

ZENTRALE TESTS UND PRÜFUNGEN IM DIENSTE
SCHULISCHER ENTWICKLUNG:

Erfahrungen in Deutschland sowie Beobachtungen in England,
Frankreich und den Niederlanden

Summary: This paper is concerned with the description of the various types of central assessments and examinations that are conducted on a large scale, either on a regular or irregular basis, within four European countries. It also looks into the question of how the results from these different types of external evaluation are fed back to the schools and other institutions within the educational systems. The article goes on to look at the current research activities in the field, where both public and private information is used to assist in the development of schools. It starts with a look at the present position in Germany, where research into this matter has only recently begun and where, as a result of this, there is a comparatively little knowledge on the subject. Also covered within the paper is an overview of the test and examination systems that are being used in England, France and the Netherlands. These countries seem to be particularly interesting due to the integration of the assessment and examination processes into very different, historically grown, types of administration. What are the aims within the national context of these countries and what do the results of research tell us about the realisation of these intentions within the schools? Finally, in Germany there is a clear necessity for a fundamental debate on the targets, methods and strategies of using examination results for the improvement of schools if we do not want to reach an impasse with extended test procedures.

National wie international ist eine ausgeweitete Praxis verschiedener Formenshullaufbahnbegleitender Tests und zertifizierender Abschlußprüfungen zu beobachten. Sie gehen über die Leistungsmessung an einzelnen Schule hinaus, erfassen die Lernergebnisse zumeist vieler hundert Schulen und tausender SchülerInnen und haben somit einen ‚large scale‘- bzw. großflächigen Charakter. Der vorliegende Beitrag gibt eine Übersicht über solche Test- und Prüfungsformen in vier europäischen Ländern, eingebettet in eine Darstellung der jeweiligen Strukturen der nationalstaatlichen Test- und Prüfungssysteme. Im Mittelpunkt der Überlegungen steht die Frage der Verknüpfungsmöglichkeiten der Ergebnisse solcher Leistungsmessungen mit der Entwicklung von Schulen:

- Welche Intentionen lassen sich auf bildungsadministrativer Ebene im Hinblick auf schulische Strategien der Datenrezeption und -nutzung erkennen?
- Welche Forschung gibt es in den Ländern zu den tatsächlichen Strategien einzelner Schulen?

Hinsichtlich der Untersuchung dieser Aspekte zeigt der Blick auf die deutsche Situation ein Forschungsdesiderat auf. In den exemplarisch gewählten Ländern England, Frankreich und den Niederlanden hingegen ist man vertrauter mit entsprechenden Handlungs- und Sachzusammenhängen, da der Trend ausgeweiteter Test- und Prüfungsverfahren dort bedeutend früher als in Deutschland einsetzte. Der Status quo in diesen Ländern wird nachfolgend beschrieben sowie abschließend miteinander verglichen, und es werden sich daran anschließende Forschungsaufgaben skizziert.

Vergleichende Forschungsmethoden werden bisher auf dem Gebiet der Rezeptions- und Nutzungsstrategien im Rahmen großflächiger Tests und Prüfungen in Deutschland kaum genutzt. Dabei beschreibt z.B. MITTER die generelle Funktion des Vergleichs als „einen der fundamentalen Wege zur Gewinnung und zur Erweiterung von Erkenntnissen sowie zur Erhöhung des Kenntnisniveaus“ in vielen Bereichen inner- und außerhalb der Erziehungswissenschaft (MITTER 2001, S. 91). Auch die Schulleistungs- und -entwicklungsforschung können aus internationalen Erfahrungen lernen. Aus der Perspektive der deutschen Komparatistik hebt MITTER die Bedeutung aktueller Vergleichsuntersuchungen hervor, wenn er sagt, „daß zur Erweiterung der Dimensionalität in der Vergleichenden Erziehungswissenschaft auch der bisher außerhalb der ‚etablierten‘ Disziplin gebliebene internationale Schulleistungsvergleich als Untersuchungsgegenstand wachsender Aufmerksamkeit wert ist“ (MITTER, 2001, S.97).

1. Erfahrungen in Deutschland und im Umgang mit großflächigen Leistungsmessungen

Die schulische Qualitätsdebatte konzentrierte sich in Deutschland viele Jahre lang auf die sogenannte Kontext- und Prozeßqualität, nämlich auf

- Merkmale der gegebenen schulischen und außerschulischen Umwelt einschließlich der materiellen und personellen Ausstattung der Schulen (Kontext- oder Inputqualität) sowie auf
- Merkmale des Unterrichts, der Klassenführung und der LehrerIn-SchülerIn-Interaktion (Prozeßqualität).

Mittlerweile liegt der Schwerpunkt national wie international auf der Qualität fachlicher und überfachlicher Wirkungen des Unterrichts (Wirkungs- oder Outputqualität), vor allem aber auf kognitiven Lernerträgen. Neben der Beschäftigung mit schulischer Qualität an sich geht es vor dem Hintergrund aktueller Autonomisierungstendenzen und Globalisierungstendenzen (vgl. KLEMM 2000) verstärkt um die Frage ihrer effektiven Steuerung. Helmke faßt die Entwicklungen im Primar- und Sekundarschulbereich unter dem Begriff des „Paradigmenwechsels“ zusammen: Das derzeit vorherrschende erkenntnisleitende Forschungsinteresse sei durch – ökonomisch formuliert – den Wandel von der „Input- zur Output-Orientierung“ (HELMKE 2000, S. 135/HELMKE 2001) auf einer empirischen Basis charakterisiert (vgl. zur „empirischen Wende“ LANGE 1999).

Ein wichtiges Steuerungselement schulischer Qualität ist dabei die Evaluation. Die Etymologie dieses Begriffs läßt eine Verbindung zum ‚value‘-Begriff erkennen, der wiederum die Nützlichkeit und Qualität von etwas, aber auch das Ergebnis einer Messung meint. Tests und Prüfungen, die zentral z.B. durch eine Behörde oder/und eine universitäre Forschungsgruppe in ihrer Entwicklung, Durchführung und Auswertung gesteuert werden, stellen eine Form externer Evaluation dar. Mit diesen Zentraltest und -prüfungen wird der Nutzen und die erreichte Qualitätsstufe schulischen Unterrichts erfaßt, analysiert und bewertet. Sie erlangen im Rahmen des beschriebenen Paradigmenwechsels zusehends Prominenz, was sich an der Vielzahl der bundesländerinternen und -übergreifenden Schulleistungstests wie LAU, dem Bayerischen Mathematiktest, MARKUS und QuaSUM sowie BIJU, PISA-E und DESI ablesen läßt (vgl. zur Erklärung der Akronyme Abb.1). Diese wurden in der Folge der internationalen Large-Scale-Assessments mit deutscher Beteiligung (CivEd, PISA, PIRLS, TIMSS...) zu deren Ergänzung entwickelt (vgl. die Übersicht bei van ACKEREN/KLEMM 2000, S. 12–13).

Bei der Darstellung des derzeit ausgeweiteten Interesses an externen Evaluationsformen wird allerdings selten erwähnt, daß es einen zweiten wichtigen Evaluations-„Strang“ gibt, bei dem ebenfalls Externe von zentraler Stelle aus bereits seit Jahrzehnten in beinahe der Hälfte der Bundesländer schulische Arbeitsergebnisse beurteilen. Diese Form schulischer Evaluation ist unter dem Schlagwort ‚Zentralabitur‘ am bekanntesten (vgl. KLEMM 1998). In den sieben Bundesländern, die diese Prüfungsform derzeit praktizieren, bearbeiten die SchülerInnen in bestimmten Schulformen und Fächern auch zum Abschluß der Sekundarstufe I zur gleichen Zeit die gleichen Aufgabenstellungen. Das Thema dezentraler versus zentraler Organisation regelmäßiger Abschlußprüfungen hat man bei der aktuellen Qualitätsdiskussion vergleichsweise weniger stark als die schullaufbahnbegleitenden Leistungstests im Blick. Die Diskussionen – vor allem um das Zentralabitur und seine Funktion hinsichtlich Qualitätssicherung und Vergleichbarkeit – werden

eher auf bildungspolitischer denn auf wissenschaftlicher Ebene geführt. Eine Literaturrecherche zeigt, daß sich die politischen Meinungsbilder kaum auf entsprechende wissenschaftliche Forschung stützen können. Lediglich im Rahmen der TIMS-Studie beschäftigt man sich eingehender mit dem Zusammenhang von Prüfungsform und erzielter Leistung aus deutscher Perspektive – speziell im Kontext von Qualitätsentwicklung und -sicherung, weniger im Bereich der Vergleichbarkeits-Frage (vgl. BAUMERT/WATERMANN 2000). Mit dieser Teiluntersuchung wurden beide Typen externer Evaluation in dem Sinne miteinander verknüpft, daß das Evaluations-Instrument ‚Zentralprüfung‘ mit Hilfe der TIMSS-Ergebnisse evaluiert wurde.

