

Viele europäische Nachbarländer sind deutlich »testerfahrener« als Deutschland. Was lässt sich von ihnen lernen? Wie kann man vermeiden, dass standardisierte Tests zu Verengungen des Unterrichts führen? Wie lässt sich die professionelle Urteilsfähigkeit von Lehrkräften für den Umgang mit Tests nutzen und unterstützen? Wie gelingt es, Tests für eine individuelle Lernberatung fruchtbar zu machen?

Vom Daten- zum Informationsreichtum?

Erfahrungen mit standardisierten Vergleichstests in ausgewählten Nachbarländern

ISABELL VAN ACKEREN

In der Wirtschaft sind solide Daten über Kunden, Warenbestände und Umsätze grundlegend für Entscheidungsprozesse. Eine Übertragung der wirtschaftlichen Bedeutung von Daten auf die Schule widerstrebt jedoch; schließlich sind Schüler keine Waren eines wirtschaftlichen Produktionsprozesses für abnehmende Märkte. Dennoch erfordern die Reformen im Bildungswesen auch in Schulen den Umgang mit Daten (Stichwort »landesweite Vergleichsarbeiten«). Angesichts der vielfältigen Testtypen ist die Zeit gekommen, die Beziehung zwischen standardisierten Tests, ihren Gebrauchsmöglichkeiten und ihrer Nutzung grundsätzlich zu überdenken. Denn: Zentrale Systemstrategien können zwar die Bedingungen für gute Lernergebnisse schaffen, sie können diese aber nicht direkt beeinflussen, da in jeder einzelnen Schule entschieden wird, wie auf externes Intervenieren reagiert wird. Der Blick

auf testerfahrene Länder bietet die Möglichkeit, sich mit Chancen und Grenzen dieses Reformelementes auseinander zu setzen und zu fragen, inwieweit Tests tatsächlich den Unterrichtsalltag erreichen. Exemplarisch werden nachfolgend Erfahrungen aus England, Frankreich und den Niederlanden skizziert. Die Erkenntnisse, die auf Literaturrecherchen und Experteninterviews basieren (vgl. van Ackeren 2003), werden eingeleitet durch grundsätzlichere Überlegungen zu unterschiedlichen Testformaten als Bildungsreformelement. Den Beitrag abschließend werden Handlungsperspektiven skizziert.

Tests als Bildungsreformelement

Nachfolgend wird nur von solchen Tests die Rede sein, die überhaupt geeignet sind, Schul- und Unterrichtsentwicklung zu unterstützen. Es geht nicht um Vergleichsstudien wie TIMSS und PISA; mit dem Ziel des Bildungsmonitorings sind sie vor allem auf der Systemebene nützlich. Es geht aber auch nicht um ad hoc erstellte informelle Tests, die einzelne Lehrkräfte für ihren Unterricht zu-



sammenstellen. Vielmehr sind so genannte formelle Testformate gemeint, die durch ein hohes Maß an Standardisierung bei der Durchführung und auch bei der Auswertung und Interpretation, möglichst nach testpsychologischen Gütekriterien, gekennzeichnet sind. Sie sind häufig in umfangreicheren Studien normiert und stellen Totalerhebungen dar, um die Position eines Individuums relativ zu einer vergleichbaren Gruppe oder einem früheren Messzeitpunkt bestimmen zu können (vgl. Leutner 2001). Sie stehen im Dienste unterschiedlicher pädagogischer Entscheidungen auf einer rationalen Basis. In diesem Sinne stellen sie Diagnoseinstrumente dar, die einerseits auf Auslese hin orientiert sein können (etwa bei Übergangentscheidungen) und die andererseits auf Förderung angelegt sind. Insbesondere bei der Förderdiagnostik können Defizite nicht nur konstatiert, sondern Maßnahmen zu ihrer Behebung abgeleitet werden, da die Daten für die aktuell betroffene Schülerpopulation genutzt werden können. In diesem Sinne fällt im Folgenden der Blick vor allem auf schulaufbahnbegleitende formelle Tests.

