung und Schulentwicklung

Erfahrungen mit zentralen Tests, Prüfungen und Inspektionen in England, Frankreich und den Niederlanden

Waxmann Münster / New York
München / Berlin

Bibliografische Informationen der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

Gedruckt mit Unterstützung des Bundesministeriums für Bildung und Forschung.

Bundesministerium
für Bildung
und Forschung

Diese Arbeit wurde 2003 als Dissertation von der Universität Duisburg-Essen (Standort Essen) angenommen.

Studien zur International und Interkulturell Vergleichenden Erziehungswissenschaft, Band 2

ISSN 1612-2003
ISBN 3-8309-1377-X

Postfach 8603, D-48046 Münster

http://www.waxmann.com
E-Mail: info@waxmann.com

Umschlag: Pleßmann Kommunikationsdesign, Ascheberg
Druck: Zeitdruck GmbH, Münster
Gedruckt auf alterungsbeständigem Papier, DIN 6738

Inhalt

One of the potentially most powerful mechanisms
for achieving change in education
is the external examination system.

(Harold J. Noah und Max A. Eckstein 1992)

Einführung

Das wissenschaftliche, das bildungsadministrative, das bildungspolitische und das allgemeine öffentliche Interesse an Bildung und Erziehung hat sich aufgrund verstärkter einzelschulischer Autonomie und der unterschiedlich verlaufenden Schulentwicklung in den Bundesländern während der 1990er Jahre immer deutlicher auf Qualitätsaspekte von Schule und Unterricht konzentriert. Diese Entwicklung wurde bekanntermaßen durch die Vorlage der Ergebnisse der insbesondere zwischen 1995 und 2001 veröffentlichten internationalen Leistungsvergleichsstudien und die damit verknüpften empirischen Belege mangelnder Qualitätsstandards schulischer Leistungsergebnisse auf Schulsystem- und Schulebene im deutschen Bildungswesen im Vergleich zum internationalen Maßstab verstärkt.

Verknüpfung von Messung und Steuerung

Die Wahrnehmung der aufgezeigten Missstände lässt in ihrer Folge eine ganze Nation nach möglichen Maßnahmen der Behebung fragen. Das Interesse an den Fragen, welche Leistungen Schulen zum einen hervorbringen und wie sie zum anderen ihre Arbeitsresultate bei festgestellten Defiziten verbessern können, ist enorm. Die zu bewältigende Aufgabe lässt sich wie folgt fassen: Es gilt, vom Prüfen schulischer Qualität auch zu ihrer Entwicklung zu gelangen. Dabei haben sich die diskutierten und bislang umgesetzten Maßnahmen von der Steuerung des Kontextes von Schule auf eine deutliche Akzentuierung von Maßnahmen im Bereich der Prozess- und vor allem der Wirkungssteuerung verschoben. Dementsprechend wird in Deutschland von einem Paradigmenwechsel der Schulsteuerung, der sich im Laufe der 1990er Jahre vollzogen hat, gesprochen (vgl. u.a. Helmke 2000). Dieser relativ neue Fokus drückt sich auf der Schulsystemebene u.a. in der – nach Jahren der Abstinenz – deutlichen Präsenz bei internationalen Leistungsvergleichen aus. Aber auch innerhalb der Bundesländer sind die Aktivitäten überregionaler Leistungsmessung zahlreich geworden.

Die Zahl der Absichten, die mit solch unterschiedlichen Formen großflächig angelegter, extern gesteuerter Leistungserhebungen verknüpft sind, hat sich in diesem Kontext in dem oben beschriebenen Sinne ausgeweitet, nämlich von einer reinen Bestandsaufnahme durch Leistungsmessungen und ihrer Bewertung hin zu schulreformierenden Möglichkeiten. Entsprechende Handlungsoptionen sollen einerseits durch verschiedene Formen der Leistungsmessung – vor allem in ihrer Kombination mit der Erhebung erklärungsrelevanter Kontextvariablen – identifiziert werden. Sie können aber andererseits, so die Hypothese, auch selbst als Maßnahme der Qualitätssicherung und -entwicklung wirken und als solche bewusst eingesetzt werden. Letzteres ist in zwei grundsätzlichen Modellen denkbar:

- Zum einen können Einzelschulen direkt und vertraulich über ihre Ergebnisse entsprechender überregionaler Tests, Studien etc. unterrichtet werden. Ein solches Feedback soll den Ausgangspunkt eines informationsbasierten Entscheidungshandelns vor Ort zugunsten der Qualitätsentwicklung darstellen.
- Zum anderen existiert – in der Diskussion und Praxis weniger hierzulande als in anderen europäischen Ländern – die Idee, über die Etablierung marktähnlicher Strukturen in Bildungssystemen, in denen die Veröffentlichung einzelschulischer Daten eine wesentliche Rolle spielt, in den schulischen Institutionen einen von außen gesetzten, eher indirekt wirkenden Anreiz zur Umsetzung von Maßnahmen der Qualitätsentwicklung wirksam werden zu lassen.

Die Verknüpfung von empirischer Leistungsforschung und der Entwicklung des Schulsystems insgesamt gewinnt somit in unterschiedlich ausgestalteten Reformbestrebungen an Bedeutung. Das beschriebene veränderte Qualitätsbewusstsein lässt sich dabei bis auf die einzelschulische Ebene erkennen. Diskussionen um die Wirksamkeit von Schule entspringen nicht mehr nur einem nationalen, systemischen Denken, sondern finden immer stärker in den Köpfen der Betroffenen am Schulstandort statt. So fragen sich Eltern – verstärkt durch die aktuellen PISA-Ergebnisse –, ob eine bestimmte Schule für ihre Kinder gut genug ist, und in den Schulen wächst das Bewusstsein, sich mit den Aspekten der Qualität ihrer schulischen Arbeit in unterschiedlichsten Bereichen auseinander zu setzen. In dieser Hinsicht wird auch nach Verknüpfungsmöglichkeiten zwischen großflächigen externen Evaluationsformen und der Entwicklung der Einzelschule gefragt: Wie lässt sich das Verhältnis von System- und Schulebene über Tests, Prüfungen und Studien definieren und möglicherweise steuern?

Externe Evaluation als ein Element von Bildungsreform

Der Blick über die nationalen Grenzen zeigt, dass in vielen Schulsystemen Europas die Einzelschulen in den letzten Jahren immer stärker dazu angehalten werden, Zeugnis über ihre Leistungen abzulegen. Zentral organisierte Leistungsmessungen und Evaluationsprogramme spielen dabei eine wichtige Rolle. Mit ihrer Hilfe werden Auskünfte über die Qualität der Einzelschule eingeholt und öffentlich oder/und vertraulich verschiedenen Zielgruppen zur Verfügung gestellt. Großflächige externe Evaluationsformen stellen ein wichtiges Element grundsätzlicher, auf die Anhebung des Leistungsniveaus eines Bildungssystems ausgerichteter Reformen dar: Die Schulen vieler Länder sind zum Subjekt einer Gesetzgebung und Politik geworden, die große Veränderungen in Bereichen wie ‚Curriculum', ‚Leistungsmessung' und ‚Schulautonomie' mit sich brachten. So sind die curricularen Forderungen restriktiver geworden und Testverfahren sowie regelmäßige Schulinspektionen wurden ausgeweitet, Bildungspolitik wurde bei gleichzeitiger Dezentralisierung der Verantwortung für ihre Implementierung stärker zentralisiert und verschiedene Regierungen haben versucht, Marktelemente im Bildungsbereich einzuführen, indem Eltern verstärkt Schulwahlmöglichkeiten eingeräumt wurden und die Finanzierung der Schulen an die eingeworbene Schülerzahl gekoppelt wurde. In diesem Kontext ist ein Denken in betrieblichen Begriffen auch im schulischen Bereich nicht mehr ungewöhnlich. Unterricht wird durchaus als ein Produktionsprozess gesehen, mit dem Inputs in Outputs transformiert werden.[1] Die Aus-

1 ‚Output' ist eine Bezeichnung, die der informationstechnologischen bzw. ökonomischen Fachsprache entlehnt ist. Im ersten Fall ist mit ‚Output' die Ausgabe von Daten bzw. die Gesamtheit

prägung solcher Reformen variiert von Land zu Land aufgrund historischer, kultureller, institutioneller und politischer Faktoren. Es gibt einige gemeinsame Elemente, die sich über die Ländergrenzen hinweg identifizieren lassen; großflächige, extern gesteuerte Evaluationsformen sind ein Beispiel dafür, obgleich solche Reformelemente im Kontext des jeweiligen nationalen Bildungssystems einschließlich seines historischen und gegenwärtigen sozialen Zusammenhangs zu betrachten sind.