Zentral gesteuerte, großflächig angelegte Leistungsüberprüfungen sind in allen deutschen Bundesländern ein aktuelles Thema – ob in Form der beschriebenen regelmäßigen Zentralprüfungen oder in der Gestalt vielfältiger Large-Scale-Assessments. Sie finden Eingang in zumeist voneinander getrennte Debatten, was an den intentionalen Unterschieden liegen mag:

- Zentralprüfungen unterliegen in Deutschland im wesentlichen der Zertifizierung eines qualifizierenden Abschlusses im Berechtigungswesen.
- Leistungstests dienen vor allem der Standortbestimmung und der Rechenschaftslegung durch ein Systemmonitoring sowie zunehmend der Identifizierung schulischen Handlungsbedarfs.

Im Rahmen der letzteren Evaluationsform werden die Testresultate nicht mit der individuellen SchülerInnenbeurteilung verknüpft. Eine Ausnahme stellen an dieser Stelle der bayerische Mathematik- und Deutschttest dar (vgl. z.B. www.isb.bayern.de/ghs/bmt.htm). Hier fließen Testergebnisse in die Jahresfortgangsnote ein. Umgekehrt werden die Ergebnisse zentraler Abschlußprüfungen nicht als Ausgangspunkt für das Sichtbarmachen ungleicher Ergebnisse und möglicher erklärender Zusammenhänge genutzt. Ein solches Handlungsinteresse wird derzeit nur im Kontext der Schulleistungsstudien verfolgt. Bei ihrer chronologischen Entwicklung ist ebenfalls – dies gilt zumindest für die länderinternen Untersuchungen – ein Paradigmenwechsel festzustellen: von der Datenbereitstellung für systemisches Steuerungswissen hin zu einer additiven Ermittlung und Bereitstellung von Ergebnissen im Hinblick auf einzelschulische Nutzungsmöglichkeiten. Dementsprechend werden bei Untersuchungen wie LAU, QuaSUM oder MARKUS nicht bloß Lernstandsmessungen in einzelnen Fächern durchgeführt, sondern es werden mit Hilfe von Fragebögen individuelle, klassenbezogene, schulische- und außerschulische Kontextmerkmale erhoben. Sie werden mit den Leistungsergebnissen durch statistischen Methoden ‚verrechnet‘, um ein faires Leistungsbild sowie Hin-

weise auf mögliche Ursache-Wirkungs-Zusammenhänge zu erhalten und die Daten somit für die jeweilige schulische Institution handlungsrelevant zu machen.

Ähnliche Absichten lassen sich in europäischen Nachbarländern nicht nur im Kontext von Schulleistungsstudien, sondern ebenso für den Bereich regelmäßiger schullaufbahnbegleitender Zentraltests, die wir in dieser Form in Deutschland nicht kennen, und im Rahmen zentraler Abschlußprüfungen identifizieren. Hier werden auch für dieses Evaluationsverfahren mehr oder weniger differenziert Hintergrundvariablen erhoben, mit Test- und Prüfungsergebnissen korreliert und den Schulen, oftmals auch der Öffentlichkeit mitgeteilt. Hinter den nationalen Test- und Prüfungswesen stehen jeweils unterschiedliche, historisch gewachsene Absichten. Es läßt sich jedoch immer wieder das Grundprinzip regelmäßiger Tests und Prüfungen auf der einen und zumeist unregelmäßiger und in dieser Form oftmals jüngerer Tests der Gattung ‚large-scale‘ auf der anderen Seite erkennen. Elemente beider Evaluationsformen sind im internationalen Vergleich unterschiedlich stark verknüpft. Dies wird nachfolgend mit Blick auf die für die innerdeutsche Debatte besonders interessanten Länder England, Frankreich und die Niederlande verdeutlicht. Solche Vernetzungen in anderen Ländern haben sich aus einem verstärkten Interesse an einer intensivierten Zusammenarbeit von Schulleistungsforschung und Schulentwicklungsforschung mit dem Ziel, empirische Daten handlungsrelevant zu machen, entwickelt. Die Frage wird auch in Deutschland in jüngster Zeit aufgeworfen (vgl. u.a. KLEMM/SCHRATZ 2001, KLIEME/BAUMERT/SCHWIPPERT 2000, MARITZEN 2001, PEEK 2001, ROLFF 2001, TILLMANN 2001), doch mangelt es bislang an Forschungsergebnissen, um entsprechende Antworten geben zu können. Abbildung 1 gibt einen Überblick über intendierte Datennutzungs-Strategien und die Forschung über ihre Realisierung vor Ort im Rahmen der Studien LAU, QuaSUM und MARKUS. Bei diesen Untersuchungen hat man bisher eingehender über schulindividuelle Rückmeldungen und Nutzungsmöglichkeiten nachgedacht und tut dies auch noch. Es wird aber zudem deutlich, daß es nur zwei Forschungsprojekte sowie eine kleinere Studie von KLUG/REH und so gut wie keine Ergebnisse zu diesem Thema in Deutschland gibt. Bezogen auf die deutsche Situation bleibt ein Forschungsdesiderat festzuhalten: Es bedarf der Erschließung eines neuen Forschungsfeldes, in dem die Rezeptions- und Nutzungsstrategien der Schulen im Anschluß an Ergebnisrückmeldungen untersucht werden, um administrative und einzelschulische Handlungsstrategien im Sinne schulischer Entwicklung zu optimieren.

Abbildung 1 (Isabell van Ackeren): Ergebnis-Nutzung im Kontext schulischer Leistungsuntersuchungen in Deutschland am Beispiel von LAU, QuasUM und MARKUS (Stand: 1/12/2001)

Bezeichnung der Studie Bundesland, Erhebungsjahr, Altersgruppe	Intentionen		Realisationen	
	intermediäre Strategien auf Systemebene (vgl. u.a. die Ergebnisberichte)	intermediäre Strategien auf einzel-schulischer Ebene (vgl. u.a. die Ergebnisberichte)	Studien/Veröffentlichungen	Evolution von Nutzungsstrategien
LAU Aspekte der Lernungs-gänge und der Lernentwicklung Hamburg 1996, 1998, 2000 5., 7., 9. Klassen (früher)	Schaffung verbesserter Grundlagen für die Beratung und Unterstützung der Schulen durch das Amt für Schule - Nutzung der Ergebnisse für die Lernentwicklung - für die Lehrerfort- und Weiterbildung - für die Diskussion von Standards (gilt auch für QuasUM, vgl. Peck 2001, S. 62)	Anwendung der literarischen Rückmeldungen im Kontext des Gesamtberichts durch die Kollegen mit Unterstützung der Schulfachsicht und auf weitere Experimenten im Hinblick auf Konsequenzen für die pädagogische Arbeit, z.B. für - Stärkung der Klusurvorbereitung, - Kriterien der Notengebung - Entwicklung von Förderstrategien - die Schulprogrammentwicklung dennoch: Umgang mit Daten liegt in der Verantwortung der Schulen	Interviews mit den Schülern von acht Hamburger Gymnasien zur Rezeption und Nutzung der LAU-7-Ergebnisse (Klug/Rh 2000): - Welche Personen in den Schulen beschäftigen sich mit diesen Umständen mit dem Daten? - Gibt es Konflikte? - für die Lehrerfort- und -Weiterbildung - Was sieht das Elterninteresse aus? - Konkrete Handlungs-konsequenzen? ... Eine systematische Evaluation steht aber noch aus	unterschiedliches Interesse an weiterem Ausmaß der Nutzung mit Daten (verstärktes Interesse meist bei pflichtigen Abweichungen) - Probleme bei der Bewertung der Daten und der Ableitung von Handlungs-konsequenzen - Beibehaltung der schulischen Diskus-sionen über Unterrichtsqualität generell, aber: Laubbäume werden durch Differenzierung gefördert, Unterrichtszeit
QuasUM Qualitätsentwicklung an Schulen zum Unterricht in Mathematik Bremen 1999; 3. und 6. Klassen	Beschäftigung der Schulleitung mit den zentralen Lernzielen und den Hinweisen auf Konsequenzen, welche die Leistungs-gänge beeinflussen - bessere Diskussionsgrundlage für die öffentliche Debatte	Interdisziplinäre, intensive Diskussion in den Schulen, z.B. über - Klusurvorbereitung, - Notengebung, - Förderstrategie und - weitere Evaluation Dennoch einschicken Schulen selbst über Berechtigten Form und Umfang der Auseinandersetzung mit den Daten.	Projekt „Schulminimierung und Schulentwicklung“ - Untersuchung zum Umgang mit den Daten für die Entwicklung von Schulen: - Evaluation schulischer Rezipienten- und Nutzungsstrategien (Formen und Inhalte) - Ziel: Methodische Systematisierung schulischer Nutzung einzelner Evaluationsergebnisse - zusätzlich: qualitative Fallstudie an sechs Gymnasien	Das Projekt orientiert sich Untersuchungs-gegenstände der schulischen Auseinandersetzung mit entsprechenden Informationen - Erste Ergebnisse bleiben noch abzu-warten
MARKUS Markusmittels-Orientierung Rheinland-Pfalz, Kompetenzen, Unterrichts-konzepte, Schulstruktur Rheinland-Pfalz 2000 8. Klassen	Breites Qualitätsmanagement in den Bereichen - „Unterricht“ - „Personal“ und - „Organisation“ in zentralen Modulen der Lehrerfort- und -weiterbildung, aber auch schon in beiden Phasen der Lehrerfort- und -weiterbildung	Standortbestimmung mit Blick auf Fachleistungen und Kontextbedingungen als Wissensgrundlage für die zielorientierte Schul- und Unterrichts-entwicklung auch in Zusammenarbeit mit anderen Schulen	Projekt WALZER („Wirktungsanalyse der Leistungs-evaluation: Zielerreichung, Ertrag für die Bildungsqualität der Schule und die Rückmeldung von Evaluations-ergebnissen“) Kombination von quantitativen (Survey-) und qualitativen (Tiefeninterviews, Schul-portraits) Methoden	bisher noch keine direkten Veröffent-lichungen aus dem Projekt