Erfahrungen in England

In England werden die »national curriculum«-Tests auf bestimmten Altersstufen (»key stages«) durchgeführt. Die Schüler werden – abgesehen von den zentralen Abschlussprüfungen am Ende der Schulkarriere – im Alter von 7, 11 (kurz vor dem Wechsel von der Primar- in die Sekundarstufe) und 14 Jahren extern getestet.

Zentrale Systemstrategien können zwar die Bedingungen für gute Lernergebnisse schaffen, können diese aber nicht direkt beeinflussen.

Mit Blick auf die Nutzbarmachung von Testergebnissen werden immer wieder taktische Mechanismen der Verbesserung erreichter Test- und Examensresultate als kurzfristige Reaktionen auf den sich weiter etablie-

renden englischen Bildungsmarkt beschrieben. Solche Effekte gehören zu den nicht-intendierten Reaktionen, da sie kaum etwas mit einer echten Schul- und Unterrichtsentwicklung zu tun haben und Benachteiligungen und Ungleichheiten zwischen Schulen und Schülern befördern:

- Schulen betreiben ein verstärktes »teaching for the test«, indem der Unterricht auf das fokussiert bzw. – kritisch formuliert – reduziert wird, was im Test erwartet und abgefragt wird.
- Zwar erhalten Eltern Entscheidungsfreiheit bei der Schulwahl, doch haben Schulen im Kontext verstärkter Autonomiebestrebungen zugleich größere Spielräume bei der Auswahl ihrer Schüler erhalten. Damit sind vor allem »gute« Schulen in die Lage versetzt, das Eingangsniveau ihrer Schüler zu regulieren.
- Der permanente Unterrichtsauschluss störender Schüler, deren Verhalten das Lernen und die Leistungen einer Klasse und schließlich den Rang in den Rankings beeinflussen kann, hat seit 1988 kontinuierlich zugenommen.

- Es wird immer wieder davon berichtet, dass sich Schulen auf ›mid-ability students‹ konzentrieren, um ihr Leistungsniveau und die entsprechende Darstellung in den Rankings zu verbessern. Spezielle Maßnahmen (Mentorenprogramme, Hausaufgabenhilfe, ergänzender Unterricht) sollen das Niveau heben. Kritiker verweisen auf die Übervorteilung dieser Schülergruppe zum Nachteil der starken und der sehr schwachen Schüler.
- Es wird von Strategien berichtet, wonach Schüler nicht zum Examen zugelassen werden, wenn es unmöglich scheint, ›high grades‹ zu erreichen.

.....

Immer wieder werden taktische Mechanismen der Verbesserung von Test- und Examensresultaten beschrieben (England).

.....

Durch ein solches Taktieren erfolgt keine direkte Nutzung der Daten. Mittlerweile wurden in vertiefenden Studien andere Handlungsmöglichkeiten beschrieben und man hat herausgearbeitet, was solche Schulen auszeichnet, die die Testinformationen in ihr eigenes Entscheidungsverfahren überführen. Eine 1998 an 415 Schulen durchgeführte Studie von Ashby und Sainsbury (2001) zeigt, dass Ergebnisse insbesondere dazu verwendet werden, schulische Stärken und Schwächen zu identifizieren, diese in Ziele für bestimmte Fachgebiete auf Schul-, Klassen-, Gruppen- (Anteile erfolgreich lernender Jungen/Mädchen, Migrantenkinder ...) und individueller Ebene umzuwandeln sowie in die curriculare Planung und in die Personalentwicklung einzubauen. Zudem werden die Resultate an den meisten (90 Prozent der befragten) Schulen in das Schulentwicklungsprogramm implementiert, was wiederum vom Schulinspektorat überprüft wird. In den meisten Schulen werden Vergleiche zu den Ergebnissen der Vorjahre gezogen. Knapp die Hälfte der Schulen vergleicht sich mit anderen Schulen. Etwa 40 Prozent der Schulen evaluiert die angemessene Behandlung der Inhalte des nationalen Curriculums im Unterricht.