Vor diesem Hintergrund streben Regierungen vieler Länder der ganzen Welt nach Möglichkeiten, Wissen und Kompetenzen zu testen, und sie interessieren sich dafür, ob und wie dadurch das Lehren und Lernen mit den entsprechenden Arbeitsergebnissen kontrolliert werden kann. Die Etablierung von Tests, Prüfungen und Studien wird als wichtiger Steuerungsmechanismus in vielen Ländern angesehen: Alle SchülerInnen werden demnach zur Teilnahme an standardisierten Tests und Prüfungen verpflichtet. Das Erreichen bzw. Nicht-Erreichen der geforderten Standards ist nicht selten mit positiven bzw. negativen Sanktionen, z.B. finanzieller Art, verknüpft, um in Schulen und Klassenräumen Änderungen zu erreichen und dadurch insgesamt Reformen in Bildungssystemen umzusetzen. Die erhofften Möglichkeiten einer testbasierten Reform scheinen dabei auch aus Kostengründen attraktiv. Tests sind offensichtlich weniger teuer als Reformen, die direkten Einfluss auf das Unterrichtsgeschehen nehmen wollen (vgl. Hamilton/Stecher/Klein 2002).

Forschungs- und Entwicklungsbedarf

Das Interesse am datenbasierten Wissensmanagement als Ausgangspunkt für Innovation ist folglich groß. Die Literatur zu psychometrischen Verfahrensweisen ist ebenso extensiv wie zu der Frage, welchen Einfluss standardisierte Leistungskontrollen auf SchülerInnen, Lehrkräfte und andere Bildungspraktiker haben. Der generelle Einfluss solcher Test- und Prüfungssysteme im großen Stil ist jedoch deutlich weniger klar beschrieben: Welche Effekte haben großflächige, extern organisierte Evaluationsformen auf die Institution Schule sowie auf das Schulsystem? Wie und in welchen Arbeitsbereichen werden testbasierte Informationen in der Einzelschule sowie in Bildungsadministration und -politik genutzt? Welche Wirkungen und Möglichkeiten eines effektiven Einsatzes können auf unterschiedlichen Ebenen des Schulsystems aufgezeigt werden? Die aktuelle internationale wie nationale Schwerpunktlegung auf externe Tests, Prüfungen und Studien und die damit verbundene Hoffnung ihrer positiven Wirksamkeit, deren Nachweis sich als Forschungsdefizit darstellt, lässt es sinnvoll erscheinen, sich dem Thema zu nähern und die u.a. bei Leithwood, Jantzi und Mascall (2002, S. 9) formulierte These für eine Analyse vorliegender Forschungsergebnisse aufzugreifen:

> „[...] the international context provides encouragement for reform-minded governments to view education as a source of solutions to many of their economic and social problems, to adopt reform strategies which assume that greater school accountability will improve student performance, and to implement these strategies on a large scale very quickly."

---

der ausgegebenen Daten und Informationen gemeint, im letzten Fall wird eine produzierte Ware, ein Ergebnis beschrieben. In der Erziehungswissenschaft spricht man eher von ‚Wirkungen' sowie von ‚Kontext', wenn es um den gegensätzlichen Begriff ‚Input' geht. An dieser Stelle erscheinen die Begriffe ‚In-' und ‚Output' griffiger für die Darstellung ihres Zusammenhangs – allerdings mit erziehungswissenschaftlichem Konnotat.

Terhart (2002, S. 69) bringt den Forschungsbedarf aus deutscher Sicht auf den Punkt: Er sieht das aktuelle Problem in der Frage, wie man mittels der zahlreichen und vielgestaltigen Leistungsvergleiche zu einer „Verbesserung des dergestalt ausgemessenen Systems kommt".

Grundlegende Fragestellung und Ziele der Untersuchung

In der vorliegenden Arbeit wird der Frage nachgegangen, welche bildungspolitischen Strategien es im Bereich des großflächig angelegten externen Evaluierens gibt und welche lokalen wie systemischen Effekte diese erzielen können. Als zentraler Begriff erweist sich dabei die international verwandte Wendung des ‚evidence-based planning' als Ausdruck eines begründeten Entscheidungshandelns sowohl auf der Systemebene als auch in der Einzelschule: Wie können unterschiedliche Bildungsakteure mit Informationsressourcen ausgestattet werden, so dass sie ihre Praxis mit entsprechenden positiven Effekten darauf abstimmen können? Evaluation erscheint in diesem Sinne als pragmatisches Mittel, politisches wie schulisches Vorgehen und Handeln zu legitimieren und zu rationalisieren. Was lässt sicher aber tatsächlich davon erwarten? Führen Informationen, die mit großflächigen, externen Evaluationsformen geliefert werden, zu gezielten Maßnahmen, die die aktuelle Situation in der Weise beeinflussen, dass diese dichter an den angestrebten Zustand herangebracht wird?

Ziel der vorliegenden Untersuchung ist es, der nationalen Bildungsdiskussion unter dem Gesichtspunkt der Handlungsorientiertheit weitere Bewertungsmaßstäbe hinsichtlich der Möglichkeiten, Grenzen und auch der Gefahren der externen Bestandsaufnahme von Leistung an die Hand zu geben. Zentral sind folgende Aspekte mit Blick auf überregionale externe Evaluationsformen:

- Sind großflächig eingeführte Reformelemente wie standardisierte Lernstandserhebungen effektiv, d.h. ist es möglich, zentrale Initiativen zu starten, die Bildungspraktiker dazu veranlassen, ihre Praxis im Sinne der Reforminitiativen nachhaltig zu ändern?
- Welche Faktoren sind bei der Verknüpfung der Steuerung des Gesamtsystems und der Entwicklung der Einzelschule über solche Elemente zu berücksichtigen und mit welchen Effekten ist im Einzelnen zu rechnen?

Um einer Beantwortung dieser grundsätzlichen Fragen näher zu kommen, soll der Stellenwert von externen Test- und Prüfungsformen einschließlich ihrer Verwendung und ihrer Effekte in solchen Ländern untersucht werden, die über einen breiteren Erfahrungshorizont in diesem Bereich verfügen. Dabei geht es eher am Rande um die Frage, wie die technische Qualität der praktizierten Tests, Prüfungen und Studien einzuschätzen ist. Zentral ist vielmehr der Aspekt ihrer Einordnung in das Gesamtsystem einschließlich der möglichen Verfahren, der Gebrauchsmöglichkeiten und der empirisch erfassten Wirkungen. Dabei werden sowohl solche Evaluationen in die Betrachtungen einbezogen, die die Leistungen von Individuen und Populationen standardisiert und großflächig erheben (Stichwort ‚Large Scale Assessments') als auch Evaluationen, die sich standardisiert und überregional mit Leistungen, Aufgaben und Strukturen von Organisationen auseinandersetzen (Stichwort ‚Schulinspektionen'). Beide Formen lassen sich unter dem Begriff ‚externe großflächige Evaluationen' zusammenfassen. Sie stellen in vielen Ländern einander ergänzende Reformelemente dar, die folgende wesentliche Ziele auch in ihrer Kombination verfolgen können:

- Evaluation zur Steuerung des Schulsystems durch die Schaffung einer integrativen Sicht auf dieses,
- Evaluation zur Rechenschaftslegung gegenüber der Öffentlichkeit auf der Schulsystem- und der Einzelschulebene sowie
- Evaluation zur Entwicklung der Einzelschule.