Quelle: Informationsgesprächen, Pressenachrichten, Zwischen- und Ergebnisberichte der angeführten Studien sowie Publikationen von Klug/Rh 2000, Maritzen 2001, Peck 2001 und zum Projekt „Schulminimierung und Schulentwicklung“ die Internetseite: <http://www.zar.uni-bonn.de/infod/empir/66/Markus.htm>

2. Beobachtungen aus dem Ausland im Kontext externer Evaluationen

Einige unserer Nachbarn verfolgen eine breite, wenn auch jeweils unterschiedliche Praxis der Rückmeldung im Kontext verschiedener Test- und Prüfungsverfahren, die auf langjährigen Erfahrungen basiert. Die Strategien der Ergebnisnutzung in England, Frankreich und den Niederlanden erscheinen mit Blick auf die im europäischen Vergleich jeweils einmaligen Verwaltungstypen besonders interessant:

- Das englische Beispiel zeigt die Einbindung eines zentralisierten Prüfungswesens in eine marktorientierte Schulentwicklung eines ansonsten traditionell lokal verwalteten Systems.
- Mit Frankreich wird ein stark zentralistisch organisiertes Land in die Betrachtung einbezogen, an dem sich die Intention einer auf Gleichheit setzenden republikanischen Schulpolitik herausarbeiten läßt.
- Die Niederlande stehen schließlich für den Versuch, die Vielfalt einer in hohem Maße privat getragenen, durch Autonomie gekennzeichneten Schullandschaft mit dem Interesse vergleichbarer Schulleistungen in Einklang zu bringen.

In diesen drei Ländern kann ein gemeinsamer Problemansatz als Vergleichsgrundlage und übergeordnete Bezugsgröße identifiziert werden: die Schwierigkeit der bildungsadministrativ intendierten Ergebnisnutzung auf der einen und der einzelschulischen Realisierung auf der anderen Seite. Dabei ist der vorliegende, problembezogen angelegte Vergleich, der einen Ausschnitt der nationalen Bildungssysteme erfaßt, zugleich auch im Zusammenhang der nationalstaatlichen, historischen Gewachsenheit zu sehen. So muß die spezifische Fragestellung der Datennutzung in den Kontext des jeweiligen Bildungssystems gestellt werden, ohne dabei eine systemische Totalanalyse zu leisten. Im folgenden werden deshalb wesentliche nationale Entwicklungen skizziert, um Trends im Test- und Prüfungswesen, speziell im Bereich von Optimierungsstrategien im Sinne einer verbesserten einzelschulischen Entwicklung, nachvollziehbar einordnen zu können.

2.1 England – Strukturen der Datennutzung im Kontext marktorientierter Entwicklungen

Ausgangspunkt radikaler Veränderungen in der englischen Schullandschaft war der unter der konservativen Regierung THATCHER verabschiedete Education Reform Act von 1988. Er war von der Überzeugung getragen, daß ein Wettbewerb zwischen Schulen eine Verbesserung qualitativer Standards und in der Konse-

- die ‚National Literacy and Numeracy Projects‘ aber auch
- Value-Added-Projekte in LEAs (‚Local Education Authorities‘) wie Lancashire, Surrey und Hampshire.

Zudem gibt es folgende schulische Informationssysteme:

- ‚Performance Indicators in Primary Schools‘ (PIPS),
- das ‚Year Eleven Information System‘ (Yellis),
- das ‚Middle Years Information System‘ (MidYIS) sowie
- das ‚A Level Information System plus Advanced GNVQ‘ (ALIS).

Die zuletzt genannten Projekte werden unter der Leitung des ‚Curriculum, Evaluation and Management Centre‘ (CEM) an der Universität Durham durchgeführt. Hierbei können Schulen – und nicht nur, wie sonst üblich, Institutionen auf bildungsadministrativer Ebene – selbständig eine wissenschaftliche Untersuchung ihrer Lernergebnisse in Auftrag geben:

„We develop, produce and provide tests and questionnaires to be completed by students under standardised conditions. We analyse these and provide clear graphical feedback, and comparisons with many hundreds of other schools and colleges. Data on pupil progress (value added) is provided when outcome measures become available. At each stage we try to measure what matters, be it attitudes, safety, relationships, learning and teaching processes etc.“ (CEM-Homepage: <http://cem.dur.ac.uk/>).

Schulen in England sind folglich so datenreiche Institutionen, wie man es sich in Deutschland kaum vorstellen kann. Sie werden oftmals auch als ‚data driven‘ bezeichnet, wobei damit ein nächster Schritt im Evaluationsprozeß angesprochen ist, nämlich der Umgang mit den Resultaten. Die mehr oder weniger differenzierten Informationen sollen unmittelbare Optimierungsstrategien am Schulstandort ermöglichen. Daneben meint der Ausdruck ‚data driven‘ aber auch eine indirektere, stärker vermittelte als unmittelbar resultierende Entwicklungsbewegung auf schulischer Ebene. Diese vor allem bildungspolitische Intention ist bei der Betrachtung der Entwicklungen im englischen Test- und Prüfungssystem zeitlich früher anzusetzen: Seit der Bildungsreform von 1988 wird den Eltern schulpflichtiger Kinder die Möglichkeit der freien Schulwahl ohne Bindung an Einzugsgebiete (‚catchment areas‘) eingeräumt, was im englischen Sprachraum als ‚open enrolment‘ bezeichnet wird. Es gibt bisher nur wenige Untersuchungen, die der Frage nachgehen, wie und warum sich Eltern für bzw. gegen eine Schule entscheiden und

welchen Einfluß veröffentlichte Ergebnisse und Kontextdaten dabei haben. Beabsichtigt ist eine Kennzeichnung ‚guter‘ Schulen mittels publizierter Leistungsergebnisse als Referenzrahmen, die von Eltern interpretiert werden und in ihre Schulwahlentscheidung einfließen sollen, so daß ‚good schools‘ einen besonders starken Zulauf haben. Damit differieren die Zahl der SchülerInnen und in der Konsequenz auch die daran gekoppelten Finanzmittel: Seit 1988 erfolgt die Berechnung der einzelschulischen Finanzressourcen viel stärker als zuvor indikatorenbezogen auf einer Per-Capita-Grundlage (vgl. dazu STÜBIG 1998). Um am Bildungsmarkt bestehen zu können, setzt in der Folge ein scharfer Wettbewerb um SchülerInnen ein, wobei Schulen in Problemgebieten von vorneherein benachteiligt sind. Diese Schwierigkeit hat man mittlerweile erkannt und sucht nach z.B. finanziellen Ausgleichsmöglichkeiten.