Saunders und Rudd (1999) beschreiben auf der Grundlage einer Fallstudie an neun Schulen, dass Daten offensichtlich insbesondere dann genutzt werden, wenn es in der Schule eine Person gibt, die an den Daten interessiert ist, sich für ihre Weiterleitung verantwortlich fühlt und in der Lage ist, den Kollegen die technischen Aspekte zu vermitteln. Schulen waren besonders produktiv im Umgang mit Daten, wenn sie schon einmal Interpretationshilfen und Fortbildungen erhalten hatten. Als weitere Aspekte einer effektiven Schulkultur im Hinblick auf die Nutzbarmachung von Tests werden beschrieben: Eine systematische Prüfung der Testergebnisse erscheint selbstverständlich, Wandel und Innovation werden insgesamt akzeptiert, die Lehrkräfte haben einen klaren Fokus auf hohe, aber realistische Leistungserwartungen für alle Schüler, es existieren effektive interne Kommunikationsstrukturen, es findet eine Weiterleitung von Aufgaben an möglichst viele Kollegen mit einem realistischen Blick für deren Kapazität im Umgang mit Daten statt und Schüler werden aktiv in die Reflexion ihrer Ergebnisse einbezogen.

Erfahrungen in Frankreich

Seit dem Reformgesetz von 1989 ist man verstärkt um eine Evaluation der Schülerleistungen mittels auf das nationale Curriculum abgestimmter Tests bemüht, die die bilanzierenden landesweiten Abschlussprüfungen ergänzen. Seitdem werden alle Schüler im Alter von acht Jahren etwa zur Halbzeit der Primarstufe und im Alter von elf Jahren zu Beginn des Collège in Französisch und Mathematik zu Beginn eines Schuljahres ›en masse‹ getestet. Ein ähnlicher Test wurde 1992 ebenso für die 16-Jährigen zu Beginn des ›lycée‹ eingeführt. Dieser Test wurde allerdings wieder abgeschafft, »[...] da die Ergebnisse nicht ausreichend von den Lehrkräften genutzt wurden« (Bonnet 2002, S. 47). Schulen erhalten zudem Schulleitungs-, Lehrer- und Schüler-Fragebögen, um Hintergrundmerkmale zu erfassen. In Frankreich ist jede Schule selbst für die Datenanalyse verantwortlich, was die Intention unterstreicht, die Daten für die Unterrichtsplanung zu nutzen. Die Resultate können für eine Standortbestimmung zudem mit den öffentlich einzusehenden nationalen Stichproben-

ergebnissen verglichen werden.

Mit den Tests soll eine Evaluationskultur in den Schulen gefördert und den Lehrkräften sollen die Konsequenzen ihres unterrichtlichen Handelns wie einen Spiegel vor Augen geführt werden. Sie sollen sich die Heterogenität der Schülerschaft gegenwärtigen und gegebenenfalls zu einer Individualisierung ihres Unterrichts und der Hinterfragung angewandter Methoden kommen. Sie haben über diese Tests hinaus Zugriff auf eine Datenbank des Ministeriums mit konkreten Aufgaben, die laufend im Unterricht eingesetzt werden können. Die Tests haben keinen zertifizierenden, sondern diagnostischen Charakter für Lehrkräfte und Eltern, zumal sie zu Beginn des Schuljahres durchgeführt werden.