Feinziele der Studie

Zu den Feinzielen der Untersuchung gehört die Beantwortung der Frage, welche Schwerpunktlegungen hinsichtlich der Evaluationsziele in anderen Ländern vorzufinden sind, wie sie weiter differenziert werden können, welche Interdependenzen es zwischen den Zielen möglicherweise gibt und ob und wie sie schließlich erreicht werden. „Je nach Art, Umfang und Anlage einer Studie ergeben sich unterschiedliche Arten von Informationen, und diese wiederum führen zu unterschiedlichen Steuerungsmöglichkeiten [...]" (Terhart 2002, S. 72). Dementsprechend verfolgt die Arbeit die Absicht, unterschiedliche Evaluationsstudien, ihre Möglichkeiten und ihre Wirkungen zu analysieren.

Differenzierter formuliert ergeben sich folgende *Untersuchungsbereiche* und damit verknüpfte Unteraspekte:

- Systematische Darstellung der Organisationsstrukturen standardisierter, überregionaler und zentral gesteuerter, regelmäßig und unregelmäßig durchgeführter Tests, Prüfungen und Studien (‚Large Scale Assessments') sowie Schulinspektionen als externe Evaluationsformen während und am Ende der Schullaufbahn,
- Beschreibung von Feedback-Systemen dieser Evaluationsformen und damit verknüpfter Intentionen: Veröffentlichung von Daten im Kontext von Rechenschaftslegung und Rückmeldung vertraulicher Daten als grundlegende Rückmeldestrategien sowie
- Analyse der Strategien und Effekte der Datennutzung zur Steuerung des Schulsystems aus bildungspolitischer Sicht sowie zur Schulentwicklung aus einzelschulischer Perspektive.

Ländervergleich als methodischer Ausgangspunkt

Den Ausgangspunkt der Arbeit bildet der Forschungs- und Entwicklungsbedarf aus deutscher Sicht. Aber auch das Untersuchungsziel richtet sich auf das deutsche Schulsystem, indem Entwicklungsperspektiven für eine deutsche Evaluationspolitik im Bereich der Primar- und Sekundarstufe – weniger jedoch für den Hochschulbereich, die berufliche Bildung und das Sonderschulwesen – abgeleitet werden. Der Hauptteil der Arbeit besteht aus vergleichenden Analysen zu den genannten Themenbereichen in wichtigen Vergleichs- und Nachbarländern. Die exemplarische Analyse internationaler Erfahrungen im Bereich großflächiger Leistungsmessung richtet sich auf die test- und inspektionserfahrenen Länder England, Frankreich und die Niederlande. Die Ergebnisse internationaler Leistungsvergleiche haben in den gewählten Ländern während der 1980er Jahre und zu Beginn der 1990er Jahre zur Verstärkung extern gesteuerter Evaluationsformen geführt. Deutschland hingegen nimmt seine qualitativen Mängel im Bildungswesen im Vergleich zu anderen Nachbarländern viele Jahre später wahr und befindet sich mit der Verstärkung externer Evaluationselemente im internationalen Trend. Zu welchen Effekten wird die Bundesrepublik aber mit den eingeleiteten und

geplanten Maßnahmen kommen und was können wir im Vorfeld solcher Konsequenzen aus den Erfahrungen in Nachbarländern lernen?

Der internationale Vergleich bietet in Teilen den Ersatz für ein Experiment, doch wird im Kontext der vorliegenden Untersuchung sehr wohl berücksichtigt, dass ‚Evaluation' zwar ein transnationales Schlagwort geworden ist, sich jedoch die Umsetzung der dahinter stehenden Absichten und Ziele sehr deutlich über die Ländergrenzen hinweg unterscheidet. In den Untersuchungsländern nähert man sich dem Thema der Evaluation sehr verschieden, was sich als Ergebnis verschiedener Traditionen, Verwaltungstypen etc. darstellt. Unter Berücksichtigung der Grundsätze der Vergleichenden Erziehungswissenschaft wird der Vergleich nicht abstrakt, sondern in engem Bezug zur vorliegenden Datenbasis beantwortet, die in den gewählten Ländern unterschiedlich differenziert ist. Die vergleichenden theoretischen und praktischen Analysen stützen sich auf Literatur- und Materialrecherchen sowie auf die Ergebnisse von in den Untersuchungsländern durch die Verfasserin durchgeführten Experten-Ratings.

Aufbau der Arbeit

Vor dem Hintergrund der skizzierten defizitären Forschungslage, der darauf gestützten Fragestellung der Untersuchung und der ersten Darstellung des methodischen Vorgehens stellt sich der Aufbau der Arbeit wie folgt dar:

Die Untersuchung gliedert sich in drei Teile, mit denen
- zunächst eine Bestandsaufnahme der Strukturen und Strategien der Implementierung großflächiger Evaluationsformen ausgehend von Deutschland und dann mit Blick auf die drei Untersuchungsländer geleistet wird,
- nachfolgend Forschung zu den tatsächlichen Effekten großflächiger Leistungsuntersuchungen in England, Frankreich und den Niederlanden zusammengetragen und vergleichend analysiert wird und schließlich
- in einem dritten Arbeitsteil auf der Grundlage einer zusammenfassenden Übersicht über die recherchierten Effekte der Evaluationsformen mögliche Perspektiven für das deutsche Schulsystem in diesem Bereich hergeleitet werden.

Im Einzelnen stellen sich diese drei Untersuchungsteile wie folgt dar: Der *erste Teil* der Arbeit nimmt mit *Kapitel A* die deutsche Situation mit ihrem wachsenden Datenbestand durch zahlreiche und vielgestaltige großflächige Tests, Prüfungen und Leistungsstudien auf und leitet aus der Forschungslage zu diesen Evaluationsformen hinsichtlich der Rückmeldung und Nutzung der auf diese Weise generierten Daten Forschungsbedarf ab. Die Betrachtung der Test- und Prüfungspraxis einschließlich entsprechender Forschung in den gewählten Untersuchungsländern erweist sich als Chance, mehr über die Möglichkeiten und Effekte unterschiedlicher Evaluationsformate im Sinne der Grundlagenforschung, aber auch hinsichtlich ihres Anwendungsbezuges zu lernen. In *Kapitel B* werden dementsprechend die Möglichkeiten des Vergleichs in der Erziehungswissenschaft im Allgemeinen sowie des Vergleichs der ausgewählten Länder England, Frankreich und Niederlande im Speziellen erläutert und damit die Länderwahl begründet. In den nachfolgenden Unterkapiteln wird die Test- und Prüfungspraxis der einzelnen Untersuchungsländer in den jeweiligen Kontext der nationalen Schulentwicklung eingebettet, denn eine isolierte Betrachtung externer Evaluationselemente könnte deren Intentionen und Wirkungen nicht im Zusammenwirken ver-

schiedener Reformelemente herleiten sowie verstehend analysieren und interpretieren. Zu den *wesentlichen länderübergreifenden Reformelementen*, die im Zusammenhang mit den jeweiligen Test-, Prüfungs- und Inspektionsstrukturen zu betrachten sind, gehören, wenn auch mit unterschiedlicher länderspezifischer Ausprägung:

- die Festschreibung von Standards, deren Erreichung durch standardisierte Leistungsmessungen kontrolliert wird,
- die Veröffentlichung von Ergebnissen externer Evaluationen,
- die Möglichkeit einer freien Schulwahl und damit eine handlungsrelevante Orientierung der Öffentlichkeit an transparenten Leistungsdaten sowie
- das Ausmaß der einerseits zunehmenden Autonomisierung der Einzelschule, deren Verpflichtung zur Rechenschaftslegung über erreichte Arbeitsresultate auf der anderen Seite verstärkt wird.