Grundlegendes Ziel bleibt aber die extrinsische Motivierung gefährdeter Schulen, ihr Niveau zu steigern und zur nationalen Standardsicherung beizutragen. In englischen Schulen erfolgt Schulentwicklung oftmals im Kontext dieser marktähnlichen Strukturen, die durchaus als Motor schulischer Entwicklung geschaffen wurden. Folgenden Maßnahmen der Schulen als Reaktion auf ihre aktuelle Marktposition sind u.a. in der Literatur beschrieben:

- verstärkte Werbemaßnahmen für die Schule etwa durch ‚glossy‘ Schulbrochüren,
- der Ausschluß unterrichtsstörender SchülerInnen,
- besondere Fördermaßnahmen für SchülerInnen, die sich auf der Grenze zwischen guter und schlechter Leistung befinden und als ‚borderliner‘ bezeichnet werden,
- Zulassungsbeschränkungen zu Prüfungen für leistungsschwache KandidatInnen oder
- der Wechsel zu einer Awarding Body, die aus Sicht der Schule einfachere Test- und Prüfungsmaterialien entwickelt (vgl. FOXMAN 1997).

Seit der verbesserten Versorgung der Schulen mit detaillierteren Informationen soll – so die dahinter stehende Absicht – auch die Entwicklungsplanung an den Schulen besser zu steuern sein. Fraglich ist jedoch, ob und wie die rückgemeldeten Daten vor Ort genutzt werden. Forschungslücken in diesem Bereich werden vielfach beschrieben (vgl. z.B. SAUNDERS 1999, SCHAGEN 1998) und damit erklärt, daß einem Großteil englischer Schulen erst seit kurzem ‚value-added‘-Indikatoren zur Verfügung stehen. „[...] whilst it is true that value added measures of performance have been developed in methodological and analytical terms, the other end of the process, so to speak – of support for schools in managing and using such

data – is still relatively under-researched“ (SAUNDERS 1999, S. 64). Die bisher vorhandenen, wenigen Studien beziehen sich sowohl auf die Ergebnisnutzung im Rahmen der regelmäßigen Assessments und Examinations (z.B. FITZ-GIBBON 1997 und ASHBY/SAINSBURY 2001) als auch auf regional begrenzte Zusatzstudien (z.B. SAUNDERS/RUDD 1999, die sich auf QUASE beziehen). Die Ergebnisse können an dieser Stelle nicht im einzelnen referiert werden, doch ergibt sich bei der Durchsicht der Forschungsliteratur der Eindruck, daß Schulen sehr unterschiedlich mit den Resultaten umzugehen wissen. Dies betrifft z.B. die curriculare Planung, die Nutzung für die Schulprogrammarbeit oder auch leistungsdifferenzierende Maßnahmen. Es wird auch deutlich, daß mehr Hilfen im Umgang mit immer mehr und immer komplexeren Daten benötigt werden. Eine detaillierte Wahrnehmung, Analyse und Weiterverfolgung der englischen Erfahrungen ist lohnend mit Blick auf die weitere Planung einer deutschen Evaluationspolitik, denn dort versucht man u.a. folgende, wichtige Fragen zu beantworten: „Is the effort worthwhile? Do value-added approaches help to improve education in practice?“ (Saunders 1999, S. 64). –

2.2 Frankreich – Datengenerierung und Datennutzung innerhalb einer zentralisierten und auf Gleichheit setzenden Schulpolitik

„On a visit to a school in the provinces the French Education Minister is reputed to have said ‚Today is Tuesday and this is the second afternoon period, so they must be doing home economics‘“ (BONNET 1997, S. 295). Diese nicht nur in Frankreich bekannte Anekdote spielt auf den traditionellen Zentralismus des französischen Bildungswesens an, der sich u.a. in einem national verbindlichen Curriculum und nationalen Tests und Prüfungen äußert. Dennoch sind auch in Frankreich Dezentralisierungstendenzen zu erkennen, die sich etwa in einer erweiterten schulischen Flexibilität im Umgang mit Zeitressourcen oder der individuellen Akzentuierung von Lehrplaninhalten äußert (vgl. zur Frage der Dezentralisierung in Frankreich DUTERCQ 2000). Im Bereich des französischen Test- und Prüfungswesens hat zudem die fortlaufende Bewertung durch die individuelle Lehrkraft immer noch eine enorme Bedeutung, zumal der Einfluß des sogenannten ‚brevet des collèges‘ am Ende der Sekundarstufe I im Vergleich zur herausragenden Bedeutung des abituradäquaten ‚baccalauréat‘ geschmälert ist: Das Bestehen der zentralen Prüfungen in Mathematik, Französisch, Geschichte und Geographie, deren Ergebnisse zusammen mit den Noten der letzten beiden Schuljahre in das Brevet einfließen, ist nicht notwendig, um die Oberstufe besuchen zu können, obgleich ein solches Zertifikat auf dem Arbeitsmarkt Bedeutung hat. Von vielen SchülerInnen wird die Prüfung als gute Übung von Examenstechniken mit Blick auf das

Baccalauréat gesehen, das BONNET als „ultimate school leaving certificate and the compulsory passport to higher education“ beschreibt (vgl. BONNET 1997, S. 297). Im übrigen wird selten erwähnt, daß die Examensaufgaben zwar nach zentralen Rahmenvorgaben, aber in bestimmten Regionen individuell entwickelt werden. So bearbeiten etwa beim Baccalauréat alle KandidatInnen landesweit zur gleichen Zeit nicht unbedingt gleiche, aber dennoch vergleichbare Aufgaben. Auch bei der Aus- und Bewertung ist man in Frankreich nicht ganz so zentral organisiert, wie vielleicht erwartet: Die Prüfungsbögen werden extern durch Lehrende anderer Schulen der Region ausgewertet, nicht aber durch eigens dafür ausgebildete PrüferInnen. Die Bedeutung des Zentralitäts- bzw. Dezentralisierungsthemas ist in den Ländern unterschiedlich gewichtet und differenziert, vor allem im Hinblick auf damit verbundene Absichten, was sich im Test- und Prüfungswesen widerspiegelt.

Die zentralen französischen ‚examens‘ haben im wesentlichen einen summativ-zertifizierenden Charakter, obwohl auch in Frankreich die Brevet- und Baccalauréat-Ergebnisse regelmäßig sowohl durch die Presse (z.B. www.lemonde.fr/education/) als auch von behördlicher Seite publiziert werden. Die offiziellen Veröffentlichungen sind wohl auch eine korrigierende Reaktion auf die ersten Pressepublikationen nicht-adjustierter Prüfungsergebnisse im Jahr 1982. Sie liegen in der Hand einer 1987 für die Evaluation des Schulsystems gegründeten Abteilung innerhalb des Bildungsministeriums. Sie heißt mittlerweile ‚Direction de la Programmation et du Développement‘ (DPD) und hat weitere bildungsökonomische und -statistische Aufgaben (www.education.gouv.fr/dpd/). Das Hauptziel der zentralen Prüfungen besteht darin, die Bildungschancen aller SchülerInnen überall in Frankreich dadurch anzugleichen, daß die Leistungsfähigkeit der Schulen möglichst dicht beieinander gehalten wird. Die soll nicht zuletzt durch diesen permanenten und öffentlichen Vergleich untereinander geschehen, der mittlerweile ‚fairer‘ gestaltet wird: Es werden Erwartungswerte (‚taux attendu‘) unter Einbeziehung sogenannter ‚catégories socio-professionnelles‘ der Eltern und des Alters der SchülerInnen be- und verrechnet sowie als Mehrwert- bzw. ‚valeur ajoutée‘-Informationen mitgeteilt. Inzwischen versucht man auch in Frankreich früher erreichte Leistungen in die Mehrwert-Berechnungen einzubeziehen. Die Veröffentlichungen führen zwar tendenziell zu einem Wettbewerb unter den Schulen, nicht aber zu einer nennenswerten Umorientierung der Schulanmeldungen. Der Grund hierfür liegt in den festen Schulbezirken, die in den meisten Regionen seit der ‚carte scolaire‘ zu Beginn der sechziger Jahre existieren. Deshalb findet man in Frankreich nicht die marktähnlichen englischen Strukturen wieder, obgleich seit den achtziger und neunziger Jahren verstärkt Schulwahlmöglichkeiten eingeräumt werden (vgl. z.B. van ZANTEN 1996). BONNET beschreibt zudem die Veröffent-

lichung von Resultaten als Ausdruck einer intendierten Versachlichung notwendiger und durch die Publikationen initiiertes Diskussionen über Standards unter allen, direkt oder indirekt an Erziehungs- und Bildungsprozessen beteiligten Akteuren. Den Schulen selbst stehen seit 1995 noch differenziertere ‚valeur ajoutée‘-Informationen zur Verfügung als in den öffentlichen Rankings, die auch als ‚palmarès‘ bezeichnet werden, enthalten sind. Aufgrund der immer noch stark zentralisierten Strukturen im französischen Bildungswesen sind auf nationaler Ebene viele Informationen zur Lehrer- und Schülerschaft verfügbar. Auf diese Weise können den Schulen 21 Indikatoren, sogenannte ‚indicateurs pour le pilotage des établissements scolaires‘ (IPES), zur Verfügung gestellt werden. Sie sollen ihnen als Referenzrahmen für ihre eigenen, computergestützten Auswertungen dienen, um sich mit anderen Schulen auf lokaler, regionaler und nationaler Ebene vergleichen und Handlungskonsequenzen ableiten zu können. Diese Indikatoren beziehen sich im wesentlichen auf die in Frankreich so definierten vier Kategorien ‚Inputindikatoren‘ (SchülerInnenmerkmale), ‚Outputindikatoren‘ (Test- und Prüfungsergebnisse), ‚schulische Ressourcen und Merkmale des Schulmanagements‘ sowie auf ‚außerschulische Hintergrundvariablen‘ (vgl. DEMAILLY u.a. 1998, S. 84 ff.). „Die DEP hat die Absicht, daß dies ein Mittel der Selbstevaluation für die Schulleitung ist [...]“ (DEMAILLY u.a. 1998, S. 29, aus dem Französischen übersetzt, l.v.A.).