Anhand der Literaturlage wird deutlich, dass man wenig über den tatsächlichen Gebrauch der unterschiedlichen Daten auf der Schul- und Unterrichtsebene weiß. Die administrative Idee der Initiierung einer Evaluationskultur wurde offensichtlich in den Schulen bisher kaum aufgegriffen. Informationen zur Nutzungsfrage finden sich u. a. bei Bonnet (1997), der auf der Grundlage von Interviews mit mehreren hundert Lehrkräften folgendes berichtet: 90 Prozent der Primarschul- sowie 85 Prozent der Sekundarschullehrenden nutzen die Tests. Sie werden in erster Linie für die Informierung der Eltern sowie zur Bildung homogener Leistungsgruppen genutzt. Kaum Verwendung finden sie hingegen für die in Frankreich obligatorische Schulprogramm-Entwicklung. Abschließend sei noch auf einen nicht unbedeutenden Steuerungseffekt hingewiesen: So werden Effekte beschrieben, wonach national festgestellte Schwächen, z. B. in der Geometrie, zu ihrer Verbesserung im darauf folgenden Jahr führten, weil sich die Lehrkräfte offensichtlich besonders auf diesen Aspekt konzentrierten. Damit ist eine deutliche Steuerungswirkung beschrieben.

Erfahrungen in den Niederlanden

In den Niederlanden gibt es – neben den teilzentralen Abschlussprüfungen – nur einen obligatorischen schullaufbahnbegleitenden Test am Ende der Grundschulzeit. Es ist den Schulen freigestellt, welchen Test sie einsetzen, jedoch wählen 80 Prozent den Test des zentralen Testinstituts

CITO. Er wird schulintern durchgeführt, was immer wieder mal zu Problemen führt, da am Testtag an manchen Schulen überdurchschnittlich viele leistungsschwächere Schüler krank gemeldet sind oder unzulässige Hilfen bei der Testbearbeitung gegeben werden. Deshalb wird eine verstärkte externe Kontrolle diskutiert. Die eigentliche Funktion dieses Tests wird wie folgt beschrieben: Zunächst einmal geht es um das Testen der individuellen Kompetenzen jedes einzelnen Schülers. Dies bietet ihm, seinen Eltern sowie den Lehrkräften eine objektivierte Basis für die weitere Schullaufbahnorientierung. Darüber hinaus soll er die schulische Selbstevaluation durch die Daten, die eine nationale Bezugsbasis haben, anregen und ermöglichen.

Weiterhin besteht die viel genutzte Möglichkeit, entwickelte Lernzuwächstests einzusetzen, die den Schulen von CITO zur Verfügung gestellt werden. Sie bieten der einzelnen Lehrkraft einen externen, nationalen Vergleichsmaßstab. Das »leerlingvolgsysteem« (Schülerbegleitsystem) ist ein weiteres, den niederländischen Grundschulen – auf der Primarstufe liegt ein besonderer Akzent evaluativer Maßnahmen – vertrautes Evaluationsinstrument. Es ist dazu gedacht, das Unterrichtsangebot besser abzustimmen und das methodisch-didaktische Handeln auf die individuellen Bedürfnisse zuzuspitzen. Solche Systeme sind auf die Situation des einzelnen Schülers gerichtet. Die Daten können aber auch auf Klassen- und Schulebene aggregiert werden. Neben den Testpaketen zu Basisfertigkeiten gibt es Materialien, um hinter den Testergebnissen liegende Ursachen besser zu erkennen, und es gibt konkrete Materialien, die bei identifizierten Schwächen zur systematischen Lernentwicklung eingesetzt werden können.

Roeleveld, Otter und Blok (2002) berichten zum Umgang mit solchen Schülerbegleitsystemen über folgende Nutzungsstrategien: Schlechte Testergebnisse einzelner Schüler werden fast durchgängig analysiert, die Resultate werden durchweg beim Übergang in die nächste Klassenstufe weitergegeben und die erzielten Mittelwerte der Lerngruppe werden in jeder zweiten Schule mit den nationalen Ergebnissen verglichen. Allerdings stelle sich die Testabnahme und Auswertung als sehr zeitintensiv dar.

Bildungspolitische und -praktische Handlungsperspektiven

Innovation ist ein Prozess, kein einmaliges Ereignis. Reformen bedürfen der Stärkung der Kompetenz, den eigenen Wandel vor Ort zu organisieren. In diesem Zusammenhang erscheint es wichtig, die Fähigkeit des Evaluierens, den Umgang mit Qualitätsentwicklungssystemen und eine begründete Akzeptanz bereits in der Ausbildung zu fördern, eine Überlastung der Lehrkräfte zu vermeiden und Motivation zu fördern, z. B. durch die Schaffung eines ständigen Angebots, Wissen und Kompetenzen – individuell und als Organisation – weiterzuentwickeln und Anstrengungen in die Ausbildung von Beratern und Moderatoren zu investieren. Darüber hinaus werden Lehrkräfte den Daten gegenüber nicht offen sein, wenn sie das Gefühl haben, dass

*Die professionelle
Urteilsfähigkeit von
Lehrkräften muss genutzt und
unterstützt werden.*

Botschaften der Daten in Entscheidungen einfließen, die sie direkt betreffen (leistungsorientierte Bezahlung und Aufstiegsmöglichkeiten ...). Es bedarf zwar einer leistungsorientierten Kultur, die aber gleichwohl Vertrauen in die professionellen Fähigkeiten von Lehrern setzt. Die professionelle Urteilsfähigkeit von Lehrkräften muss genutzt und unterstützt werden, indem ihnen nicht bei der Auswertung von Tests und Prüfungen das Mandat entzogen wird. In umgekehrter Denkrichtung bedeutet dies auch, dass Lehrer taktische Verbesserungsmaßnahmen (vgl. England) vermeiden und sich innovationsbereit und offen zeigen. Viele Entwicklungen der letzten Jahre im Bereich der Testtheorie wurden von Praktikern kaum zur Kenntnis genommen; wohl u. a. wegen der statistischen Komplexität der Verfahren. Dennoch sei vor einer kurzschlüssigen Testkritik gewarnt. Gesunde Skepsis ist sicher angebracht; um jedoch das Ziel einer effektiven daten-geleiteten Schulentwicklung zu erreichen, sind alle Akteure im Bildungs-

system mit ihrem spezifischen Blick für Chancen und Gefahren gleichermaßen gefordert. Als größtes Problem stellt sich die »ownership« der Daten dar, das Gefühl, Anteil an ihnen zu haben und auf sie Einfluss nehmen zu können. Es muss gelingen, Vertrauen und Kompetenz in bzw. bei der Dateninterpretation in Schulen aufzubauen. Dabei müssen beide Seiten, Bildungsadministration und Bildungspraxis, eingestehen, dass sie Grenzen bei der objektiven bzw. fairen Messung von Leistung haben; in diesem wechselseitigen Eingeständnis liegt Potenzial für den Weg vom Daten- zum Informationsreichtum.

Literatur

Ackeren, I. van: Evaluation, Rückmeldung und Schulentwicklung: Erfahrungen mit zentralen Tests, Prüfungen und Inspektionen in England, Frankreich und den Niederlanden. Münster 2003

Ashby, J./Sainsbury, M.: 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. Slough 2001

Bonnet, G.: Profiles of Educational Assessment Systems World-wide: Country Profile from France. Assessment in Education 4 (1997) H. 2, S. 295-306
Bonnet, G.: Schulentwicklung in Frankreich. PÄDAGOGIK 54 (2002) H. 4, S. 44-48

Leutner, D.: Pädagogisch-psychologische Diagnostik. In: D. H. Rost: Handwörterbuch Pädagogische Psychologie, Weinheim 2001, S. 521-529

Roeleveld, J./Otter, M.E./Blok, H.: Leerlingvolgsystemen in de jaren 90: Secundaire analyses op gegevens uit PRIMA en IST. Amsterdam 2002

Saunders, L./Rudd, P.: Schools' use of »value added« data: A Science in the service of an art? Slough 1999

Dr. Isabell van Ackeren, Jg. 1974, vertritt derzeit die Professur für Allgemeine Didaktik an der Universität Duisburg-Essen.

Adresse: Universität Duisburg-Essen, Fachbereich Bildungswissenschaften, 45117 Essen, E-Mail: i.vanackeren@uni-essen.de