Diese Aspekte erweisen sich als zentral für die Wirkungsentfaltung externer, überregionaler Formen der Leistungserhebung, will man nicht allein die Effekte der direkt an Schulen rückgemeldeten Daten und die Effekte eines solchen Feedbacks untersuchen, sondern den Wirkungsbegriff breiter fassen und auch indirekt erzeugte Mechanismen im Blick haben. Am Ende dieses Kapitels stehen schließlich eine Zusammenfassung sowie ein Vergleich der Entwicklungstrends in den Untersuchungsländern. In *Kapitel C* wird die Entwicklung, Durchführung, Auswertung, Rückmeldung und Nutzung von Daten im Zusammenspiel von Ministerien, Testinstituten und Schulinspektoraten dargestellt und vergleichend betrachtet, denn in allen Untersuchungsländern sind verschiedene Instanzen mit Evaluationen befasst. Vieles ist deskriptiv, um die Möglichkeiten der Ausgestaltung von institutionalisierten Evaluationssystemen aufzuzeigen, die dem deutschen Betrachter der ausländischen Situationen eher fremd erscheinen, sich aber durchaus in den internationalen Trend einordnen. Zugleich wird bei einer solchen Betrachtung wichtiger Einrichtungen der Test- und Prüfungssysteme ihre politische Dimension verdeutlicht, die sich als wesentlich erweist, zumal der Impetus für eine testbasierte Bewegung in Bildungssystemen im Wesentlichen von Politikern und Akteuren außerhalb des Schulalltags ausgeht. In diesem letzten Kapitel des ersten Arbeitsteils werden die nationalen Test- und Prüfungseinrichtungen in den Ländern zunächst länderweise vorgestellt und anschließend verglichen. Ihre Anbindung an die nationalen Bildungsministerien und die Zusammenarbeit zwischen diesen Einrichtungen findet dabei ebenfalls Berücksichtigung.

Der *zweite Teil* der Studie stellt die bisherigen wissenschaftlichen Befunde des Wirkungsgrades unterschiedlicher überregionaler Evaluationen den im ersten Teil beschriebenen Intentionen gegenüber. In *Kapitel A* wird zunächst auf der Grundlage entsprechender Literatur ein theoretisches Rahmenkonzept der Evaluationswirkungen entwickelt. Dies erweist sich als notwendig, um zum einen die vorgefundenen empirischen Belege aus den einzelnen Ländern mit dem Ziel ihres Vergleichs zu systematisieren und um zum anderen Forschungsdefizite durch die Gegenüberstellung von wissenschaftlich fundierten theoretischen Annahmen und empirisch abgesicherten Erkenntnissen besser identifizieren zu können. Bei der Entwicklung eines solchen Rahmenkonzeptes stehen nicht allein verschiedene Modelle möglicher Nutzungsformen rückgemeldeter Ergebnisse im Vordergrund. Um die Komplexität des Themas zu erfassen, müssen vielmehr auch die Wechselwirkungen mit kontextuellen Einflussgrößen berücksichtigt werden, die im ersten Teil bereits angedeutet und an dieser Stelle

systematisch aufgegriffen werden. *Kapitel B* orientiert sich für jedes einzelne Länderkapitel an der erarbeiteten Gliederung des vorangegangenen Kapitels und systematisiert für England, Frankreich und die Niederlande die vorgefundenen empirischen Befunde, die aufgrund der Forschungslage in den Ländern unterschiedlich differenziert sind. Ein Vergleich der Nutzungsstrategien und Effekte schließt dieses Kapitel ab. Mit dem *dritten Teil* der vorliegenden Arbeit werden in *Kapitel A* die wesentlichen Befunde der Untersuchung vor allem aus dem zweiten Untersuchungsteil zusammengefasst. Im abschließenden *Kapitel B* werden auf dieser Grundlage Empfehlungen für die Gestaltungen bildungspolitischer Maßnahmen im Bereich der Wirkungssteuerung aus deutscher Sicht formuliert und es wird der weitere Forschungsbedarf skizziert, der sich auf einen neuen zu entwickelnden Forschungstyp in Deutschland richtet.

Der Aufbau der Untersuchung macht deutlich, dass sich der Begriff der ‚Wirkung' externer Evaluationen sowohl auf die intendierten als auch auf die erzielten Effekte bezieht. Die alleinige Betrachtung von Effekten, ohne diese den Intentionen gegenüber zu stellen, ließe es nicht zu, externe Leistungsmessung umfassend bewerten zu können, die Möglichkeiten der Steuerung zu beleuchten und vor allem die Validität der Testsysteme zu beurteilen. Der Forschungsbedarf ist deutlich formuliert: „Die Frage, wie man vergleichende Leistungsinformationen produktiv nutzen kann, wie ein solcher Rückmelde- und Einbringungsprozess zu gestalten ist und was geschieht, wenn man extern erzeugte Informationen in interne Prozesse einbringt, ist innerhalb der empirischen Schul- und Unterrichtsforschung weitgehend ungeklärt" (Terhart 2002, S. 87). Ziel der Arbeit ist es, den Forschungsstand in anderen Ländern möglichst umfassend zu erheben, obgleich dabei – zumal bei drei Untersuchungsländern einschließlich der Analyse der deutschen Situation – wichtige Literatur und Dokumente unbeabsichtigt der Aufmerksamkeit der Verfasserin entgangen sein mögen bzw. bei einem irgendwann zu ziehenden Schlussstrich nicht mehr berücksichtigt werden können. Was sich an dieser Stelle bereits festhalten lässt, ist die Tatsache, dass die vorliegende Literatur- und Materialrecherche einschließlich der erarbeiteten ersten Systematisierung der Wirkungen externer überregionaler Leistungserhebung nur den Ausgangspunkt einer dringend benötigten weiteren empirischer Forschung darstellen kann.

I Strukturen und Strategien der Implementierung großflächiger externer Evaluationsformen

‚Evaluation' und ‚Schulentwicklung' stellen Begriffe dar, die als eng miteinander verknüpft erscheinen. Dies galt bislang eher für die Verbindung interner Evaluationsformen mit der Entwicklung von Schule. In jüngerer Zeit ist mit dem verstärkten Wiedereinstieg der Bundesrepublik in unterschiedliche Formen von Leistungsstudien auch die Frage der Wirkung externer Evaluationsformen für die Schulentwicklung gestellt worden; dies wird mit dem Titel der Arbeit aufgegriffen. Es ergibt sich zunächst ein grundsätzlicher Klärungsbedarf hinsichtlich des semantischen Gehaltes der Begriffe ‚Evaluation', ‚Entwicklung von Schule' aber auch mit Blick auf die Begriffe der ‚Rückmeldung' sowie der ‚Wirkung', um den Untersuchungsgegenstand deutlicher zu kennzeichnen.

Zum Evaluationsbegriff

Eine etymologische Betrachtung des Begriffs ‚Evaluation' lässt eine Verbindung zum englischen ‚value'-Begriff erkennen. Er umschreibt den Wert, die Nützlichkeit und Qualität von etwas, aber auch das Ergebnis einer Messung. Evaluation meint in einem weitesten Sinne alle Typen systematischer und sachlicher Feststellung und Wiedergabe von Fakten, etwa über das Funktionieren von Schlüsselaspekten von Bildungssystemen. Posch und Altrichter (1997, S. 29) definieren Evaluation als „kontinuierliche Überprüfung der Zweckmäßigkeit von Handlungsvollzügen" und verweisen damit auf ein weiteres Kennzeichen von Evaluation: Die Bewertung auf der Basis einer Analyse der vorgefundenen Daten ist zentrales gemeinsames Moment unterschiedlichster Evaluationsformen. Dieser letztere Aspekt erweist sich als nicht unproblematisch, da – ausgehend von einem Maßstab – eingeschätzt wird, wie gut oder schlecht eine Leistung, ein Verhalten etc. ist. Dies kann zu Konsequenzen für die evaluierte Einrichtung oder einzelne Evaluierte führen, zumal Evaluation über eine Bestandsaufnahme hinaus bei festgestellten Defiziten auch auf Änderung zielen kann. Dabei kann sich ein Spannungsverhältnis zwischen einer vertraulichen oder gar öffentlichen Rechenschaftslegung über erreichte bzw. nicht erreichte Arbeitsergebnisse und dem Entwicklungsbedarf aus der Perspektive organisatorischen Lernens ergeben. Dieses, an dieser Stelle nur angedeutete und später noch aufzugreifende Problem, wird vor allem externen Evaluationsformen zugeschrieben. Externe bzw. Fremd-Evaluation durch von außen kommende Evaluatoren soll nicht mehr nur die Selbstreflexion von z.B. schulischen Einrichtungen mit dem Ziel einer Objektivierung und der Einhaltung fachlicher Standards ergänzen. Die Absichten unterschiedlicher externer Evaluationsformen, vorallem derer, die überregional und standardisiert angelegt sind, sind zahlreicher geworden, wobei die erhofften Effekte durchaus im Widerspruch zueinander stehen können.