Neben den genannten Abschlußprüfungen kennt man auch in Frankreich schullaufbahnbegleitende Prüfungen (vgl. zu diesen Evaluationsformen z.B. THÉLOT 1993 und VOGLER 1996). Seit dem Reformgesetz von 1989, dem sogenannten ‚loi d’orientation‘, ist man verstärkt um eine detailliertere Evaluation der SchülerInnenleistungen mittels spezieller, auf das nationale Curriculum abgestimmter Tests bemüht. Im Zuge dessen werden alle SchülerInnen seit 1989 im Alter von acht und elf Jahren (CE2 bzw. ‚cours élémentaire‘ und 6ème) in Französisch und Mathematik zu Beginn eines Schuljahres ‚en masse‘ getestet. Dahinter steht die Absicht, eine ‚culture de l’évaluation‘ in den Schulen zu fördern und den Lehrkräften die Konsequenzen ihres unterrichtlichen Handelns wie einen Spiegel vor Augen zu führen (vgl. BONNET 1997, S. 299). Ein ähnlicher Test wurde 1992 auch für die 16jährigen in der ‚seconde‘ zu Beginn des ‚lycée‘ eingeführt, der neben Französisch und Mathematik auch die ersten großen modernen Fremdsprachen (Englisch, Deutsch oder Spanisch) testet. Somit wurden jährlich jeweils etwa 800.000 SchülerInnen auf den unteren beiden Altersstufen und etwa 500.000 ältere SchülerInnen getestet. Da den LehrerInnen mittlerweile genügend Materialien zur Verfügung stehen, um eigene Tests durchzuführen, aber auch aus Kostengründen, wird abwechselnd nur noch eine Alterstufe jährlich getestet. Als schullaufbahnbegleitende Leistungsmessungen haben diese Tests keinen zertifizierenden, sondern eher

formativ-diagnostischen Charakter für die Lehrkräfte und Eltern innerhalb einer Schule. Es werden zwar nicht sämtliche extern entwickelten Tests auch extern ausgewertet, doch gilt dies für eine nationale, repräsentative Stichprobe auf allen drei Altersstufen. Ziel ist ein nationales Systemmonitoring – etwa im Bereich der Identifizierung besonderer Schwächen der getesteten Fachinhalte –, das aber auch im Rahmen zusätzlicher, stichprobenartiger Leistungsuntersuchungen erfolgt. Neben den reinen Leistungstests erhalten die Schulen Schulleitungs-, LehrerInnen- und SchülerInnen-Fragebögen, um Hintergrundmerkmale zu erfassen. In Frankreich ist jede Schule selbst für die Analyse der Daten mittels einer speziellen Software verantwortlich, was die Intention unterstreicht, die Daten direkt für die Unterrichtsplanung zu nutzen. Die Resultate können für eine Standortbestimmung zudem mit den öffentlich einzusehenden nationalen Stichprobenergebnissen (<http://msg.education.gouv.fr/Evace26/debut.htm>) – auch unter Einbeziehung der IPES (s.o.) – verglichen werden. Die Schulen, die in das nationale Sample einbezogen waren, werden über ihre Ergebnisse, die der Öffentlichkeit nicht zugänglich gemacht werden, informiert und erhalten zudem einen pädagogischen Kommentar. Sie haben über diese Tests hinaus Zugriff auf eine Datenbank des Ministeriums mit konkreten Aufgaben, die laufend im Unterricht eingesetzt werden können.

Neben dieser Form regelmäßiger Tests und Prüfungen gibt es in Frankreich – wie bereits angedeutet – zudem eher unregelmäßig bzw. nicht jährlich durchgeführte, stichprobenartig, aber dennoch großflächig angelegte Schulleistungsuntersuchungen. Zu diesen Studien gehört ein Sample Testing, das sich seit 1984 etwa alle fünf Jahre auf im Durchschnitt 15jährige SchülerInnen der ‚troisième des collèges‘ bezieht. Es werden die Fachleistungen aller unterrichteten Fächer aber auch schulische Einstellungen und Werte untersucht. Daneben gibt es Längsschnittstudien („des panels d’élèves“), welche SchülerInnenkohorten über mehrere Jahre, vom 11. Lebensjahr bis zum Ende der Schul- und Hochschulzeit und dem Eintritt in die Arbeitswelt, begleiten. Dies geschieht u.a. auf der Basis nationaler Test- und Prüfungsergebnisse, Schulleitungs-, Eltern- und SchülerInnenfragebögen sowie -interviews. Die Ziele dieser Untersuchungen liegen eher auf systemischer Ebene: Öffentliche Debatten über Bildungsfragen sowie bildungspolitische Entscheidungen sollen durch empirische Daten unterfüttert werden. Es geht um Inhalte wie Trends im Kontext von Schulkarrieren oder die Entwicklung von Qualitätsstandards im Laufe der Zeit.

In Frankreich werden folglich – wie beschrieben – schulische Lernergebnisse auf verschiedene Art und Weise zentral erhoben und rückgemeldet. Wird darüber hinaus auch kontrolliert, wie die Daten vor Ort genutzt werden? Wie stark ist das Interesse an schulischen Nutzungsstrategien in Frankreich? Werden Daten und die

Verpflichtend ist zudem die Teilnahme an den zum Schuljahr 1994/95 eingeführten Abschlußprüfungen am Ende der Sekundarstufen I und II in allen Schulformen, die jedoch nicht zu 100 % in die Zeugnisnote eingehen: Die Fachnote setzt sich je zur Hälfte aus von CITO entwickelten und im Bereich der Multiple-Choice-Aufgaben zentral ausgewerteten Tests sowie aus schulinternen Prüfungen zusammen. Letzteres soll es den Schulen – als Kompromiß zwischen schulischer Autonomie und zentraler Regelung – ermöglichen, eigene inhaltliche Akzente zu setzen. Die Schulen haben jedoch für die Entwicklung des schulinternen Prüfungsteils auch Zugriff auf eine Aufgabedatenbank von CITO. Insgesamt herrsche eine hohe Zufriedenheit mit diesem System (van Bruggen 2000, S. 30). Aber auch im niederländischen Modell gibt es nicht bloß punktuelle Überprüfungen, die in Form von Abschlußprüfungen mit einer Zertifizierung von Leistungen zusammenfallen. Es besteht weiterhin die Möglichkeit, eine Vielzahl gut entwickelter Lernzuwachstests einzusetzen, die den Schulen von CITO zur Verfügung gestellt und von vielen Lehrkräften für die eigene Evaluationsarbeit genutzt werden und somit prozessual bei der Aufdeckung von Qualitätsmängeln helfen sollen.

Bis vor kurzem war die Veröffentlichung von Test- und Prüfungsergebnissen undenkbar. Schulen, Schulaufsicht und Ministerium wurden durch neue Entwicklungen überrascht: Die bis dahin von der Schulaufsicht vertraulich behandelten und als Diskussionsgrundlage mit den Schulen genutzten Daten wurden per Gerichtsentscheid unter Berufung auf das Recht der Informationsfreiheit von einer Zeitung eingeklagt. Aufgrund des großen öffentlichen Interesses nach den ersten Veröffentlichungen von Ranglisten in den Medien im Oktober 1997 – die Beilage der *Trouw* war in Rekordzeit ausverkauft – stimmte das Inspektorat einer regelmäßigen Publikation zu:

„Public interest in school performance appeared to be great. This was a bit of a surprise but not an unpleasant one. More than in the past, the inspectorate is working in a glass house. Working under these circumstances is a challenge” (KUWT 1999, S. 10).