Großflächige externe Evaluationsformen können unterschiedliche Absichten verfolgen:

- Unterstützung von Lernprozessen durch Diagnose (formative/diagnostische Evaluation)
  Beispiel: diagnostische Tests während der Schullaufbahn

- Zertifizierung einer erreichten Leistung und damit verknüpft die Selektion beim Übergang in abnehmende Institutionen (summative bzw. bilanzierende Evaluation)
  Beispiel: Zentrale Abschlussprüfungen am Ende von Schulkarrieren
- Rechenschaftslegung über den Wirkungsgrad eingesetzter Ressourcen gegenüber der Öffentlichkeit (ebenfalls eine Form bilanzierend-summativer Evaluation)
  Beispiel: Zentrale Abschlussprüfungen, standardisierte Schulinspektionen
- Monitoring von Bildungsstandards als systematisches und regelmäßiges Verfahren des Sammelns von Daten zu wichtigen Aspekten des Bildungswesens (nicht nur Outcome-Variablen), um ein Bild der Gesamtsituation zu erhalten.
  Beispiel: Schulleistungsstudien, Vergleichsarbeiten, Schulinspektionen

Mit dieser Auflistung der möglichen Funktionen sind zugleich beispielhaft Formen überregionaler externer Evaluation benannt, die einer Systematisierung bedürfen. Welche Evaluations-Typen finden in der Untersuchung Berücksichtigung und welche nicht? Sie sollen nachfolgend aus den denkbaren Möglichkeiten begründet ausgewählt und kategorisiert werden.

Formen überregionaler externer Evaluation

Sowohl national als auch im internationalen Vergleich ist derzeit eine Ausweitung der Praxis verschiedener Typen standardisierter schullaufbahnbegleitender Tests und Studien sowie zertifizierender Abschlussprüfungen zu beobachten. Diese durchaus sehr unterschiedlichen Formen externer Evaluation haben im Kern Folgendes gemeinsam:

- Sie gehen über die Leistungserhebung an der schulischen Einzelinstitution hinaus.
- Sie beziehen sich auf Lernergebnisse, die oftmals an vielen hundert Schulen und für tausende SchülerInnen erfasst werden.
- In dieser Hinsicht haben sie einen international so bezeichneten ‚large scale'- bzw. großflächigen Charakter, da sie sich auf eine Vielzahl von in die Untersuchung einbezogenen Personen und Institutionen beziehen und eine Region und nicht nur einen einzelnen Ort berücksichtigen.
- Sie unterliegen zudem einer zentralen Steuerung, da sie von einem Punkt aus, z.B. durch eine Behörde oder/und eine universitäre Forschungsgruppe oder durch einen kommerziellen Anbieter standardisiert in ihrer Entwicklung, Durchführung und Auswertung gesteuert und begleitet werden.[2]

Auf der Grundlage dieser Kennzeichen wird die weitere Analyse folgender externer Evaluationen begründet, die man zudem auch international kennt, was den Blick ins Ausland zu diesem Themengebiet ermöglicht und sinnvoll erscheinen lässt. In die Analyse werden folgende Evaluationstypen einbezogen:

- Large Scale Assessments als Untersuchungsschwerpunkt:
  - standardisierte, überregionale Tests während der Schullaufbahn, die systematisch das Erreichen eines bestimmten Leistungsniveaus messen,
  - zentral organisierte Abschlussprüfungen und
  - stichprobenartige und flächendeckende Schulleistungsstudien auf der Makroebene des Schulsystems (Lehmann 2001b spricht von „Systemevaluationen") sowie

2 vgl. zur Begriffsbestimmung auch van Ackeren 2003

- standardisierte, überregionale Inspektionen oder Audits, die sich auf die einzelschulische Institution insgesamt sowie ihre MitarbeiterInnen beziehen und durch unabhängige ExpertInnen durchgeführt werden.

Der Begriff ‚Large Scale Assessments' hat in der aktuellen Qualitätsdebatte Prominenz erlangt, obgleich die inhaltliche Füllung dieser Wortgruppe nicht immer eindeutig erscheint. Unter Einbeziehung des internationalen Begriffsverständnisses und unter Berücksichtigung der oben benannten Charakteristika erweist sich der Ausdruck als offen für eine Vielfalt von Typen und Varianten von Leistungsvergleichen.[3] Sie lassen sich mit den Schlagworten ‚Tests', ‚Prüfungen' und ‚Studien' als Assessment-Formen weiter klassifizieren, wobei letztere als so genannte ‚Schulleistungsstudien' dem allgemeinen Begriffsverständnis von Large Scale Assessments wohl am nächsten kommen. Dennoch sollen auch die beiden anderen Formen, standardisierte Tests und Prüfungen, vor dem Hintergrund der gegebenen Begriffsdefinition berücksichtigt werden.

Regelmäßige Schulinspektionen, bei denen alle Schulen eines Landes systematisch besucht und mit einem möglichst gleichen Bewertungsmaßstab in ihren Arbeitsergebnissen bewertet werden, stellen eine traditionelle Form externer Evaluation dar. Hinsichtlich der in vielen europäischen Ländern praktizierten Verfahren und formulierten Zielsetzungen solcher Inspektionen unterscheiden sich diese allerdings deutlich vom Verständnis und der Organisation der Schulaufsicht in Deutschland. Schulinspektionen als externe Evaluationsformen sollen deshalb in diese Untersuchung einbezogen werden, da sie die oben genannten Kriterien der Ziele und der Organisation überregionaler externer Evaluation erfüllen. Sie haben darüber hinaus eine doppelte Funktion:

- Zum einen leisten sie selbst eine Bestandsaufnahme, Analyse und Bewertung von Arbeitsergebnissen. Zudem wird schulische Qualität breiter gefasst als im bloßen Bezug auf Testergebnisse, indem schulische Organisationsstrukturen sowie der einzelschulische Kontext stärker berücksichtigt werden.
- Zum anderen geben sie Hinweise auf die Effekte der anderen benannten externen Evaluationen, sozusagen als Meta-Evaluation der externen Evaluationen.

Aus diesen Gründen werden Schulinspektionen an dieser Stelle berücksichtigt, aber nicht den Large Scale Assessments untergeordnet. Durch ihre Abgrenzung davon wird vielmehr ihre besondere Rolle im Wechselspiel mit Formen von Large Scale Assessments herausgestellt. Die Charakterisierung großflächiger Leistungsmessung macht zugleich deutlich, welche Formen der Erhebung nicht in die Untersuchung einbezogen werden sollen, die möglicherweise mit Begriffen wie ‚Leistung', ‚Messung', ‚Erhebung' oder ‚Kontrolle' assoziiert werden können: Nicht berücksichtigt werden die über acht Millionen Klassenarbeiten, die jährlich in den ca. 410.000 Klassen und Kursen an deutschen Schulen geschrieben werden (vgl. van Ackeren/Klemm 2000). Diese Form der Leistungskontrolle bezieht sich nicht auf eine Region und findet auf der Mikroebene der Interaktion im Klassenraum statt. Sie ermöglicht zwar den individuellen Ver-

---

3 Eine der wenigen, in der Literatur vorzufindenden Definitionen des Begriffes ‚large-scale' findet sich u.a. bei Hamilton/Stecher/Klein 2002, S. 2:
„*Large-scale tests* [Hervorhebung im Original, IvA] are administered to large numbers of students across classrooms, schools, and districts. They are developed and mandated by parties external to a particular classroom or school. Large-scale tests include commercially developed tests that are administered as part of a district's or state's testing program, as well as tests that are developed by districts or states themselves."

gleich innerhalb einer Klasse oder – wie im Fall der so genannten Parallelarbeiten – den Vergleich zwischen Parallelklassen, um vergleichbare Standards zu sichern. Doch gehen diese Beispiele nicht über die Einzelschule hinaus und haben somit keinen großflächigen Charakter. Ebenfalls werden als Varianten externer Evaluation keine Peer-Reviews – als Einladung so genannter ‚kritischer Freunde' durch die Schulen selbst – einbezogen.

Zur Begriffsbestimmung der ‚Wirkung' externer Evaluation

Der erste Teil der Untersuchung beleuchtet zunächst die deutsche Situation im Bereich der in der beschriebenen Weise verstandenen großflächigen Tests und Prüfungen und geht den folgenden Fragen nach:

- Welche Formen dieser großflächigen Leistungserhebungen kennt man in den einzelnen deutschen Bundesländern und länderübergreifend auf Bundesebene?
- Welche Forschung gibt es zur tatsächlichen und möglichen Nutzung bzw. – genereller gesprochen – zu den Effekten der Ergebnisse für Entwicklungsprozesse auf der administrativen Systemebene und in der einzelnen Schule?

Die Antwort auf die Frage nach den derzeit in Deutschland genutzten Möglichkeiten der vergleichenden Leistungsmessung und -kontrolle zeigt einen breites Spektrum praktizierter Leistungserhebungen in allen Ländern; extern gesteuerte, großflächige Leistungserhebungen sind bundesweit ein Thema, wie nachfolgend gezeigt wird. Hinsichtlich der Untersuchung der Effekte und der Nutzbarmachung der Leistungsmessung zeigt der Blick auf die deutsche Situation allerdings ein Forschungsdesiderat auf. Die vorliegende Studie legt ihren Schwerpunkt auf die Verfahren der Auswertung der externen Evaluationsformen, vor allem aber auf ihre Rückmeldung, Rezeption und die Ableitung bzw. Erarbeitung von Handlungsstrategien. Unter der für die Untersuchung zentrale Bezeichnung ‚Wirkung' werden in diesem Zusammenhang

- sowohl direkte als auch indirekte Effekte verstanden:
  - Informationen führen direkt zu Handlungskonsequenzen bzw.
  - Daten führen in einem komplexen Steuerungssystem im Zusammenwirken bestimmter Strukturen im Schulsystem zu Effekten.
- Effekte werden auf der Systemebene sowie auf der Schulebene berücksichtigt:
  - mit Blick auf Bildungsadministration und Bildungspolitik sowie
  - hinsichtlich der schulischen Institution bzw. bestimmter Personengruppen und
- es geht um die Unterscheidung von Rezeption auf der einen und von Nutzung extern erhobener Evaluationsergebnisse auf der anderen Seite:
  - Wahrnehmung rückgemeldeter Daten sowie
  - ‚Handlungsstrategien' im Sinne der Umsetzung der Daten in gezieltes Handeln.
- Schließlich interessieren sowohl intendierte als auch tatsächliche Effekte.

Der Untersuchungstitel macht zudem deutlich, dass Wirkungen großflächiger Evaluation vor allem im Bereich der Entwicklungsmöglichkeiten von Schule in diesem Zusammenhang gesucht werden sollen. Es geht im Wesentlichen um die Frage, wie die Ergebnisse großflächiger Tests, Prüfungen und Studien für die Schulentwicklung genutzt werden bzw. genutzt werden können.

Zur zentralen Frage der Schulentwicklung durch großflächige externe Evaluationen

Der Bedeutungsinhalt des Begriffes ‚Schulentwicklung' hat seit Ende der siebziger Jahre einen Perspektivwechsel erfahren: von der Betrachtung und Untersuchung des Schulsystems als Ganzheit mit damit verknüpften „System- und Globalstrategien" (Buhren/Killus/Müller 1998, S. 237) hin zur Einzelschule, die – auch durch die Systemtheorie beeinflusst – spätestens seit Beginn der neunziger Jahre in den Mittelpunkt dieser Forschungsrichtung rückte. Dahinter steht die Erkenntnis, dass die individuelle Schule durch das administrative System kaum geradlinig und in einer direkten Art und Weise zu steuern ist. Für einzelschulische Entwicklungen sind „[...] in erster Linie die Lehrpersonen und die Leitung selbst verantwortlich" (Rolff 1998, S. 297); dies gilt um so mehr im Kontext aktueller Autonomisierungs-Tendenzen, denen externe Evaluationsformen, etwa in Form zentraler Tests und Prüfungen, als funktionales, der Gefahr der Auseinanderentwicklung entgegentretendes Äquivalent gegenüber stehen. Im Sinne der modernen Schulentwicklungs-Forschung wird der Begriff der ‚Schulentwicklung' in der vorliegenden Untersuchung relativ breit gefasst. Folgende Aspekte erscheinen zentral:

- Im Fokus der Forschung zur Schulentwicklung steht die einzelne Schule am individuellen Schulstandort (vgl. u.a. Rolff 1998). Schulentwicklung zielt auf die bewusste Weiterentwicklung von Schule, auf ihre Fähigkeit, sich selbst zu organisieren, zu reflektieren und zu steuern. Sie versteht sich dabei als „Synthese von Organisations-, Unterrichts- und Personalentwicklung" (Kempfert/Rolff 1999, S. 22). Diesen Aspekten der Entwicklung der Einzelschule soll weiter nachgegangen werden.
- Es geht aber auch um die Rahmenbedingungen der Unterstützung von Schulentwicklung (vgl. Kempfert/Rolff 1999, S. 22), d.h. die einzelschulische Entwicklung soll im vorliegenden Untersuchungszusammenhang vom Gesamtsystem her gedacht werden und die Verknüpfung zwischen Einzelschulen und dem administrativen System durch die Ergebnisnutzung großflächiger Tests und Prüfungen untersucht werden. Schulentwicklung wird folglich auch im Systemzusammenhang betrachtet. Die Einzelschule wird als in sich geschlossen und zugleich offen gesehen (vgl. Rolff 1998).

Gegenüber einer schulbezogenen Verwendungsmöglichkeit großflächiger Tests und Prüfungen über die Bereitstellung von Handlungswissen auf der Schulsystemebene hinaus zeigen sich viele Akteure sowohl auf der administrativen als auch auf schulischer Seite nicht mehr zu skeptisch; zumindest sind sie an entsprechenden Möglichkeiten interessiert. Dies gilt international sehr viel deutlicher als auf nationaler Ebene. Das Nachdenken über und die Realisierung einer Koppelung von Leistungsergebnissen, für die zudem häufig erklärende Hintergrundvariablen erhoben und mitgeteilt werden, mit Optimierungsstrategien an den untersuchten schulischen Einrichtungen sowie auf der Systemebene stellt die Begriffe ‚Rückmeldung von Evaluationsergebnissen', ‚Wirkung' und ‚Schulentwicklung' immer mehr in einen Zusammenhang. Qualitäts- und Schulentwicklung beginnen zu verschmelzen, indem es nicht mehr nur um Verbesserung von Lernresultate geht, sondern auch um die Stärkung der Fähigkeit, den eigenen Wandel zu managen (vgl. Hopkins u.a. 1994). Die Untersuchung dieser Wechselbeziehung bedarf in Deutschland – wie nachfolgend dargestellt wird – weiterer Forschung.

Externe Evaluation im Kontext von ‚school effectiveness' und ‚school improvement'

Aufgrund gemeinsamer Kernfragen ist international ein Prozess in Gang gekommen, der die Forschungsansätze der international so bezeichneten ‚school effectiveness research' und der ‚school improvement research' einander näher kommen lässt. Dieser Paradigmenwechsel betrifft vor allem den Bedarf, Schulentwicklung stärker als in der Vergangenheit in Relation zu den schulischen Arbeitsergebnissen zu setzen.