In den Niederlanden versucht man mittels sogenannter Qualitätskarten (www.kwaliteitskaart.nl) von Anfang an ein gerechteres Leistungsbild zu vermitteln. Wesentlicher Punkt dieser für jede Schule veröffentlichten Karten ist die Bestimmung des Mehrwerts, der als ‚toegevoegde waarde‘ bezeichnet wird. Er wird u.a. auf der Grundlage ethnischer und sozio-ökonomischer Hintergrundmerkmale der Schülerschaft nach elaborierten statistischen Methoden berechnet (vgl. u.a. BOSKER/SCHREERENS 1999). Eltern, die die Daten bewerten möchten, können dies im nationalen und regionalen Vergleich tun, und sie können Schulen, die unter ähnlichen Umständen arbeiten, miteinander vergleichen. Bei der Entwicklung entsprechender Konzepte hat man bewußt auch ins Ausland geschaut, um aus den dortigen

langjährigen Erfahrungen rechtzeitig lernen zu können (vgl. KARSTEN/VISSCHER/DE JONG 1999).

Über diese regelmäßigen Leistungsfeststellungen hinaus kennt man auch in den Niederlanden externe Evaluationen in Form von Studien, die stichprobenartig schulische Wirkungsqualität erfassen. Dabei geht es zunächst einmal um das Sammeln von Daten zur Qualität des Schulsystems insgesamt und in der Folge für die Steuerung des Systems. Zu diesen Studien gehört PPON (‚Periodieke Peiling van het Onderwijsniveau‘), bei der in einem Zyklus von fünf Jahren seit 1986 Unterrichtsinhalte und Lernergebnisse der Primarstufe unter die Lupe genommen werden. Dies geschieht in allen Fächern – nicht bloß in Niederländisch und Mathematik – stichprobenartig durch CITO im Auftrag des Ministeriums (vgl. www.cito.nl/po/ppon/alg/eind_fr.htm). Die Ergebnisse werden u.a. in Relation zu den Kernzielen der Primarstufe gesetzt, um zu erkennen, inwieweit sie erreicht wurden bzw. geändert werden sollten. Durch den periodischen Untersuchungscharakter können weiterhin Entwicklungen von Standards im Zeitverlauf betrachtet werden. Zielgruppe der Ergebnisrückmeldungen sind vor allem Lehrerausbildungszentren, die Schulaufsicht, Didaktiker et cetera. Mittlerweile hat fast jede niederländische Schule schon einmal an der Untersuchung teilgenommen, bei der die SchülerInnen an Leistungstests teilnehmen und die Schulen Angaben zu Hintergrundmerkmalen ihrer Schülerschaft machen. Die Ergebnisberichte zu den einzelnen Fächern können bei CITO bestellt werden. Ein Daten-Feedback an einzelne, beteiligte Schulen gibt es bei PPON nicht; dies erlaubt die Anlage der Studie nicht (vgl. VAN DAM/VAN DER SCHOOT 1999, S. 35). Somit bleibt der Nutzen dieser Tests ein vermittelter.

Ein anderes Projekt verbirgt sich hinter dem Akronym PRIMA (‚landelijk cohortenonderzoek primair onderwijs‘), das seit 1994 vom Institut für angewandte Sozialwissenschaften IST an der Katholischen Universität Nijmegen und vom SCO-Kohnstamm Institut an der Universität von Amsterdam durchgeführt wird (vgl. www.its.kun.nl/2000ITS/Thema/omk/Prima.htm). In dieser Längsschnittstudie werden die Schullaufbahnen von ungefähr 60.000 SchülerInnen aus 700 Grundschulen mit dem Ziel, ein aktuelles Bild von der Primarstufe zu erhalten, alle zwei Jahre untersucht. Eine Ausweitung auf die Sekundarstufe ist geplant. Die Leistungen der SchülerInnen werden getestet, und Hintergrundmerkmale werden mittels Eltern-, Lehrkräfte- und Schulleitungsfragebögen erfaßt. Auch für die PRIMA-Studien gibt es Gesamtberichte, und darüber hinaus können die teilnehmenden Schulen die rückgemeldeten schulischen und außerschulischen Ergebnisse ihrer SchülerInnen mit denen ihrer Altersgenossen – auch im Zeitverlauf – landesweit vergleichen; dennoch geht es bei PRIMA nicht um eine individuelle Diagnostik (vgl. JUNGBLUTH 1995). Sowohl bei PPON als auch bei PRIMA handelt es

sich um Evaluationsstudien auf einem landesweiten ‚macro-niveau‘ (VAN DAM/VAN DER SCHOOT 1999). Da den Schulen durch verschiedene andere Tests ausreichend Daten für die SchülerInnen- und Gruppenevaluation auf einem ‚micro-niveau‘ und für die schulische Evaluation auf einem ‚meso-niveau‘ zur Verfügung stehen, wird die einzelschulische Ergebnisrückmeldung und -nutzung im Kontext von Studien wie PPO und PRIMA kaum untersucht: „Was die Schulen mit den Ergebnissen – die sie auf vertraulicher Basis rückgemeldet bekommen – anfangen, muß im Prinzip jede Schulleitung für sich selbst bestimmen“ (JUNGLUTH 1995, S. 13, aus dem Niederländischen übersetzt, I.v.A.).

Zentrale Tests und Prüfungen gehören zum Alltag niederländischer Schulen. Im Rahmen der PRIMA-Tests ist man bestrebt, die schulische Mehrbelastung durch die Stichprobenuntersuchung zu reduzieren: In Zukunft sollen die Testergebnisse des CITO-‚Leerlingvolgsystem‘ (LVS) für die Längsschnittuntersuchung genutzt werden, um Vernetzungen zwischen den Evaluationsformen zu schaffen. Das LVS ist ein weiteres, den niederländischen Grundschulen – auf der Primarstufe liegt ein besonderer Akzent evaluativer Maßnahmen in den Niederlanden – vertrautes Evaluationsinstrument (vgl. www.cito.nl/po/lvs/alg/eind_fr.htm). Über dieses u.a. von CITO entwickelte SchülerInnen-Informationssystem erhalten Schulen Daten zum Lernprozeß ihrer Schülerschaft. Diese Informationen sind für die direkte Verwendung in der Einzelschule gedacht. Mit Hilfe von Testmaterialien und einer speziellen Software kann die einzelne Lehrkraft Stärken und Schwächen identifizieren, letztere mittels weiterer Tests genauer analysieren und schließlich Aktionspläne erstellen. Das Leerlingvolgsystem zielt auf drei Ebenen der Ergebnisverwendung:

- 1) Die Testresultate sollen ein maßgeschneidertes didaktisches Handeln für den einzelnen Schüler/die einzelne Schülerin ermöglichen.
- 2) Überdies kann ein Profil der gesamten Lerngruppe erstellt werden, mit und an dem die Lehrkraft entsprechend arbeiten kann.
- 3) Schließlich sollen die Daten den Schulen ein umfassendes Bild ihrer Schulqualität insgesamt vermitteln, das durch die anderen beschriebenen Informationsquellen ergänzt werden kann.

Die Ergebnisse der PRIMA-Untersuchung, die u.a. auch schulische Erfahrungen mit Leerlingvolgsystemen evaluiert, zeigen bisher, daß die Informationen kaum auf der zuletzt beschriebenen Ebene verwendet werden. Eine Studie des SCO-Kohnstamm Instituut geht derzeit der Frage nach, wie Schulen die Daten in ihr Qualitätsmanagement im Moment integrieren bzw. in Zukunft einarbeiten können.

In den Niederlanden – so zeigt die Übersicht – unterscheidet man im wesentlichen zwischen öffentlich zugänglichen und vertraulichen Daten und den jeweiligen Effekten. Auf vertraulicher Basis mitgeteilte Daten sollen Schulentwicklungsprozesse direkt fördern. Publierte Indikatoren dienen zwar ebenfalls diesem Ziel, sollen aber zusätzlich zur schulischen Rechenschaftslegung (‚accountability‘) nach außen und zur Unterstützung elterlicher Schulwahlentscheidungen beitragen (vgl. BOSKER/SCHEERENS 1999). Eine Anfrage bei der Universität Twente (durch die Verfasserin) macht deutlich, daß man sich bisher vor allem um geeignete Rückmeldestrategien beider Formen von Informationen an die Schulen bemüht (vgl. auch DIJKSTRA u.a. 2001). Allerdings weiß man noch wenig über die Effekte in den Schulen, wozu aber erste Forschungsprojekte angelaufen sind, deren Ergebnisse man auch hierzulande wahrnehmen sollte.

3. Zusammenfassendes Fazit der Erfahrungen und Beobachtungen aus den Ländern

Der Überblick über das englische, französische und niederländische Test- und Prüfungswesen zeigt, daß Vergleichbarkeits- und Qualitätssicherung mittels zentral gesteuerter Leistungsmessungen unterschiedlicher, in nationalen Traditionen stehender Ausprägungen außerhalb Deutschlands weit verbreitet und nahezu normal ist (vgl. Abb. 2 für eine Übersicht über die Typen und Varianten externer Evaluation im Ausland im Vergleich zu Deutschland).

Zu den Zielen der verschiedenen Evaluationstypen können in unterschiedlicher Maße folgende Aspekte gehören:

- eine nationale Standortbestimmung und eine datengestützte Schulpolitik mittels eines Systemmonitorings,
- die Unterstützung elterlicher Schulwahlentscheidungen unter Berufung auf das Informationsrecht und zur Schaffung bildungsmarktähnlicher Strukturen,
- in umgekehrter Perspektive die Rechenschaftslegung der einzelnen Schule gegenüber der Öffentlichkeit,
- die einzelschulische Entwicklung vor Ort und als grundlegendes Ziel
- die Erreichung und Sicherung von Qualität und Vergleichbarkeit.

Zu den entsprechenden Instrumenten gehören in den untersuchten Ländern (vgl. Abb. 2):

- regelmäßig durchgeführte, schullaufbahnbegleitende Tests sowie Prüfungen zum Abschluß der Sekundarstufen I und II, aber auch

Abbildung 2: Beispiele großflächiger Test- und Prüfungsformen in Deutschland, England, Frankreich und den Niederlanden

	Nächstdeckende, regelmäßig durchgeführte zentrale Tests und Prüfungen		stichprobenartige, d.h. zeitlich und/oder regional begrenzte Leistungsstudien, (erfassen Hintergrundmerkmale)	
	schlußprüfungsähnliche Tests	Berechnung addiert value	Abschlußprüfungen	Berechnung addiert value
Deutschland	Bayrischer Mathematiktest (BMT) und Deutschtest <i>Testergebnisse fließen in die Zeugnisausgabe ein</i>	nein	zum Abschluß der Sekundarstufe I und II in Baden-Württemberg, Bayern, ostdeutschen Bundesländern (Brandenburg - noch - ausgenommen)	nein
England	National Curriculum Tests: auf den Key stages 1, 2, 3 im Alter von 7 und 11 Jahren (jeweils in Mathematik und Englisch) und mit 14 Jahren (zusätzlich in Naturwissenschaften)	value added: keine berechnete Informationen	am Ende der S.I. General Certificate of Education (GCSE) und am Ende der S.II. General Certificate of Education (GCE) gilt auch für berufsbildende Abschluss	value added: keine berechnete Informationen
Frankreich	évaluation en masse: - cours élémentaires: 6jährige - 6ème: 7 Jährige (jeweils Mathematik und Französisch) - seconde: 8jährige (Mathematik, Französisch, erste Fremdsprache)	nein	zum Abschluß der S.I. brevet des collèges zum Abschluß der S.II. baccalaurat	value ajoutée: indicateurs pour le pilotage des établissements scolaires (IPES)
Niederlande	zum Abschluß der Primarstufe: zweifach bissonderweils im Alter von 12 Jahren (Leistungsvollsystem (LVS): nicht obligatorisch, aber weit verbreitet in den Niederlanden)	nein	zentrale Prüfungen am Ende der Sekundarstufen I und II	begevoegde (kwaliteitskaart?)
				„den paschi d'elèves“ vom 11. Lebensjahr bis zum Eintritt in das Berufleben Stichprobenuntersuchung in der „treisième des collèges“ etwa alle fünf Jahre
				„periodieke peiling van het onderwijsniveau“ (PPON) alle 5 Jahre, Primarstufe, alle Fächer

Weitere Lehrlinien sowie Literaturhinweise finden sich auf Test, (Stand: 1/2004)

- regelmäßig und unregelmäßig stattfindende, flächendeckende oder stichprobenartige Leistungsstudien, die zumeist von der Bildungsadministration initiiert werden, aber durchaus auch von der einzelnen schulischen Einrichtung in Auftrag gegeben werden können.

Teilweise sind diese Formen externer Evaluation multifunktional, teilweise dienen sie vor allem einem der Ziele, wobei auch nach Vernetzungsmöglichkeiten gesucht wird. Eng verknüpft mit der Frage nach der Funktion dieser Formen der Leistungsfeststellung ist die Koppelung mit weiteren Steuerungsmechanismen. Dazu gehören die Veröffentlichung bzw. die Wahrung der Vertraulichkeit von Informationen und die Schaffung bzw. Vermeidung elterlicher Schulwahlmöglichkeiten in ihrer Verknüpfung mit der Pro-Kopf-Finanzierung von Schulen. Die Wege, die in den Test- und Prüfungswesen beschritten werden, sind im jeweiligen nationalen Kontext zu verstehen: In England geht es im wesentlichen darum, Leistung im marktformigen Wettbewerb zu steigern, in den Niederlanden will man Qualität und Vergleichbarkeit in einem auf Schulautonomie setzenden Land sichern, und in Frankreich sucht man nach der Stärkung von Chancengleichheit in einem zentral agierenden Staat.

Sowohl mit den Zielen als auch mit den Instrumenten des Testens und Prüfens, vor allem aber mit der Nutzung der Ergebnisse für die Schulentwicklung wird sich die innerdeutsche Debatte grundsätzlich auseinandersetzen müssen, wenn sie auf dem Weg regelmäßigen Testens und Prüfens sinnvoll fortschreiten will. Zudem wird man sich speziell in Deutschland mit einer weiteren Funktion der Leistungsmessungen beschäftigen müssen: Aufgrund der Kulturhoheit der Länder ergibt sich eine zusätzliche Dimension, nämlich der Vergleich der Länderpolitiken. TILLMANN führt uns eine mögliche Zukunft, die am niederländischen Modell orientiert ist, vor Augen. Er fordert eine entsprechende Diskussion der Chancen und Risiken einer Übertragung (TILLMANN 2001), will man nicht von Entwicklungen und unbeabsichtigten Effekten überrascht werden. Daß eine bestimmte Prüfungsform, ob sie nun zentral oder dezentral organisiert ist, nicht zu besseren Leistungsergebnissen im internationalen Vergleich führt, haben die TIMSS-Ergebnisse für den Bereich regelmäßiger Abschlußprüfungen gezeigt (vgl. BAUMERT/BOS/WATERMANN 1999, S. 118 ff). Klemm macht allerdings auf die Funktion der Schaffung und Sicherung von Vergleichbarkeit aufmerksam, die Zentralprüfungen zu einem sinnvollen Instrument der Schulentwicklung machen können:

„Gerade das Sichtbarmachen ungleicher Ergebnisse und ihres Zusammenhangs mit der Unterschiedlichkeit der Voraussetzungen für die Erbringung geforderter Leistungen und für deren Zertifizierung setzt eine Entwicklung in Gang, die

Gleichheit der Lernbedingungen fördern kann. [...] Es ist ein Irrtum zu unterstellen, daß Gleichheit dadurch befördert würde, daß ungleiche Ergebnisse nicht zur Kenntnis genommen werden“ (KLEMM 1998, S. 294).

So bleiben viele Fragen zu stellen, die man anderswo bereits zu beantworten sucht: Es bleibt zu klären, auf welche Art und Weise welche Daten an Schulen rückgemeldet werden sollten, so daß sie verstanden und in einem weiteren Schritt in Handlungsstrategien sowohl auf administrativer als auch auf schulischer Ebene umgesetzt werden. So können die Zentraltests und -prüfungen in ihrer Verknüpfung mit Hintergrundvariablen vielleicht doch über das Ziel der Vergleichbarkeit hinaus zur Qualitätsentwicklung und -sicherung beitragen. Der Erfahrungsvorsprung anderer Länder kann dahingehend genutzt werden, daß Test- und Prüfungsverfahren mit Large-Scale-Charakter einschließlich der Rückmelde- und Nutzungsstrategien außerhalb Deutschlands eingehender betrachtet und die im Ausland gewonnenen Erkenntnisse für die Schulentwicklung in Deutschland genutzt werden.

Literatur

- Ackeren, Isabell van/Klemm, Klaus: „TIMSS, PISA, LAU, MARKUS und so weiter. Ein aktueller Überblick über Typen und Varianten von Schulleistungsstudien“. In: Pädagogik 12/2000, S. 10–15
- Ashby, John/Sainsbury, Marian: How Do Schools Use National Curriculum Test Results? A survey of the use of National Curriculum test results in the management and planning of the curriculum at key stages 1 und 2. Online unter: www.nfer.ac.uk/summary/aic.htm
- Baumert, Jürgen/Bos, Wilfried/Watermann, Rainer: TIMSS/III. Schülerleistungen in Mathematik und den Naturwissenschaften am Ende der Sekundarstufe II im internationalen Vergleich. Zusammenfassung deskriptiver Ergebnisse. Berlin 1999
- Baumert, Jürgen/Watermann, Rainer: „Standardsicherung durch die Abiturprüfung: Zentralabitur oder dezentrale Prüfungsorganisation?“. In: Baumert, Jürgen/Bos, Wilfried/Lehmann, Rainer: TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 2. Opladen 2000, S. 341–351
- Bars, Stéphanie Le: „L'évaluation des établissements. La rumeur a la vie dure“. In: Le Monde de l'Éducation. September 1996, S. 46–49
- Bonnet, Gérard: „Profiles of Educational Assessment Systems World-wide. Country Profile from France“. In: Assessment in Education 2/1997, S. 295–306
- Bosker, Roel J./Scheerens, Jaap: „Openbaare prestatiegegevens van scholen; nuttigheid en validiteit“. In: Pedagogische Studien 1/1999, S. 61–73

- Broadfoot, Patricia: „Assessment Developments in French Education“. In: Educational Review 3/1992, S. 309–326
- Broadfoot, Patricia: Empowerment or Performativity? English Assessment Policy in the late twentieth Century. Brighton, 1999. Online unter: www.leeds.ac.uk/educol/documents/00001216.htm
- Bruggen, Johan C. van: „Qualitätsentwicklung durch zentrale Tests? Das Beispiel Niederlande“. In: Pädagogik 12/2000, S. 28–31
- Cousin, Olivier/Felouzis, George: „Qu'est-ce qu'un lycée efficace?“. In: Le Nouvel Observateur, 26.3.–1.4. 1998, S. 6–10
- Dam, Paul van/Schoot, Frank van der: „Professionele instrumenten voor professionele organisaties“. In: School & Begeleiding Dezember 1999, S. 29–36
- Demailly, Lise u.a.: Évaluer les établissements scolaires. Enjeux expériences, débats. Paris und Montréal 1998
- Dijkstra, Anne B. u.a. (Hg.): Het oog der natie: scholen op rapport. Standaarden voor de publicatie van schoolprestaties. Assen 2001
- Dutercq, Yves: Politiques éducatives et évaluations. Querelles de territoires. Paris 2000
- Fitz-Gibbon, Carol Taylor: The Value Added National Project Final Report. Feasibility studies for a national system of value-added indicators. Durham 1997. Online unter: www.qca.org.uk/ca/5-14/durham_report.pdf
- Foxman, Derek: Educational league tables: For promotion or relegation? A review of the issues. London 1997
- Helmke, Andreas: „Von der externen Leistungsevaluation zur Verbesserung des Lehrens und Lernens“. In: Trier, Uri Peter: Bildungswirksamkeit zwischen Forschung und Politik. Chur und Zürich 2000, S. 135–164
- Helmke, Andreas: „Internationale Schulleistungsvergleichsforschung. Schlüsselprobleme und Perspektiven“. In: Zeitschrift für Pädagogik 2/2001, S. 155–160
- Jungbluth, Paul: „Leerlingprestaties en -vaardigheden in een vergelijking tussen basisscholen. De terugkoppeling van toetsgegevens in het 'PRIMA'-cohort-onderzoek“. In: Basisschool Management April 1995, S. 9–13
- Karsten, Sjoerd/Visscher, Adrie/Jong, Tim de: Handle with care! Buitenlandse ervaringen met het publiceren van schoolprestatiegegevens. Achtergrondstudie. Onderwijsraad, Den Haag 1999
- Klemm, Klaus: „Steuerung der Schulentwicklung durch zentrale Leistungskontrollen?“. In: Rolff, Hans-Günter u.a. (Hg.): Jahrbuch der Schulentwicklung – Band 10. Weinheim 1998, S. 271–294
- Klemm, Klaus: „Large scale assessments in einem modernisierten Bildungssystem“. In: Die Deutsche Schule 3/2000, S. 329–328
- Klemm, Klaus/Schratz, Michael: „Leistungstests und Schulentwicklung“. In: Journal für Schulentwicklung 2/2001, S. 4–8
- Klieme, Eckhard/Baumert, Jürgen/Schwippert, Knut: „Schulbezogene Evaluation und Schulleistungsvergleiche – Eine Studie im Anschluss an TIMSS“. In: Rolff, Hans-Gün-

- ter u.a.: Jahrbuch der Schulentwicklung. Band 11. Weinheim und München 2000, S. 387–419
- Klug, Christian/Reh, Sabine: „Was fangen die Schulen mit den Ergebnissen an? Die Hamburger Leistungsvergleichsstudie aus der Sicht ‚beforschter‘ Schulen“. In: Pädagogik 12/2000, S. 16–21
- Kuijt, Paul: Towards a quality card for secondary schools. Netherlands Inspectorate of Education publishes results of secondary schools. o.O. 1999
- Lange, Hermann: „Qualitätssicherung in Schulen“. In: Die Deutsche Schule 2/1999, S. 144–159
- Maritzen, Norbert: „Schulleistungsforschung und Schulentwicklung, zwei Seiten einer Medaille“. In: Journal für Schulentwicklung 2/2001, S. 55–64
- Mitter, Wolfgang: „Der Vergleich in der Erziehungswissenschaft und die Vergleichende Erziehungswissenschaft. Oskar Anweiler zum 75. Geburtstag“. In: Bildung und Erziehung 1/2001, S. 91–103
- Peek, Rainer: „Rückmeldestrategien als Element der Qualitätsentwicklung durch großflächige Tests“. In: Journal für Schulentwicklung 2/2001, S. 55–64
- Rolff, Hans-Günter: „Was bringt die vergleichende Leistungsmessung für die pädagogische Arbeit in Schulen?“. In: Weinert, Franz E. (Hg.): Leistungsmessungen in Schulen. Weinheim und Basel 2001, S. 337–352
- Saunders, Lesley: 'Value added' measurement of school effectiveness: a critical review. NFER, 1999
- Saunders, Lesley/Rudd, Peter: Schools' use of 'value added' data: a science in the service of an art?. Brighton 1999. Online unter: www.nfer.ac.uk/conferences/valueadd.htm
- Schagen, Ian: Adding Value with Value-Added. London 1998. Online unter: www.nfer.ac.uk/conferences/value.htm
- Slater, John: „National tests; Parliamentary questions“. In: TES, 16.03.2001. Online unter: www.tes.co.uk/
- Stübiger, Heinz: „'Local Management of Schools' – ein Beitrag zur Verbesserung der Qualität von Schule?“. In: Die Deutsche Schule 1/1998, S. 93–105
- Thélot, Claude: L'évaluation du système éducatif. Coûts, fonctionnement, résultats. Paris 1993
- Thélot, Claude: Les lycées sous le feu de l'évaluation. La lettre grise. Lausanne 1999. Online unter: www.unil.ch/penombre/lg/04_02.htm
- Tillmann, Klaus-Jürgen: „Leistungsvergleichsstudien und Qualitätsentwicklung – oder: Auf dem Weg zu holländischen Verhältnissen?“. In: Journal für Schulentwicklung 2/2001, S. 9–17
- Vogler, Jean u.a.: L'évaluation. Paris 1996
- Whetton, Chris/Twist, Elizabeth/Sainsbury, Marian: National Tests and Target Setting: Maintaining Consistent Standards. New Orleans 2000. Online unter: www.leeds.ac.uk/educol/documents/00001422.htm
- Zanten, Agnes van: „Market trends in the French School System: overt policy, hidden strategies, actual changes“. In: Oxford Studies in Comparative Education 1/1996, S. 63–75

Der Recherchestand für die im Text und im Literaturverzeichnis genannten Internetdokumente ist November 2001.

Kurzbiographie

Isabell van Ackeren, geb. 1974, wissenschaftliche Mitarbeiterin, Studium der Biologie, Germanistik und Erziehungswissenschaft für die Sekundarstufen I und II in Essen mit Aufenthalten in England und den Niederlanden. 1999 Staatsexamen, seit 2000 Mitarbeiterin in der Arbeitsgruppe Bildungsforschung/Bildungsplanung an der Universität Essen. Veröffentlichungen in Mediävistik und Linguistik sowie im Bereich der Schulleistungsforschung. Aktuelles Projekt: Zur Nutzung großflächiger Tests für die Schulentwicklung. Gefördert vom Bundesministerium für Bildung und Forschung (BMBF).

Anschrift: Universität Essen, Fachbereich 2, AG Bildungsforschung/-planung, Universitätsstr. 11, 45117 Essen, E-Mail: isabell.vanackeren@uni-essen.de