‚School improvement research' meint – in Ergänzung zur oben erwähnten deutschen Forschungstradition in diesem Bereich – eine Forschungsrichtung, die sich in den 60er Jahren gegen extern und top-down vorgegebenen Wandel in curricularen und organisatorischen Fragen wandte. In den frühen 80er Jahren entwickelte sich die ‚school improvement'-Forschung schnell weiter. Es ging um ein neues Paradigma, das wie folgt gekennzeichnet ist (vgl. Gray u.a. 1999, S. 21 ff)

- Im Vordergrund stand und steht bis heute die Idee einer ‚bottom-up'-Orientierung, d.h. die individuelle Schule und ihre MitarbeiterInnen sollen für schulische Entwicklung zuständig sein bzw. ‚top down'- und ‚bottom up'-Strategien sollen sich sinnvoll ergänzen. In umgekehrter Perspektive müssen extern eingeleitete Reformen dem individuellen Schulcharakter in seiner Wechselwirkung mit der Umgebung Rechnung tragen.
- Bislang wurden vor allem Entwicklungsprozesse, die systematisch geplant und gemanaged werden müssen, und weniger schulische Arbeitsergebnisse betont.
- Die Forschungsmethodik ist eher qualitativ ausgerichtet, so dass Verbesserungen und Entwicklungen weniger mit Daten belegt werden.
- Es gibt vor allem ein Interesse, Schulen als dynamische Institutionen zu sehen, weshalb in dieser Forschungstradition der Zeitaspekt eine wichtige Rolle spielt, um Entwicklungen und Wandel nachvollziehen zu können.
- Der Fokus liegt zudem auf der Schulkultur und weniger auf der Schulstruktur.

Eine zweite wichtige internationale Forschungstradition wird als *‚school effectiveness research'* bezeichnet. Sie hat in ihren Ursprüngen ebenfalls einen eher reaktiven Charakter und wandte sich gegen die These, dass Schulen keinen Unterschied hervorrufen (‚schools made no difference', vgl. Coleman u.a. 1966) und Unterschiede allein auf Hintergrundmerkmale der SchülerInnen zurückzuführen sind. In diesem Sinne konzentriert sich die Schuleffektivitäts-Forschung vor allem

- auf schulische Lernergebnisse, wobei zunehmend gefragt wird, wie Schulen diese erreichen und wie sie sich verbessern können,
- auf die Organisation von Schule und weniger auf ihre Kultur,
- auf Charakteristika von Schulen, die sich in ihrer Effektivität verbessert haben und
- sie bedient sich dabei vor allem quantitativer Methoden unter Zuhilfenahme von Datenanalyse-Techniken.

Seit dem Beginn der 1990er Jahre existiert ein klareres Bild davon, wie schulische Leistungsergebnisse am besten evaluiert werden. Zu den wichtigen Fragen und Aufgaben seit dieser Zeit gehören u.a. (vgl. Gray u.a. 1999, S. 27 ff): die Stabilität schulischer Effektivität im Laufe der Zeit, schulische Wirksamkeit in verschiedenen Bereichen (unterschiedliche Fächer, kognitive und affektive Leistungen...), die unterschiedliche schulische Effektivität für verschiedene Schülergruppen (z.B. mit unterschiedlichem sozio-

ökonomischem/ethnischem Hintergrund), das Ausmaß der gemessenen Leistungsunterschiede auf der Grundlage unterschiedlicher statistischer Methoden, Prozesse in ‚ineffective schools' und Faktoren, die schulische Effektivität fördern können. Zu diesem letzten Aspekt gehören die Führungsrolle der Schulleitung, die Einbindung von Mitarbeitern in Entscheidungsfragen, die Erwartungshaltung hinsichtlich der möglichen Schülerleistung, Strategien zur Einbindung der Eltern sowie die Schaffung organisatorischer Kohäsion, um Informationsflüsse und die Einbindung der MitarbeiterInnen zu erleichtern. In diesen Bereich fällt auch die Frage der Rolle externer Evaluation. Mit ihr werden zum einen Schülerleistungen und weitere Kennzeichen von Schule im Sinne der Schuleffektivitäts-Forschung erhoben, zum anderen wird ihnen eine Funktion im Rahmen der Entwicklung und Innovation von Schule zugesprochen.

Es scheint zu einem nicht unerheblichen Anteil der Gebrauch unterschiedlicher Forschungsansätze durch Politiker zur Verbesserung ihrer Praxis zu sein, der beide Forschungsrichtungen einander näher gebracht hat und bringt, wobei dieser Prozess der ‚merging traditions' in anderen Ländern fortgeschrittener ist als hierzulande. Einen Forschungsbedarf auf dem Gebiet des Umgangs mit Daten und Ergebnissen externer Evaluation hat man in anderen europäischen Ländern für die eigene nationale Situation bedeutend früher erkannt: In den exemplarisch gewählten Ländern England, Frankreich und den Niederlanden ist man vertrauter mit entsprechenden Handlungs- und Sachzusammenhängen, da der Trend ausgeweiteter Test- und Prüfungsverfahren dort bedeutend früher als hierzulande einsetzte.

In Abgrenzung zur zunächst dargestellten deutschen Situation gibt dieser erste Untersuchungsteil eine Übersicht über die Formen der Leistungserhebung in diesen Ländern, eingebettet in eine Darstellung der jeweiligen Strukturen der nationalstaatlichen Evaluationssysteme. Das Kapitel zur methodischen Anlage der Untersuchung leitet schließlich zur darstellenden Analyse der administrativen Strukturen und Intentionen der Ergebnisnutzung sowie zur Forschung zu den realisierten, tatsächlichen Strategien am einzelschulischen Standort im zweiten Untersuchungsteil über.

A Die deutsche Situation – Bedarf der Nutzbarmachung von Test- und Prüfungsergebnissen

In Deutschland konzentrierte sich die schulische Qualitätsdebatte viele Jahre lang auf die so genannte Kontext- und nachgeordnet auch auf Prozessqualität. Die Kontextqualität, die auch – ökonomisch formuliert – als ‚Inputqualität' bezeichnet wird, bezieht sich auf Merkmale der gegebenen schulischen und außerschulischen Umwelt einschließlich der materiellen und personellen Ausstattung der Schulen. Zu den personellen Eigenschaften sind zum Beispiel auf Seiten der Lehrkraft die Anzahl der LehrerInnen an einer Schule – auch im Verhältnis zur Schülerzahl –, ihre Qualifikation, aber auch pädagogische Einstellungen, Werthaltungen oder das persönliche Engagement zu zählen. Auf der anderen Seite gehören auch die SchülerInnen zur personellen Ausstattung einer Schule; sie sind beispielsweise charakterisiert durch ihre Anzahl (Klassenfrequenz), die geschlechtsspezifische Zusammensetzung, durch verschiedene Leistungsniveaus, einen bestimmten häuslichen Hintergrund und durch den Anteil von SchülerInnen mit einer Migrationsgeschichte. Diese personellen Merkmale stellen genauso wie die materiellen Kennzeichen – zu diesen gehören beispielsweise die sächliche Ausstattung an Schulen (Lehr- und Unterrichtsmaterial etc.), die Ausgestaltung des Bildungssystems sowie die Bildungsausgaben einschließlich ihrer Verteilung – mehr oder weniger starke Einflussgrößen auf die schulische Leistungserbringung dar. Dies gilt auch für die Prozessqualität, die Merkmale der Klassenführung, der Lehrer-Schüler-Interaktion und des Unterrichts meint, der beispielsweise durch die didaktisch-methodische Gestaltung, durch die Wahl von Arbeitsformen, durch die Zeitnutzung (‚time on task') und durch diagnostische Kompetenzen geprägt wird und dabei wiederum von Inputqualitäten beeinflusst ist.

Zweifel an der Wirksamkeit bisheriger Qualitätssicherung und -entwicklung

Offensichtlich hat die Konzentration auf diese Aspekte schulischer Qualität einschließlich der Betonung ihrer Steuerung nicht zur Sicherung und Entwicklung von Qualität in ausreichendem Maße beitragen können. Die gegenwärtigen Schulleistungsstudien, die einen „unterschiedlich hohen Auflösungsgrad" (Terhart 2002, S. 71) erreichen und Leistungsdaten auf unterschiedlichen Ebenen des Schulsystems generieren, kommen alle zum gleichen Ergebnis: Es gibt Zweifel an der Wirksamkeit des deutschen Bildungssystems, die mehreren Blickwinkeln entspringen:

- Aus *Sicht des deutschen Bildungssystems* erzielen andere wichtige Vergleichsländer auf der einen Seite deutlich bessere Leistungsmittelwerte und auf der anderen Seite ist es anderen Nationen zugleich möglich, die Streuung der erzielten Ergebnisse um diesen Leistungsmittelwert deutlich geringer zu halten (vgl. Deutsches PISA-Konsortium 2001 und OECD 2001). Es gelingt folglich nicht, auf nationaler, schulsystemischer Ebene vergleichbare Standards zu erzielen.
- Auch auf der *Ebene der Bundesländer* steht derzeit – erstmals auf der Grundlage einer ausreichend großen Stichprobe der Länder bei PISA-E – die Vergleichbarkeit und Qualität schulischer Arbeitsresultate zur Diskussion. Die Spannweite zwischen dem stärksten und dem schwächsten Land ist mit 53 bis 64 Testpunkten in den Kompetenzbereichen beachtlich. Hinsichtlich vieler Aspekte schulischer Leistungsresultate einschließlich wichtiger Kontextbedingungen mangelt es an Vergleichbar-

keit. Der internationale Bezug macht zudem für die einzelnen Länder deutlich, dass es noch großer Anstrengung bedarf, um international gute bis hervorragende Leistungen zu erreichen (vgl. Deutsches PISA-Konsortium 2002).

- Aus der *Perspektive der einzelnen Schulformen* wird ersichtlich, dass es trotz deutlicher Unterschiede zwischen den schulformspezifisch erreichten Mittelwerten breite Überlappungsbereiche gibt. Diese weisen dort auf vergleichbare schulische Resultate hin, wo aufgrund vorgeschalteter Selektionsmechanismen beim Übergang in die Sekundarstufe I keine derart vergleichbaren Ergebnisse erwartet werden (vgl. z.B. Deutsches PISA-Konsortium 2001). Zugleich legen sowohl die internationalen als auch die in einzelnen Bundesländern durchgeführten Ländervergleiche dar, dass gerade in Deutschland sozialstrukturelle Merkmale gegenüber Leistungskennzeichen an wichtigen Nahtstellen der Bildungskarrieren im Vordergrund stehen. Somit werden durch die Vergleichsstudien nicht allein die gemessenen Leistungswerte in den Mittelpunkt gerückt; im Fokus steht auch die Bewertung von Leistungen, die sich offensichtlich nicht an übergeordneten vergleichbaren Standards orientiert (vgl. z.B. Lehmann/Peek/Gänsfuß 1997 und 1999). Auf das Missverhältnis von gemessenen Leistungswerten, ihrer Bewertung in Form der schulischen Notengebung und der damit verbundenen Lebenschancen machte beispielsweise auch die Studie von Köller, Baumert und Schnabel zu Mathematikleistungen von Oberstufenschülern der gymnasialen Oberstufe an Gesamtschulen und an Gymnasien aufmerksam (vgl. Köller/Baumert/Schnabel 2000).
- Aus dem *Blickwinkel der Einzelschule* wird ähnliches ersichtlich: Auch den Schulen innerhalb einzelner Schulformen wird bescheinigt, dass die Unterschiede groß sind. Unter dem Gesichtspunkt der Verteilungsgerechtigkeit und der optimalen individuellen Förderung ergibt sich ein dringender Diskussionsbedarf.
- Für die *einzelnen Lehrkräfte* wurde im Rahmen einer kleineren explorativen Fallstudie im Rahmen von PISA diagnostiziert, dass sie die Kompetenzen ihrer Schülerinnen und Schüler gerade im unteren Leistungsbereich schlecht diagnostizieren. Nach Schrader und Helmke belegen Untersuchungen, dass die diagnostische Kompetenz als ein wesentlicher Kompetenzbereich professionellen Lehrerhandelns – vor allem das Vermögen, Leistungsunterschiede abzuwägen – eine Katalysatorfunktion für den Lern- und Unterrichtserfolg durch die Abstimmung diagnostischer Informationen mit didaktisch-methodischen Unterrichtsmaßnahmen hat (vgl. Schrader/Helmke 2001). Bei fehlenden oder falschen diagnostischen Informationen auf schülerindividueller Ebene ergeben sich aber in der Summe erhebliche Konsequenzen auch für das Bildungs- und Beschäftigungssystem insgesamt. Lehrkräfte sollen nämlich mit ihren diagnostischen Ergebnissen nicht nur individuelles Lernen optimieren (so genannte ‚Förderdiagnostik'), sondern sie sollen auch im gesellschaftlichen Interesse Lernergebnisse feststellen (so genannte ‚Selektionsdiagnostik') und den Übergang in verschiedene Lerngruppen, Kurse, oder Bildungswege nach vorgegebenen Kriterien verbessern (vgl. Ingenkamp 1997).

Paradigmenwechsel in Richtung Wirkungsqualität und -steuerung

Vor diesem Hintergrund liegt der Schwerpunkt der Diskussionen sowohl national als auch international mittlerweile vor allem auf der Qualität fachlicher und überfachlicher Wirkungen des Unterrichts. In diesem Zusammenhang spricht man von der ‚Wirkungs'- oder ‚Outputqualität'. Im Mittelpunkt der Debatten stehen vor allem kognitive Lerner-

träge. Neben einer Bestandsaufnahme schulischer Qualität in Verbindung mit der Frage nach der Definition und Bewertung von Qualität an sich geht es zudem im Kontext aktueller Autonomisierungs-Tendenzen auf der einen und der Globalisierungsentwicklungen auf der anderen Seite (vgl. Klemm 2000) verstärkt um die Frage der effektiven Steuerung von Qualität. Helmke fasst diese Entwicklungen im Primar- und Sekundarschulbereich, die im Hochschulbereich im Übrigen früher einsetzten und diskutiert wurden, unter dem Begriff des ‚Paradigmenwechsels' zusammen: Das derzeit vorherrschende erkenntnisleitende Forschungsinteresse sei durch – auch dies ist wieder ökonomisch formuliert – den Wandel von der „Input- zur Output-Orientierung" (Helmke 2000, S. 135 und Helmke 2001) charakterisiert. Diese Entwicklung gründet auf einer empirischen Basis und ist zugleich durch sie gekennzeichnet (vgl. zur ‚empirischen Wende' Lange 1999). Im Kontext dieses Paradigmenwechsels ist die nachfolgend beschriebene Ausweitung extern gesteuerter, großflächig angelegter Evaluationsformen zu sehen.

1. Ein wachsender Datenbestand durch zahlreiche und vielgestaltige großflächige Tests, Prüfungen und Leistungsstudien

Ein wichtiges Element der Bestandsaufnahme und Steuerung schulischer Qualität ist im Rahmen der aktuellen, an Wirkungsmerkmalen orientierten empirischen Schulforschung die externe Evaluation geworden. Formen externer Evaluation erlangen im Kontext des beschriebenen Paradigmenwechsels zusehends Prominenz. Entsprechend der in der Einleitung zum ersten Teil der Arbeit vorgenommenen Kategorisierung werden nachfolgend die unterschiedlichen Formen von Large Scale Assessments, wie sie derzeit in der Bundesrepublik existieren bzw. geplant sind, in ihrer Organisationsstruktur und den damit verknüpften Absichten vorgestellt:

- internationale Schulleistungsstudien mit deutscher Beteiligung,
- nationale Schulleistungsstudien in Deutschland (länderübergreifend und länderintern),
- Vergleichs- und Orientierungsarbeiten sowie Diagnosearbeiten und
- zentrale Abschlussprüfungen am Ende der Sekundarstufen I und II.

Schulinspektionen werden in der Bundesrepublik in keinem der Bundesländer als standardisierte, regelmäßige Schulbesuche durchgeführt. Sie sind nicht an national verbindliche Kriterienkataloge zur Bewertung der vorgefundenen Situation geknüpft und stellen hierzulande keine Form großflächiger externer Evaluation dar. Sie werden aus diesem Grund erst in den Kapiteln zu den drei Untersuchungsländern als weiteres Modell der externen Evaluation in ihrer Anlage und Wirkung aufgegriffen.

1.1 Beteiligung an internationalen Leistungsvergleichs-Untersuchungen

Schulleistungsuntersuchungen auf einer empirischen Basis können grundsätzlich folgende Anlagestruktur haben, die zugleich wesentliche Charakteristika dieser Form von Large Scale Assessments darstellen: