ARGEV

INTERKANTONALE ARBEITSGEMEINSCHAFT EXTERNE EVALUATION VON SCHULEN

Zum Stand der externen Schulevaluation in Verbindung mit Leistungsmessung

Leistungstests und Schulevaluation in der deutschsprachigen Schweiz und Blick in andere Länder

Autorin:
Vera Husfeldt

im Auftrag der ARGEV

Aarau, den 04.12.2007

Fachhochschule Nordwestschweiz
Pädagogische Hochschule

Inhaltsverzeichnis

1 Einleitung

In den letzten Jahren haben sich in den Kantonen der Deutschschweiz unterschiedliche Formen der Qualitätsbeurteilung von Schulen durchgesetzt. Neben internen Selbstbeurteilungen kommen verstärkt auch externe Evaluationsverfahren zum Tragen. So hat beispielsweise die NW EDK mit dem Projekt Qualität durch Evaluation und Entwicklung (Q2E) den Prozess zur Entwicklung einer umfassenden Feedback- und Evaluationskultur angestossen. Q2E versteht sich als Hilfsinstrument beim Aufbau eines schulischen Qualitätsmanagements, das über eine externe Evaluation als Teil das Modells überprüft wird. Zur Beurteilung der schulischen Qualität werden dabei Steuerungsqualitäten der Schulleitung sowie Aspekte der Feedback- und Evaluationskultur berücksichtigt. Die Output- bzw. Outcomebeurteilung bleibt in Modellen wie diesem zunächst auf die Ebene der Organisation und Leitung begrenzt. Nicht mit berücksichtigt sind die Leistungen und Entwicklungen der Schülerinnen und Schüler.

Bildungsstandards

Seit den Überlegungen zur Harmonisierung der obligatorischen Schule (HarmoS) und damit zur Einführung nationaler Bildungsstandards in der Schweiz kommt dem Aspekt der Messung von Schülerleistungen eine noch stärkere Bedeutung zu. Im Gegensatz zu der inputorientierten Lehrplanfestlegung ist die Setzung von Bildungsstandards auf den Output des Bildungssystems gerichtet. Die Bildungsstandards beschreiben die Lernergebnisse, die für Schülerinnen und Schüler zu bestimmten Zeitpunkten in ihrem Bildungsweg erwünscht werden. Dabei sind laut Klieme et al. (2003) folgende Aspekte Merkmale guter Bildungsstandards: Fachlichkeit, Focussierung, Kumulativität, Verbindlichkeit für alle, Differenzierung, Verständlichkeit, Realisierbarkeit. Während die Bildungsstandards beschreiben, welche Ziele erreicht werden sollen, beziehen sich die Lehrpläne darauf, mit welchen Mitteln diese Ziele erreicht werden können. Die Bildungsstandards werden deshalb auch längerfristig nicht die Lehrpläne ablösen, sondern es ist zu hoffen, dass mit ihrer Hilfe Lehrpläne und Unterrichtsprozesse verbessert werden können. Bildungsstandards haben zwei wesentliche Steuerungsfunktionen. Zum einen setzen sie Bildungsziele in konkrete Kompetenzanforderungen um und „legen fest, über welche Kompetenzen ein Schüler verfügen muss, wenn wichtige Ziele der Schule als erreicht gelten sollen." (Klieme et al. 2003, 21) Im Sinne der Eigenverantwortlichkeit der Schulen lassen sie aber die Möglichkeit, frei zu entscheiden, wie diese Ziele erreicht und wie die dazugehörigen Kompetenzen erworben werden. Zum anderen werden die Bildungsstandards als Ergebnisse von Lernprozessen „in Aufgabenstellungen und schliesslich Verfahren [konkretisiert], mit denen das Kompetenzniveau, das Schülerinnen und Schüler tatsächlich erreicht haben, empirisch zuverlässig erfasst werden kann." (Klie-

me et al. 2003, 23). Das heisst, dass das Erreichen der Standards empirisch anhand von Leistungstests überprüft werden kann. Dieses Feedback über den erreichten Kompetenzstand ist laut Klieme et al. ein unverzichtbarer Bestandteil einer kontinuierlichen systematischen Qualitätsentwicklung für Schule und Unterricht sowie für die Didaktik.

Die Beurteilung von Schülerleistung wird in der Schweiz meist unabhängig von anderen Formen der externen und internen Evaluation und bisher auch ohne Bezug zu den Bildungsstandards durchgeführt. Als Test mit einer längeren Tradition im Bereich der Leistungsmessung ist beispielsweise die Bezirksabschlussprüfung im Kanton Aargau zu nennen, die durch ein neues System von Leistungsmessungen abgelöst werden soll. In diesem neuen Konzept spielt der ursprünglich vom Kanton St. Gallen in Auftrag gegebene computerbasierte Test mit Namen Stellwerk eine grosse Rolle. Er umfasst unterschiedliche curricular relevante Testbereiche und ergänzt das ebenfalls in St. Gallen entwickelte „Klassencockpit", das auf freiwilliger Basis in den Klassen 3-9 eingesetzt werden kann, um eine Standortbestimmung für Lehrkräfte zu bekommen. Weitere Messungen von Schülerleistungen werden in Form von regional entwickelten Vergleichsarbeiten (z.B. Zentralschweiz) durchgeführt.

Leistungsmessung und Evaluation

Sowohl die externe Schulevaluation als auch die Leistungsmessung beschreiben Aspekte der Qualität von Schule. Modelle wie Q2E fördern die Schulentwicklung indem durch unterschiedliche Verfahren Erfolge oder Probleme einer Schule an bestimmten Merkmalen evaluiert und damit bewusst und veränderungsfähig gemacht werden. Wenn man davon ausgeht, dass Schülerleistung bzw. Lernentwicklung als ein Merkmal von Schulqualität anzusehen ist, dann sollte sie ebenso wie andere organisatorische und führungsrelevante Aspekte in diesen Prozess eingebunden werden. In der Praxis zeigt sich jedoch, dass bisher der Bereich der externen Schulevaluation und der Bereich der Leistungsmessung noch weitgehend unverknüpft dastehen und der Schulentwicklungsprozess in dieser Hinsicht noch verbessert werden könnte. Eine Einbindung der Ergebnisse aus der Leistungsmessung in die weitere Beurteilung der Qualität einer Schule würde auf der einen Seite der externen Evaluation ein weiteres Merkmal zur Beschreibung von Schulen an die Hand geben und damit auf der anderen Seite auch die Leistungsmessung als Instrument der Schulentwicklung stärken.

Externe Schulevaluation ist dazu geeignet, in ausgewählten Bereichen Aussagen über die Qualität von Schule zu machen. Bereiche, die Schülereigenschaften direkt betreffen, sind dabei jedoch meist nur am Rande betroffen, als würde Qualität von Schule nur durch die Fähigkeiten und Kompetenzen der Leitungs- und Lehrpersonen manifest. Durch eine Einbindung der Leistungsmessung in die externe Schulevaluation können zumindest Teilbereiche von Schülerleistungen mit einbezogen werden und damit die Basis der Beur-

teilung von Schulqualität etwas erweitern.

Ist Schülerleistung = Schulleistung?

Allein von der Leistungsmessung ausgehend kann nur in den wenigsten Fällen überhaupt ein Rückschluss auf die Qualität einer Schule zugelassen werden. Punktuelle Schülerleistungen sind, wie wir aus einer Reihe empirischer Studien (vgl. z.B. Zahner et al. 2002) wissen, sehr stark von sozialen Faktoren und nur zu ganz geringen Anteilen direkt von der schulischen Bildung abhängig. Wenn also ein Schüler sehr gute Leistungen im Lesen erbringt, muss dies nicht unbedingt ein Zeichen von hoher Qualität seiner Schule sein, ebenso muss eine schlechte Schülerleistung nicht zwingend auf mangelnden Erfolg der Schule zurückzuführen sein, es gibt eine Reihe anderer Erklärungsansätze dafür. Zu den Einflussfaktoren für Schülerleistung zählen natürlich die kognitiven Voraussetzungen, die Einstellungen zum Lernen und die Lerngewohnheiten des Schülers. Aber selbst wenn alle denkbaren Einflussfaktoren für alle Schüler konstant wären, könnte noch nicht von der Schülerleistung auf die Qualität der Schule geschlossen werden, da die Schüler bereits mit unterschiedlich ausgeprägten Leistungen an die Schule kommen.

Qualität von Schule im Bereich der Schülerleistung besteht in der bestmöglichen Förderung also der maximalen Leistungssteigerung der Kinder und Jugendlichen. Will man also Schulqualität im Bereich der Schülerleistung beurteilen, so ist es unerlässlich, zusätzlich zu der Momentaufnahme auch die Entwicklung der Schülerleistung zu betrachten. Die Leistungsentwicklung eines Schülers ist viel stärker als seine Leistung selbst von schulischen und unterrichtlichen Faktoren abhängig. Das ist dadurch zu erklären, dass die genetischen Voraussetzungen und die Einflüsse des sozialen Umfeldes einen sehr grossen Einfluss auf die Leistung eines Schülers haben. Wie aus einer Reihe von Studien bekannt ist, ist die Ausgangslage bei Eintritt in die Schule der stärkste Prädiktor für die Schülerleistung. Bei einer Betrachtung der Leistungsentwicklung hat hingegen per Definition die Ausgangslage keinen Einfluss. Das heisst allerdings nicht, dass bei der Leistungsentwicklung soziale und genetische Faktoren keine Rolle spielen. Selbstverständlich hängt auch die Leistungsentwicklung nicht nur vom schulischen Umfeld ab, sondern auch von Faktoren wie z.B. häusliche Bildungsressourcen oder Lerndispositionen. Das bedeutet, dass selbst die Betrachtung der Leistungsentwicklung der Schülerinnen und Schüler einer Schule keine sicheren Aussagen über die Qualität der Schule zulässt. Wenn aber die Daten der Leistungs- und Leistungsentwicklungsmessung mit Daten aus der externen Schulevaluation verknüpft werden, können sie sich gegenseitig stützen und einen Schritt weiter gehen, als dies bisher geschehen ist.

Zielsetzung des Berichts

Ziel dieses Berichtes ist, darzustellen, wie eine solche Verknüpfung der beiden Bereiche sinnvoll realisiert werden kann. Dazu werden im Folgenden einige in der Deutschschweiz bestehende Modelle der Leistungsmessung dargestellt und im Hinblick auf ihre Einbindung in die externe Schulevaluation genauer betrachtet. Ausgehend von der Diskussion dieser Modelle werden anschliessend Modelle zur Leistungsmessung in anderen Ländern dargestellt und hinsichtlich ihrer Aussagekraft zur Qualität von Schule beurteilt. Die Betrachtung der Modelle anderer Länder gibt auf der einen Seite genaueren Aufschluss über die Möglichkeiten und Erfolge der Implementation einer Verknüpfung von Schulevaluation und Leistungsmessung. Andererseits werden dadurch aber auch die Probleme und Schwierigkeiten deutlich, die mit einer solchen Zusammenführung verbunden sind. Um ein möglichst breites Spektrum an Modellen abzudecken werden neben der Schweiz drei weitere europäische Länder betrachtet. Diese Länder zeichnen sich durch vergleichbare Lebens- und Organisationsformen, vergleichbare Wert- und Geisteshaltungen und eine vergleichbare historische, wirtschaftliche und kulturelle Tradition aus und eignen sich damit sehr gut, um eine Reflexion über die Modelle in der Schweiz anzuregen.

Das Bildungssystem in Deutschland weist nicht nur durch die föderalen Strukturen sehr grosse Parallelen zum Schweizerischen System auf. Ähnlich wie in der Schweiz ist die Tradition der Schulevaluationsforschung und Schulleistungsmessung in Deutschland noch sehr jung. England und die Niederlande haben in diesem Bereich eine wesentlich längere Forschungstradition, doch allen Ländern gemeinsam ist die Schwierigkeit, die Ergebnisse grossangelegter Schulleistungsstudien konkret für die Schulentwicklung der Einzelschule zu nutzen. Die Bemühungen, die unternommen werden, um dieses Ziel zu erreichen und die Modelle, die dafür verwendet werden, sind in den einzelnen Ländern sehr unterschiedlich.

England hat ein sehr striktes Überprüfungsystem eingeführt, auf deren Grundlage die Schulen für die Leistungen ihrer Schüler verantwortlich gemacht werden. Das Land zeichnet sich durch ein stark marktorientiertes Prüfungssystem und den Einsatz der Tests als Mittel zur Schulsystemsteuerung aus. In England sticht der starke Einfluss des Inspektorats heraus. Der Bereich der Leistungsmessung und der externen Schulevaluation sind eng verbunden und beide Bereiche überzeugen durch sorgfältige Planung und Organisation.

Die Niederlande haben zwar im Gegensatz zu Deutschland und der Schweiz kein föderales System, sind aber mit einer extrem hohen Vielfalt unterschiedlicher Modelle in ihrem Schulsystem konfrontiert. Sowohl privat getragene Schulen als auch staatliche Schulen verfügen in den Niederlanden über eine sehr hohe Autonomie. Um trotzdem vergleichbare Bedingungen und vergleichbare Schulleistungen zu gewährleisten, setzen die Niederlande auf ein

Tabelle 1: Leitfragen

- Was wird gemessen?

  Fachkompetenzen, Methodische Kompetenzen, soziale Kompetenzen, affektive Kompetenzen

- Wie wird gemessen?

  Testarten (siehe auch Anhang B), Rolle der Testzeit

- Wer misst?

  welches Institut, welche institutionelle Einbettung, welche Kooperationen mit anderen Evaluationsstellen

- Wie läuft die Durchführung ab?

- Wie werden die Daten analysiert?

  welche statistischen Auswertungsverfahren, Skalierungen, Mehrebenenanalysen

- Wer erhält Einblick in welche Daten und Ergebnisse?

- Bei wem liegt die Rechenschaftspflicht?

- Wie werden Verknüpfungen von Schülerleistungsdaten zu Schulqualitätsdaten vorgenommen?

- Wie fliessen die Daten in den externen Evaluationsprozess bzw. ins schulinterne Qualitätsmanagement ein?

- Wie erscheinen Leistungsdaten im Evaluationsbericht?

umfassend gestaltetes Monitoringsystem, dass Schulleistungsmessung und externe Schulevaluation in beeindruckender Weise verbindet und die notwendigen Informationen für die Schulsystemsteuerung liefert.

Anhand von Leitfragen werden die Modelle der Leistungsmessung und externen Schulevaluation der ausgewählten Länder im Folgenden analysiert. Die Beschreibung der Projekte aus den unterschiedlichen Ländern soll sich grob an diesen Leitfragen (siehe Tabelle 1) orientieren.

2 Leistungstests in der deutschsprachigen Schweiz

Ein einheitliches System der Leitungsmessung liegt in der Deutschschweiz bisher noch nicht vor. Allerdings gibt es in einigen Bereichen umfangreiche Testmaterialien für die Messung von Leistung. Die Erfassung und Analyse der Daten sowie teils auch die Rückmeldung der Ergebnisse erfolgt bisweilen mit modernsten Verfahren. Dennoch wird relativ schnell deutlich, dass die Beurteilung von Schulqualität hieraus noch nicht ableitbar ist. Im Folgenden werden wichtige Praxisbeispiele für Leistungsmessungen in der Deutschschweiz vorgestellt. Diese Aufstellung beansprucht nicht, vollständig zu sein.

2.1 Check 5: Freiwillige Standortbestimmung für fünfte Klassen

2.1.1 Zielsetzung, Organisation und Inhalt

Der Check 5 ist ein freiwilliger Leistungstest der vom Departement für Bildung, Kultur und Sport des Kantons Aargau für die fünften Klassen der Primarschulen kostenlos bereitgestellt wird. Die Entwicklung, Durchführung und Auswertung obliegt dem Institut für Bildungsevaluation, Assoziiertes Institut der Universität Zürich. Der Test orientiert sich an den Lehrplänen des Kantons und beinhaltet die Bereiche Deutsch und Mathematik, die in jeweils 90 Minuten bearbeitet werden müssen. Zusätzlich testet er auch Fähigkeiten im kooperativen Problemlösen (60 Minuten) und im selbstregulierten Lernen (20 Minuten). Ziel dieses Leistungstests ist es, auf der einen Seite durch die gewonnenen Informationen über Schülerleistungen eine Förderung individueller Schülerinnen und Schüler sowie der gesamten Klasse vorzunehmen und den Unterricht damit nachhaltig weiterzuentwickeln und zu verbessern. Die Lehrpersonen können darüber hinaus auch mit den Check-5-Ergebnissen ihre eigene Beurteilungspraxis reflektieren. Auf der anderen Seite soll der Test auf der bildungspolitischen Ebene Aufschluss über die Schulleistungen von Fünftklässlern in ausgewählten Bereichen geben und ggf. Handlungsbedarf erkennen lassen.

Der Test ist in ein Evaluationsverfahren eingebunden, das die folgenden fünf Schritte umfasst:

1. Leistungen bestimmen und Test entwickeln,
2. Leistungen messen und beurteilen,
3. Testergebnisse analysieren und interpretieren,
4. Ziele setzen und Massnahmen ergreifen,
5. Massnahmen umsetzen und deren Wirkung überprüfen.

Die Lehrpersonen sind ganz konkret in den letzten drei Schritten gefordert. Sie analysieren die Ergebnisse und versuchen, die Ursachen dafür zu erkennen. Auf der Basis der neuen Erkenntnisse setzen sie Ziele und ergreifen

Massnahmen zur entsprechenden Veränderung ihrer Unterrichtsgestaltung. Über die Wirkung dieser Massnahmen sollen die Lehrpersonen dann im Abschluss reflektieren.

2.1.2 Durchführung

Im Vorfeld der eigentlichen Testdurchführung findet für alle teilnehmenden Lehrpersonen eine obligatorische Informationsveranstaltung statt. An dieser Veranstaltung werden die Lehrpersonen u.a. bezüglich der Durchführung angeleitet, es werden offene Fragen geklärt und das Testmaterial abgegeben etc. Die Schüler können den Test nach Anweisung durch die Lehrpersonen innerhalb der regulären Unterrichtszeit in rund vier Stunden während eines Testfensters von zwei Wochen im September bearbeiten. Zur Durchführung des Tests erhalten die Lehrpersonen rechtzeitig alle Testmaterialien und eine Handreichung, sie bestimmen den Zeitpunkt der Durchführung und verteilen dazu die bereits mit Namen versehenen Testbögen an die jeweiligen Schülerinnen und Schüler. Die Testbögen werden anschliessend von den Lehrpersonen eingesammelt und in einer bereits vorbereiteten frankierten und adressierten Schachtel zur weiteren Bearbeitung an das zuständige Institut für Bildungsevaluation, Assoziiertes Institut der Universität Zürich verschickt. Dort werden die Daten eingegeben und statistisch weiterverarbeitet. Für die einzelnen Kompetenzbereiche werden anhand der Verfahren der Item-Response-Theory[1] Skalen gebildet. Der Einfluss individueller Hintergrundvariablen wird bei der Analyse der Daten berücksichtigt. Ausserdem werden Fachleistungsunterschiede zwischen den teilnehmenden Klassen analysiert. Gleich nach der Testauswertung wird ein Zwischenbericht mit den ersten Ergebnissen erstellt. Nach der Massnahmenumsetzung wird sodann – gegen Ende des Schuljahres – ein Schlussbericht mit ergänzender und vertiefender Analyse vorgelegt. Zusätzlich erhalten die Lehrpersonen ca. zwei Monate nach der Testung die Testergebnisse für jede Schülerin und jeden Schüler ihrer Klasse (Prozentsatz der richtig gelösten Aufgaben) sowie eine Übersicht darüber, wie ihre Klasse im Vergleich zu den anderen Klassen dasteht (Inhaltsverzeichnis der Rückmeldung im Anhang C.1.1). Das Departement Bildung, Kultur und Sport erhält etwa zum gleichen Zeitpunkt einen Bericht mit anonymisierten Testergebnissen. Diese Rückmeldungen sollen dazu genutzt werden, den Unterricht wirkungsvoll weiter zu entwickeln (s.o., Zielsetzung). Dazu müssen die Lehrpersonen zunächst die Ergebnisse mit Hilfe ihrer Kenntnis über die Voraussetzungen und Bedingungen der Schülerinnen und Schüler interpretieren. Ein Fragebogen, der als Leitfaden konzipiert ist, um die Testergebnisse Schritt für Schritt zu evaluieren und zu interpretieren wird zur Verfügung gestellt. Dazu erhalten die Lehrpersonen eine Handreichung zur Messung und Beurteilung von Leistungen und nehmen an einer obligatorischen Weiterbildung zum Analysieren und Interpre-

[1] siehe Erklärung im Anhang A

tieren der Ergebnisse sowie zum Definieren der Ziele und zum Bestimmen von Massnahmen teil. Eine weitere fakultative Weiterbildung zum Reflektieren der Massnahmen und Ziele wird angeboten. Anhand der eigenen Interpretationen werden dann von den Lehrpersonen Massnahmen geplant und schriftlich fixiert, die die Unterrichtsgestaltung oder auch individuelle Förderung einzelner Schüler betreffen. Fünf Monate danach werden die Lehrpersonen aufgefordert über die Massnahmen und ihre Wirkung zu reflektieren.

Die unterstützenden Massnahmen sind im Zuge der Weiterentwicklung des Programms ausgebaut worden. Im folgenden wird diese Entwicklung nachgezeichnet:

Im ersten Jahr nach Einführung des Check 5 fand die Reflexion und Interpretation der Ergebnisse sowie die Ableitung konkreter Massnahmen in zwei Schritten auf schriftlichem Weg statt (1. zur Analyse der Testergebnisse und Ableitung von Massnahmen; 2. zur Reflexion der umgesetzten Massnahmen).

Im zweiten Jahr erfolgte eine schriftliche Befragung aller Lehrpersonen zur Analyse der Testergebnisse und Umsetzung der Massnahmen. Zudem wurde begleitend zur Umsetzungsphase eine dreiteilige fakultative Weiterbildung angeboten (1. Analyse der Ergebnisse, Ableitung konkreter Massnahmen, 2. Reflexion der Umsetzungsphase, 3. Überprüfung der Wirkung der Massnahmen, Überprüfung der Zielerreichung).

Seit dem dritten Jahr (2006/07) verpflichten sich die Lehrpersonen zum Zeitpunkt der Ergebnisrückmeldung zu einer durch die Pädagogische Hochschule der Fachhochschule Nordwestschweiz (FHNW) durchgeführten vierstündigen Weiterbildung. Ziel dieser Weiterbildung ist erstens, dass die Lehrpersonen die Testergebnisse verstehen und angemessen interpretieren. Zweitens formulieren die Lehrpersonen auf der Grundlage der Testrückmeldung auf ihren Unterricht abgestimmte Ziele und Massnahmen (siehe Anhang C.1.2), die sie im Anschluss an die Veranstaltung umsetzen. Die Weiterbildungen bieten den Lehrpersonen die Möglichkeit, sich mit anderen Berufskolleginnen und -kollegen ausserhalb des Schulhausteams auszutauschen und dadurch wichtige Impulse für ihren Unterricht zu erhalten. Zusätzlich wird während der Umsetzungsphase eine fakultative Weiterbildung mit zwei Kursblöcken angeboten. Diese Weiterbildung richtet sich an Lehrpersonen, die sich mit der weiteren Umsetzung der Massnahmen auseinandersetzen möchten. Damit wird ihnen die Möglichkeit geboten, ihre Massnahmen in der Umsetzungsphase zu reflektieren, ihre Erfahrungen auszutauschen und wichtige Impulse für die nachfolgende Realisierungsphase zu erhalten (1. Kursblock). Nach Abschluss der Realisierungsphase wird überprüft, ob mit den Massnahmen auch die damit verbundenen Ziele erreicht worden sind (2. Kursblock). Zusätzlich führt das Institut für Bildungsevaluation, Assoziiertes Institut der Universität Zürich im Frühjahr eine Befragung zur Wirkung der Massnahmen durch.

2.1.3 Einschätzung

Da die Teilnehmendenzahl im ersten Jahr 2004/2005 auf 140 Klassen beschränkt war, konnten nicht alle interessierten Klassen berücksichtigt werden. Mit der regulären Einführung im Schuljahr 2005/2006 steht das Angebot allen Lehrpersonen offen und es meldeten sich bereits vier Fünftel aller fünften Klassen dazu an. Im dritten Jahr des Angebots sind es noch einmal mehr Klassen, insgesamt 320. Lehrpersonen scheinen also einer freiwilligen Form von Leistungsmessung bei Schülern keineswegs negativ gegenüberzustehen, im Gegenteil, es scheint, als hätte dieses Angebot „einem echten Bedürfnis der Lehrerschaft entsprochen“ (Tresch & Moser 2005). In Zukunft soll der Check 5 in dieser freiwilligen Form weiterbestehen und jährlich durchgeführt werden. Die freiwillige Teilnahme ermöglicht den Lehrpersonen, eine Standortbestimmung zu bestimmten fachlichen und überfachlichen Kompetenzen ihrer Schüler zu bekommen, ohne dabei zu stark unter einen Rechenschaftslegungsdruck zu geraten.

Die Einbindung in den Gesamtrahmen eines kleinen Unterrichtsentwicklungsprojektes ist äusserst sinnvoll, da es die Lehrpersonen direkt mit den Ergebnissen ihrer Klasse konfrontiert und sie dazu drängt, sich näher damit zu beschäftigen und Massnahmen für Veränderungen zu entwickeln.

Die Unterstützungssysteme während der Reflexionsphase bzw. der Umsetzungs- und Realisierungsphase sind im Laufe des Bestehens dieses Programms immer weiter ausgebaut worden, so dass den Lehrpersonen inzwischen eine Fülle an Unterlagen und Hilfestellungen für ihre weitere Arbeit mit den Daten zur Verfügung steht. Sicherlich unterscheiden sich die Lehrpersonen in ihrer Fähigkeit, die Ergebnisse standardisierter Leistungstests richtig zu lesen und zu interpretieren. Der daraus folgenden Befürchtung, dass gerade die Lehrpersonen, die eine Anleitung oder Hilfe zur Verbesserung ihrer Unterrichtssituation am meisten benötigen, das Angebot des Check 5 am wenigsten nutzen können, wird mit adressatengerechten Begleitungen und Anleitungen begegnet.

Der Evaluation der Massnahmen im Schlussbericht des Projektes Check 5 (Tresch & Moser 2005) ist zu entnehmen, dass nicht alle der von den Lehrpersonen geplanten bzw. eingeleiteten Massnahmen tatsächlich im Zusammenhang mit den Leistungsergebnissen der Schüler stehen. Massnahmen wie z.B. „Ziele der Unterrichtsstunde an der Tafel festhalten“, „ein Witzheft zusammenstellen“ oder „die Prüfungsdauer im voraus genau festlegen“ können sicherlich sinnvoll sein, sind aber offensichtlich nicht aus den Ergebnissen eines standardisierten Leistungstests ableitbar. Häufig sind die Massnahmen jedoch durch die Reflexion mit den Testergebnissen initiiert und dienen einer Optimierung des Unterrichts. In diesem Sinne erscheint eine Kritik des fehlenden Zusammenhangs zwischen den gewählten Massnahmen und den Testergebnissen zweitrangig.

Laut Evaluationsbericht 2004/2005 wurden ca. 40 Prozent der Massnah-

men nach fünf Monaten tatsächlich vollständig realisiert, weitere 50 Prozent der Massnahmen waren teilweise umgesetzt. Die Umsetzung der Massnahmen scheint im ersten Jahr dennoch recht unterschiedlich gelungen zu sein. Einge Lehrpersonen haben beispielsweise ihre zeitlichen Ressourcen überschätzt und sich zu viel vorgenommen, andere wiederum die Verbindlichkeit der Massnahmen zu wenig ernst genommen. Hier zeigt sich erneut, wie wichtig eine gute Anleitung in der Phase der Reflexion sowie der Umsetzung und Realisierung der Massnahmen ist, um zu verhindern dass nur die Lehrpersonen davon profitieren, die es ohnehin gewohnt sind, kreative Massnahmen zu entwickeln und umzusetzen. Die unterschiedlichen Unterstützungsmassnahmen, die im Rahmen von Check 5 angeboten werden, versuchen, diesem Problem in angemessener Weise gerecht zu werden.

Eine Einbindung in ein Gesamtkonzept von Schulevaluation in dessen Rahmen auch begrenzte Aussagen zur Qualität von Schule getroffen werden können, ist in Check 5 nicht vorgesehen. Sie könnte zusätzlich dazu beitragen, eine Diskussionskultur in der Schule anzuregen, in der die einzelnen Lehrpersonen Hilfestellungen zur Interpretation der Testergebnisse und zur sinnvollen Planung und Umsetzung von Massnahmen erhalten könnten. Die Anonymität und Freiwilligkeit der Teilnahme hat, wie wir oben gesehen haben, grosse Vorteile, um bei den Lehrpersonen Interesse und Vertrauen zu entwickeln. Trotzdem begrenzt sie prinzipiell die Möglichkeiten, die sich aus einem Test für die Schulentwicklung ergeben könnten. Unabhängig von der Betrachtung der Einzelschule verhindert die freiwillige Teilnahme in der Regel auch eine vollständige und repräsentative Darstellung der Leistungen. Im Fall von Check 5 ist jedoch die Teilnehmendenanzahl trotz Freiwilligkeit so hoch, dass eine Repräsentativität gegeben ist. Hier kommen einmal mehr der Erfolg und die hohe Akzeptanz des Check 5 zum Ausdruck, die sich vermutlich auch in dem Angebot umfassende Unterstützungsmassnahmen begründen.

2.2 Check 8 und Stellwerk: Ein Test passt sich an

Der Check 8 ist ein Leistungstest für die achten Klassen der Realschulen, Sekundarschulen und Bezirksschulen. Auf den ersten Blick scheint er dem Check 5 sehr ähnlich: er dient einer Standortbestimmung der Leistungen von Schülerinnen und Schülern am Ende der achten Klasse mit dem Zweck der individuellen Förderung. Die Testbereiche sind Deutsch, Mathematik, Französisch und Englisch. Er wird innerhalb der regulären Unterrichtszeit durchgeführt und beansprucht 6 Stunden innerhalb eines Zeitfensters von 6 Wochen. Die Lehrpersonen entscheiden frei über den Zeitpunkt des Tests innerhalb des vorgegebenen Rahmens und erhalten danach ein Leistungsprofil für jeden Schüler und jede Schülerin zurückgemeldet. Das Departement Bildung, Kultur und Sport bekommt Informationen über den Ist-Zustand der Schülerleistung und kann auf dieser Basis die Steuerung im Bildungssystem

planen. Auf den zweiten Blick unterscheidet sich der Check 8 jedoch ganz erheblich von seinem Namensvetter für die fünften Klassen.

2.2.1 Zielsetzung, Organisation und Inhalt

Die Ausgangsidee des Check 8 ist es, einen Teil des Abschlusszertifikats Volksschule Kanton Aargau damit abzudecken. Unter dieser Einbettung soll er mehreren Zielen dienen (siehe www.ag.ch/abschlusszertifikat):

- Offizielle Zertifizierung der Leistung am Ende der obligatorischen Schulzeit,
- Standortbestimmung und Förderung der Lernenden,
- Unterrichts- und Schulentwicklung,
- Information aller Abnehmenden,
- Aufrechterhaltung der Lernmotivation im 9. Schuljahr,
- Verbesserung der Durchlässigkeit zwischen Oberstufentypen,
- Ab einem bestimmten Ergebnis als Beleg dafür, dass eine Schülerin/ein Schüler fähig ist für den Eintritt in eine weiterführende Schule.

Im Zusammenspiel mit den anderen Zertifikatsteilen des Abschlusszertifikats ist es auch möglich, damit die bis dato bestehende Bezirksabschlussprüfung im Kanton Aargau abzulösen. Das heisst, dass der Check 8 nach Abschluss einer Erprobungsphase von drei Jahren und Bestätigung durch den Grossen Rat im Jahr 2009 nicht mehr nur ein Instrument zur Standortbestimmung und Förderung der Schülerinnen und Schülern bleibt, sondern dass er dann auch zur Selektion verwendet wird und nicht mehr freiwillig ist. Im Vergleich zum Check 5 werden ihm also erheblich mehr Funktionen zugesprochen.

Aber nicht nur in funktionaler Hinsicht, sondern auch bezüglich der Durchführung und Datenaufbereitung unterscheidet sich dieser Test ganz erheblich von dem Check 5. Für den Check 8 werden Teile des webbasierten Leistungstests „Stellwerk“ eingesetzt, es handelt sich bei diesem Test um ein komplett computerbasiertes Instrument. Der von Stellwerk bereitgestellte Teil zu den Naturwissenschaften wird bei Check 8 ausgespart, weil der Referenzrahmen (siehe Anhang C.2.1) zu diesem Testbereich zu wenig dem Aargauer Lehrplan entspricht. Die anderen Teile aus Stellwerk werden direkt übernommen. Die Testung in diesen Bereichen erfolgt adaptiv, das heisst, dass jeder Schüler bzw. jede Schülerin automatisch die seinem oder ihrem Leistungsstand entsprechenden Aufgaben zugeteilt bekommt. Der Test

passt sich also den Leistungen der Schülerinnen und Schüler an. Das computerbasierte System ermöglicht es, sofort nach Durchführung des Tests das jeweilige Leistungsprofil abzurufen, ohne dass durch das Verschicken, Einlesen und Analysieren von Fragebögen weitere Zeit eingeplant werden muss. Das System „Stellwerk“ wurde im Kanton St. Gallen speziell dafür entwickelt, eine Standortbestimmung für Schülerinnen und Schüler vorzunehmen, an der sich unter anderem zukünftige Arbeitgeber bei der Besetzung ihrer Lehrstellen orientieren können. Die Leitung, die Organisation und der Vertrieb liegt beim Lehrmittelverlag St. Gallen. Die Aufgaben, die von einer Gruppe von Lehrpersonen aus dem Kanton St. Gallen entwickelt wurden, wurden im März und April 2005 an Schülern des Kantons St. Gallen geeicht und für das adaptive Testen kalibriert. Alle achten Klassen des Kantons nahmen daran teil. Der Test wird in Kooperation mit dem Institut für Bildungsevaluation, Assoziiertes Institut der Universität Zürich für die messmethodischen Fragen und der Firma Cybersystems für die Internetplattform durchgeführt. Check 8 beinhaltet in seiner ersten Durchführungsphase im Sommer 2006 zusätzlich zu den Stellwerk-Aufgaben auch einen nicht computerisierten 45-minütigen Schreibanlass in Deutsch, der von externen Experten ausgewertet wird. In einer zweiten Durchführungsphase wird der Check 8 zusätzlich auch mündliche Teilbereiche und evtl. eine modifizierte Version des naturwissenschaftlichen Bereichs aus Stellwerk einschliessen.

2.2.2 Durchführung

Der Test wird am Ende des Schuljahres in der Regel im Mai und in der ersten Hälfte des Juni durchgeführt. Die Lehrpersonen bekommen für den webbasierten Teil rechtzeitig eine Durchführungsanleitung und Pin-Nummern für die Schülerinnen und Schüler sowie ein Generalpasswort, mit dem sie die Tests nach eventuellen Systemproblemen wieder freischalten können. Die Testung kann, je nach Ausstattung der Schule, entweder im Klassenverband als auch geteilt in Einzelgruppen durchgeführt werden. Vorausgesetzt wird, dass jeder Schüler für die Zeit der Testung über einen eigenen Computerarbeitsplatz mit Internetverbindung, Audioausgang und Kopfhörern verfügt. Für jeden Testteil des Stellwerk-Tests stehen 90 Minuten zur Verfügung. Dabei ist jedoch die Zeit, die der Schüler für die Bearbeitung der Aufgaben benötigt, nicht entscheidend für das Ergebnis des Tests. Es kommt lediglich darauf an, ob die bearbeiteten Aufgaben richtig oder falsch gelöst wurden. Das adaptive Testsystem bietet jedem Schüler am Anfang eine Aufgabe des mittleren Schwierigkeitsbereiches an. Die Schwierigkeitsparameter der Aufgaben sind zuvor im Rahmen der Kalibrierung der Items an der St. Gallener Population bestimmt worden. Die Anordnung der nun folgenden Aufgaben wird durch einen unter Anwendung der Item-Response-Theory[2] provisorisch geschätzten Personenfähigkeitswert gesteuert. Das heisst, wenn ein Schüler

[2] siehe Erklärung im Anhang A

die erste Aufgabe richtig beantwortet, bekommt er als zweites eine etwas schwierigere Aufgabe. Wenn er sie hingegen falsch beantwortet hat, so sinkt das Schwierigkeitsniveau. Es werden solange Aufgaben bereitgestellt, bis das System eine befriedigende Schätzung der Personenfähigkeit abgeben kann. Auf diese Weise passt sich der Test schrittweise an die Fähigkeit der jeweiligen Schülerinnen und Schüler an.

Ein adaptiver Test hat den grossen Vorteil, dass durch diese Anpassung der Messfehler für jedes Fähigkeitsniveau gleichermassen niedrig ist. Bei klassischen Tests steigt der Messfehler hingegen im höheren und im niedrigeren Leistungsbereich. Dies liegt daran, dass der klassische Test sich an den durchschnittlichen Schülerleistungen orientieren muss, um eine relativ präzise Aussage über die Leistungen der meisten Schüler zu ermöglichen. Da klassische Papier- und Bleistifttests in der Anzahl der Aufgaben beschränkt sind (alle Schüler bzw. grosse Gruppen von Schülern müssen ja dieselben Aufgaben lösen), können sie in den Randbereichen weniger schwere oder leichte Aufgaben anbieten. Die Messungen werden also in diesen Bereichen stärker von Messfehlern belastet.

Bei adaptiven Tests passt sich hingegen der Test an die Fähigkeiten des Schülers an und stellt genügend Aufgaben bereit, die exakt in das individuelle Fähigkeitsspektrum passen. Voraussetzung für eine gute Messung ist jedoch ein entsprechend grosser Itempool, der tatsächlich die Möglichkeit bietet, für jedes Niveau die entsprechenden Aufgaben anzubieten. Jede einzelne Fähigkeitsdimension muss also durch einen entsprechend grossen Itempool abgedeckt sein und es muss sichergestellt sein, dass die verwendeten Items tatsächlich auf unterschiedlichen Schwierigkeitsstufen für dieselbe Fähigkeitsdimension repräsentativ sind. Die Anforderungen an die Qualität der Testaufgaben sind wesentlich höher als bei herkömmlichen Tests, da beim adaptiven Testen weniger Aufgaben bearbeitet werden und diese sich auch von Schüler zu Schüler unterscheiden. Wenn ein Schüler eine gute Testaufgabe bekommt und ein anderer eine qualitativ schlechtere Testaufgabe, dann kann dies zu extremen Unterschieden bei der Bewertung führen, obwohl diese Unterschiede sich nicht in den tatsächlichen Fähigkeiten widerspiegeln. Zusätzlich zu der Anforderung, Items in höchster Qualität und hoher Anzahl für jede Fähigkeitsstufe bereitzuhalten, kommt beim adaptiven Test noch hinzu, dass der Itempool genau den Bedingungen des zugrundeliegenden psychometrischen Modells entsprechen muss (vgl. Flaugher 2000) .

Nach der Durchführung der Stellwerktests, wird den Lehrpersonen direkt über das System eine Rückmeldung als pdf-Datei erteilt, die sie als Grundlage für die Förderung der einzelnen Schülerinnen und Schüler nutzen können (siehe Anhang C.2.2). Um diese Förderung gezielter zu gestalten, werden dazu Weiterbildungen angeboten und Handreichungen bereitgestellt. Auch gibt es Beispiel- und Übungsaufgaben, die über den Lehrmittelverlag St. Gallen zu den entsprechenden Fähigkeitsbereichen und Niveaustufen kostenlos im Internet angeboten werden (siehe www.stellwerk-check.ch).

Der nicht im Stellwerk enthaltene Schreibanlass Deutsch wird hingegen an zwei festgelegten Tagen durchgeführt und beansprucht 45 Minuten. Es gibt drei Schreibanlassthemen, die per Zufallsgenerator den Lernenden zugewiesen werden. Die Auswertung erfolgt anschliessend durch externe Experten, wobei auch der Schreibanlass und die Auswertungsvorgaben so konzipiert sind, dass die Ergebnisse wie die Ergebnisse der webbasierten Aufgaben kriterienorientiert und gruppenunabhängig ausgewertet werden können.

2.2.3 Einbindung in die externe Schulevaluation

Im Herbst 2006 wurde erstmals der Stellwerktest im Rahmen einer externen Schulevaluation an der Schule Suhr in jeweils zwei dritten Klassen der Real-, Sekundar- und Bezirksschule eingesetzt. Ein weiterer Einsatz ist im Rahmen der externen Schulevaluation in Wettingen für März 2007 geplant. Die Auswahl dieses Tests ist eine pragmatische, da Stellwerk im Kanton Aargau als Teil des Abschlusszertifikats in Erprobung ist.

Die Interpretation und Beurteilung ist mit den zur Verfügung stehenden Ergebnissen aus diesem Probelauf noch relativ limitiert. Hinsichtlich der kriterialen Bezugsnorm besteht einzig die Möglichkeit zum Abgleich der Schülerleistungen mit den von Stellwerk bereitgestellten Interpretationshilfen (siehe Anhang C.2.3). Die enge Anbindung dieser Interpretationshilfen an den St. Gallener Lehrplan bringt jedoch für den Einsatz im Kanton Aargau eine ernstzunehmende Schwierigkeit mit sich. Hinsichtlich der sozialen Norm können zu dem jetzigen Zeitpunkt Vergleichswerte ebenfalls nur aus dem Kanton St. Gallen bereitgestellt werden. Daneben kommt ein Vergleich lediglich zwischen den zwei Parallelklassen innerhalb einer Schulstufe, bzw. zwischen den Schulstufen in Frage. Hinsichtlich einer Beurteilung der Leistungswerte der Schülerinnen und Schüler nach der individuellen Norm gibt es zum jetzigen Zeitpunkt noch keine Möglichkeiten, da bisher die Daten nur zu einem einzigen Messzeitpunkt erhoben wurden.

Dennoch konnten einige wichtige Erkenntnisse zur Durchführung von Leistungsmessungen im Rahmen von externer Schulevaluation aus dieser ersten Anwendung gezogen werden. Für die einzelnen Schulstufen der Schule Suhr zeigen sich unterschiedliche Verteilungsformen und unterschiedlich breite Verteilungen der Leistung, die im Rahmen der externen Evaluation Anlass für gezielte Gespräche mit den Lehrpersonen gegeben haben. Eine hohe Leistungsstreuung ist dabei durch eine hohe Heterogenität der Schülerinnen und Schüler zu erklären, während eine engere Verteilung der Leistung auf eine eher homogene Schülergruppe schliessen lässt. Auch die Formen der Verteilungen (eingipflig bzw. zweigipflig, rechts- bzw. linksschief) können Anlass für vertiefende Analysen im Rahmen der externen Schulevaluation sein. Auch wenn die Lehrpersonen, die im Rahmen der Evaluation in Suhr befragt wurden, nicht alle Besonderheiten der beobachteten Verteilungen aufklären konnten, gaben ihre Antworten doch Hinweise auf eine sinnvolle Interpreta-

tion in dem vom Testdesign vorgegebenen engen Rahmen.

Da es sich jedoch nicht um Längsschnittsdaten handelt, kann eine Aussage über die Qualität der Schule oder des Unterrichts nicht getroffen werden. Es ist möglich, dass die gemessenen Leistungen lediglich auf der zufälligen demographischen Zusammensetzung der Schülerschaft beruhen. Um Aussagen über Schul- und Unterrichtsqualität, auf die es ja bei der externen Evaluation der Schule besonders ankommt, machen zu können, wird es zukünftig nötig sein, Längsschnittsdaten und Hintergrundvariablen zu erheben.

2.2.4 Einschätzung

Der methodische Ansatz des adaptiven Testens, der im Stellwerktest verwirklicht ist, ist zumindest in Europa noch relativ neu und ambitiös. Bei Beachtung der notwendigen Voraussetzungen, ein ausreichend grosser Itempool, hohe Qualität der Items und eine breite Abbildung des psychometrischen Modells, bietet diese Methode ausserordentliche Vorteile. Der Messfehler ist für alle Fähigkeitsbereiche minimal, jeder Schüler kann unterschiedliche Aufgaben bearbeiten und ist trotzdem auf derselben Metrik mit anderen Schülern vergleichbar, Schulen können im Prinzip einen eigenen passenden Unteritempool auswählen und trotzdem gleiche Kompetenzen messen lassen und schliesslich ist es wegen der grossen Anzahl der Items und der variablen Itemabfolge innerhalb der Tests schwieriger, sich im Vorwege auf den Test vorzubereiten. Wenn die Voraussetzung allerdings nicht eingehalten sind, kann es, leichter als bei klassischen Tests, zu Fehleinschätzungen über die Leistung in den jeweiligen Dimensionen kommen.

Bei Stellwerk ist die Anzahl der Items noch relativ gering, so dass zumindest die detaillierte Darstellung unterschiedlicher Fähigkeitsdimensionen in einem Fachgebiet fragwürdig bleibt. Der Itempool wird zwar laufend, teils durch Selbstkalibrierung, erweitert, ist aber mit zur Zeit knapp über 1000 kalibrierten Items in vielen Bereichen zu klein, um die unterschiedlichen Fähigkeitsdimensionen innerhalb von 5 Schulfächern (die in Check 8 getesteten zuzüglich Natur und Technik) abzubilden. Dazu kommt, dass die der Auswahl der Items zugrundeliegenden didaktischen Überlegungen in der Dokumentation zu Stellwerk nicht weiter ausgeführt sind. Ein lernzielorientierter Referenzrahmen mit Aufgabenbeispielen dokumentiert das Vorgehen, allerdings fehlt ein didaktisch begründetes Kompetenzraster als Basis des psychometrischen Modells. Nur für den Bereich der Fremdsprachen werden die Items den Vorgaben des europäischen Sprachportfolios angeglichen.

Ein weiteres Problem bei der Messung der Leistung ist das Fehlen einer Geschwindigkeitskomponente. Zwar ist der Druck auf die Schüler auf diese Weise etwas geringer, aber dennoch ist fraglich, ob die Leistungen zweier Schüler als vergleichbar eingeschätzt werden können, wenn sie zwar beide gleichermassen eine Aufgabe richtig lösen können, der eine aber 10 Minuten dafür braucht, während der andere es in einer Minute schafft.

Ähnlich wie der Check 5 ist der Check 8 nicht in einen Rahmen zur Schulevaluation eingebunden. Präzise Aussagen über die Qualität von Schule und Unterricht sind auf der Basis der Ergebnisse aus diesem Test nicht möglich, da zu viele Einflussfaktoren (z.B. zufällige demographische Zusammensetzung der Schülerschaft) für die Ursachen der punktuellen durchschnittlichen Schülerleistung in Frage kommen. Eventuelle Einflussvariablen für Schülerleistung wie soziale Indikatoren werden nicht in die Beurteilung aufgenommen. Es werden insgesamt nur drei Hintergrundvariablen (IFS-Schüler, Fremdsprachigkeit und Schulform) erhoben, jedoch bei der Rückmeldung an die Lehrpersonen im Rahmen von Check 8 nicht berücksichtigt.

Als Konsequenz der Anwendung des adaptiven Systems können nur globale Leistungswerte der einzelnen Schülerinnen und Schüler zurückgemeldet werden, d.h. die Lehrperson bekommt für jeden Fähigkeitsbereich pro Schüler einen Wert zurückgemeldet, ohne zu wissen, welche Antworten von welchen Testaufgaben in diese Berechnung eingeflossen sind. Anhand der Beispielaufgaben in den entsprechenden Schwierigkeitsbereichen kann nun die Lehrperson das Fähigkeitsniveau ihrer Schüler einschätzen. Allerdings haben die Schüler in der Regel unterschiedliche Aufgaben bearbeitet, so dass die Beurteilung und Interpretation der Ergebnisse für die Lehrpersonen eine extrem anspruchsvolle Aufgabe wird. Im Gegensatz zum Check 5 haben die Lehrpersonen hier nicht die Möglichkeit, die Aufgaben anzusehen und die speziellen Schwierigkeiten der eigenen Klasse zu jeder Aufgabe zu analysieren. Die Rückmeldung bleibt auf einem relativ abstrakten Niveau und bietet vermutlich für die Lehrpersonen nur wenig Anregung zur Optimierung ihrer Förderung und ihres Unterrichts.

Bei jeder Form der Auswertung mit probabilistischen Testmodellen ergibt sich die Schwierigkeit, dass das Zustandekommen eines Testwertes für Nicht-Experten kaum noch nachvollziehbar ist. Diese Schwierigkeit verstärkt sich jedoch bei der Anwendung von adaptiven Testmethoden noch erheblich, weil nun nicht einmal mehr ein einheitlicher, für alle Schülerinnen und Schüler gleicher Test zugrunde gelegt werden kann.

Ein Instrument wie Stellwerk eignet sich, wenn die zugrundeliegenden Dimensionen didaktisch begründet sowie ausreichend mit passenden Items abgedeckt sind und zusätzliche beeinflussende Variablen berücksichtigt werden, hervorragend zur vergleichenden Beurteilung von Schulen oder Klassen. In diesem Sinne könnte es auch evtl. als Längsschnittmessung Aussagen über die Qualität von Schule zulassen und damit für die externe Schulevaluation nutzbar gemacht werden. Wegen der hohen Abstaktion der Ergebnisrückmeldung ist es jedoch schwierig, daraus eine Grundlage für die Förderung einzelner Schüler abzuleiten. Die fehlende Nachvollziehbarkeit der Testergebnisse auf Grundlage einzelner bearbeiteter Items wird auch ein Problem bei der Verwendung dieses Tests als Selektionsinstrument darstellen.

2.3 Orientierungsarbeiten in der Zentralschweiz

Die Bildungsplanung Zentralschweiz stellt seit 2001 Orientierungsarbeiten für das 2. bis 9. Schuljahr zum Zweck der formativen Leistungsbeurteilung bereit. Mithilfe der Orientierungsarbeiten wird überprüft, ob einzelne im Lehrplan fixierte Lernziele durch die Schülerinnen und Schüler erreicht wurden. Als grösster Unterschied zwischen den beiden vorher beschriebenen Modellen der Leistungsmessung in der Schweiz ist hier die Nutzung als Diagnoseinstrument für eine stark individualisierte Förderplanung hervorzuheben. Die Lehrpersonen werten die Arbeiten selbst aus und nutzen die Ergebnisse, um den Unterricht besser auf die Bedürfnisse der Schüler abzustimmen. Darüber hinaus können auch die Lernenden anhand eines vorgegebenen Kriterienrasters ihre Fähigkeiten selbst einschätzen und erhalten darüber die Möglichkeit, ihr eigenes Lernen besser zu steuern.

2.3.1 Zielsetzung, Organisation und Inhalt

Die Orientierungsarbeiten dienen der Überprüfung des individuellen Lernstandes der Schülerinnen und Schüler und sollen den Lehrpersonen eine Rückmeldung über den Unterrichtserfolg und dadurch über die Qualität des Unterrichts geben. Damit können sie als Diagnoseinstrument die Grundlage für abzuleitende Fördermassnahmen bilden. Den Lernenden bieten sie einen Ansatzpunkt für die Weiterentwicklung ihrer Leistungen und stärken durch das Selbsteinschätzungsraster am Ende jeder Aufgabe zusätzlich die Fähigkeit zur Selbstreflexion. Die Aufgaben der Orientierungsarbeiten haben darüber hinaus Modellcharakter, so dass sie von den Lehrpersonen als Beispiele zur Entwicklung neuer Übungs- und Prüfungsaufgaben herangezogen werden können.

Die Arbeiten werden in Form von Broschüren über den Kantonalen Lehrmittelverlag Luzern vertrieben und können von den Schulen bzw. den interessierten Lehrpersonen dort bezogen werden. Die in den Broschüren enthaltenen Aufgaben sind jeweils bestimmten Lernzielen des Lehrplans zugeordnet. Sie enthalten weiterhin unterschiedliche Schwierigkeitsstufen und schliessen alle mit einem dreistufigen Lernzielraster ab, das dazu dient, die Aufgaben strukturierter auswerten und beurteilen zu können.

Bisher liegen für den Primarbereich Orientierungsarbeiten für die Bereiche Deutsch, Mathematik, Mensch und Umwelt, Musik, bildnerisches Gestalten und technisches Gestalten vor. Für einige Klassen der Sekundarstufe gibt es zusätzlich Orientierungsarbeiten im Bereich Naturlehre, Geographie und Hauswirtschaft (Geschichte und Politik ab 2007 und Lebenskunde ab 2009) (siehe Anhang C.3.1).

2.3.2 Durchführung

Beim Einsatz der Orientierungsarbeiten wird zwischen zwei Situationen unterschieden:

- Durchführung im Rahmen der Qualitätssicherung (freiwillige Teilnahme),
- Durchführung im Rahmen des Übertrittverfahrens (obligatorisch)

In der Regel ist eine Aufgabe pro Broschüre auszuwählen. Die Schülerinnen und Schüler bekommen so viel Zeit, wie sie benötigen, um alle Teile der Aufgabe lösen zu können. Das Arbeitstempo wird ähnlich wie im Rahmen des Stellwerktests als untergeordnetes Kriterium bei der Erfassung der Schülerleistung angesehen.

Qualitätssicherung im Unterricht

Als Orientierungsarbeiten im Rahmen der Qualitätssicherung im Unterricht, können die Aufgaben innerhalb des Klassenverbandes oder nur von einzelnen Schülern bearbeitet werden. Beispielsweise wird im Merkblatt des Amts für Volksschulbildung, Luzern vorgeschlagen, Orientierungsarbeiten zu Beginn einer Unterrichtseinheit, zwischendurch oder auch am Ende einer Lerneinheit einzusetzen. Wenn die Orientierungsarbeiten am Anfang einer Lerneinheit platziert werden, kann damit das Bild der Lernausgangslage gezeichnet werden, dessen Betrachtung für die Planung der Lehr- bzw. Lerneinheit nützlich ist. Die Ergebnisse aus einer Zwischenevaluation können hingegen genutzt werden, um individuelle Fördermassnahmen einzuleiten. Ein abschliessender Einsatz der Orientierungsarbeiten kann schliesslich einen Eindruck über den Grad der Zielerreichung vermitteln. Orientierungsarbeiten können mehrmals jährlich mit unterschiedlichen Aufgaben aus den verschiedenen Lernbereichen durchgeführt werden. Auch ist es möglich, sie an Stelle anderer Lernkontrollen treten zu lassen. Die Resultate können in die Berechnung der Zeugnisnoten eingehen, auch wenn die Orientierungsarbeiten nicht grundsätzlich für die Notengebung konzipiert wurden.

Qualitätssicherung der gesamten Schule

Als Instrument der Qualitätssicherung der gesamten Schule können die Orientierungsarbeiten einen Vergleich einzelner Klassen ermöglichen und in einem grösseren Rahmen das Erreichen der gesetzten Unterrichtsziele überprüfen. Dazu planen die Schulleitungen den Einsatz der Orientierungsarbeiten an ihren Schulen, so dass ein geregelter Ablauf, eine Vergleichbarkeit der Daten und eine gemeinsame Reflexion der Ergebnisse gewährleistet ist.

Übertrittsverfahren

Die Durchführung von Orientierungsarbeiten im Übertrittsverfahren ist im Kanton Luzern obligatorisch. Dazu werden im Zeitraum vom 5. zum 6. Schuljahr (erste Hälfte) mindestens sechs Aufgaben bearbeitet. Drei davon sind aus dem Bereich Deutsch zu wählen, zwei aus dem Bereich Mathematik und eine aus dem Bereich Mensch und Umwelt.

2.3.3 Einschätzung

Die Orientierungsarbeiten in der Zentralschweiz sind als Instrumente der individuellen Diagnose und der Diagnose auf Einzelschulebene konzipiert. In dieser Funktion stellen sie eine sehr gute Grundlage für die Beurteilung der Schülerleistungen hinsichtlich bestimmter Lernziele dar und helfen sowohl den Lehrpersonen als auch den Lernenden, die Leistungen besser einzuordnen. In diesem Sinne sind die Orientierungsarbeiten ein Instrument, dessen Einsatz sehr zu begrüssen ist.

Die Orientierungsarbeiten haben hingegen nicht den Anspruch, einen Vergleich zwischen unterschiedlichen Schulen herzustellen. Die Objektivität, die für ein solches Vorhaben notwendig wäre, kann bei der gegebenen Durchführung nicht erreicht werden und ist auch nicht beabsichtigt. Auch die einzelnen Aufgaben der Orientierungsarbeiten sind als Prüfungsaufgaben und nicht als Testaufgaben konzipiert und deshalb auch nicht teststatistisch abgesichert. Dennoch ist es vorstellbar, dass ein solches Instrument auch im Rahmen einer externen Schulevaluation wertvolle Informationen zum Lernstand der Schülerinnen und Schüler hinsichtlich einzelner Lernziele geben könnte. Für grösser angelegte Vorhaben, die über Bildungsmonitoring Informationen für die Schulsystemsteuerung zur Verfügung stellen, sind die Orientierungsarbeiten hingegen vermutlich nicht geeignet.

2.4 Zusammenfassung und Beurteilung

In der deutschsprachigen Schweiz gibt es zur Zeit vielschichtige Bemühungen, Schülerleistungen systematisch zu erfassen und zu verarbeiten. Drei charakteristische Modelle, die auf verschiedenen Ebenen des Bildungssystems in unterschiedlicher Weise wirksam sind, wurden vorgestellt. Auch wenn diese Auswahl keinesfalls alle in der Deutschschweiz verwendeten Modelle erfasst, unterstreicht sie deutlich die Variabilität der einzelnen Modelle hinsichtlich ihrer Wirkung auf unterschiedlichen Ebenen.

Die in der Zentralschweiz eingesetzten Orientierungsarbeiten sind ein gutes Instrument, um auf Einzelschulebene oder Klassenebene die Leistungen der Schüler in bestimmten inhaltlichen Bereichen zu diagnostizieren und Hinweise für eine individuelle Förderung zu bekommen. Durch die Anlage der Untersuchungen schliesst sich der Vergleich mit einem weiteren Umfeld (wie z.B. der Vergleich mit anderen Schulen) aus. Damit schränkt sich auch der

Wert der Verwendung dieses Modells im Rahmen von externen Schulevaluationen ein. Günstiger wären im Zusammenhang mit externer Schulevaluation objektivere Messungen, um die spezifische Situation einer Schule im Vergleich zu anderen besser beurteilen zu können.

Der Check 5 ist ein solches Instrument, dass sowohl auf der Einzelschulebene als auch schulübergreifend Informationen zur Diagnose und Förderung bereitstellt. Da die Aufgaben zwar nicht von den Lehrpersonen selbst ausgewertet werden, aber dennoch hinsichtlich ihres spezifischen Schwierigkeitsgrades für einzelne Schüler oder Klassen analysiert werden können, würden sie sich sowohl für die individuelle Diagnose und Förderung als auch für einen übergreifenden Systemvergleich eignen. Die Anzahl der Funktionen eines Leistungstest ist jedoch in der Regel begrenzt. Eine Verwendung von Check-5-Ergebnissen zum Zweck eines schulübergreifenden Vergleichs würde sicher das Vertrauen zerstören, dass Lehrpersonen bisher in dieses Modell legen. Ohne dieses Vertrauen wäre aber auch die Möglichkeit der Nutzung des Instruments zur Individualdiagnose und -förderung nicht mehr gegeben. Die Verknüpfung von externer Schulevaluation mit den Erkenntnissen aus Tests wie Check 5 ist deshalb höchst problematisch. Es ist dabei immer zu berücksichtigen, welchen Einfluss die Nutzung zusätzlicher Funktionen auf die herkömmlichen Funktionen des Instruments hat.

Mit dem Check 8 ist von vornherein die Idee des Schulsystemvergleichs und der Zertifizierung bzw. Selektion verbunden. Ob oder inwiefern dies die Funktionen der Förderung und Unterrichtsentwicklung beschränken kann, wird sich im Laufe der Verwendung dieses Modells noch genauer zeigen. Die Anlage des Tests als adaptives Messinstrument begrenzt zumindest die Möglichkeiten der Lehrpersonen, die Testergebnisse konkret nachzuvollziehen und daraus Schlüsse für eine individuelle Förderung abzuleiten. Andererseits könnte ein gut ausgebautes adaptives Testsystem mit mehreren Testzeitpunkten in der Bildungslaufbahn der Schülerinnen und Schüler und einem fundiertem didaktischen Kompetenzmodell ein sehr nützliches Instrument für die externe Schulevaluation sein, indem es gesicherte vergleichbare Informationen über Leistungsstand und Leistungsentwicklung bereitstellte. Check 8 steht als Modell erst am Anfang eines breiteren Einsatzes, mögliche Funktionen und Weiterentwicklungen des Modells sind sicherlich noch zu erwarten.

Insgesamt zeigt sich, dass in der deutschsprachigen Schweiz sehr unterschiedliche Modelle zum Einsatz kommen. Eine Einheitlichkeit, wie sie unter den Vorgaben der harmonisierten Schule notwendig wäre, ist bisher noch nicht gegeben. Da auch die externe Schulevaluation noch keine lange Tradition hat, weist sie kaum Verknüpfungspunkte zu den vorhandenen Modellen der Leistungsmessung auf. Der Blick in andere Länder kann evtl. einige Anregungen zur Gestaltung einer möglichen zukünftigen Verknüpfung dieser Bereiche liefern.

3 Blick in andere Länder

3.1 Beispiele aus Deutschland

Die Anfänge der standardisierten Leistungsmessung

In Deutschland hat die Schulleistungsmessung zwar eine längere Tradition, doch wurden die Ergebnisse aus Schulleistungsstudien bis Mitte der 90er Jahre kaum öffentlich diskutiert. Ausgelöst durch den TIMSS-Schock (Third International Mathematics and Science Study) ändert sich diese Haltung und es setzt sich allmählich die Überzeugung durch, dass Qualitätssicherung und Qualitätsentwicklung im Bildungswesen zu den vorrangigen Aufgaben der Politik gehört. Einen ersten Vorstoss macht der Hamburger Senat 1996 mit dem Auftrag, Kompetenzen von Schülerinnen und Schülern eines gesamten 5ten Klassenjahrgangs in Hamburg zu erheben (Lehmann und Peek 1997). Mit den Ergebnissen über die Fähigkeiten der Schülerinnen und Schüler am Übergang in die Sekundarstufe I sollte umfangreiches Steuerungswissen für die Bildungsplanung zur Verfügung gestellt werden. Erstmalig werden darüber hinaus klassenbezogene Ergebnisse den Einzelschulen zurückgemeldet, so dass sie den Lehrern sowohl als Information über den Leistungsstand ihrer Klasse als auch als Ausgangspunkt für eine Diskussion um Schul- und Unterrichtsentwicklung zur Verfügung stehen. Die Hamburger Studie „Aspekte der Lernausgangslage“, die auch unter dem Akronym LAU bekannt ist, wurde in den Jahrgangsstufen 5, 7, 9, 11 und 13 mit denselben Schülerinnen und Schülern weitergeführt, so dass über den reinen Vergleich der Schulleistungen innerhalb des Stadtstaates Hamburg hinaus auch eine Beurteilung der Lernentwicklung der Schülerinnen und Schüler von der 5ten bis zur 11ten Klasse möglich ist (Lehmann und Peek 1997; Lehmann, Gänsfuß und Peek 1999; Lehmann et al. 2001). Gleichfalls als Vollerhebung wurde im Jahr 2000 der 8te Jahrgang der Rheinland-Pfälzischen Schülerinnen und Schüler im Fach Mathematik getestet. Auch im Rahmen der Untersuchung „Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext“ (MARKUS) wurden die Ergebnisse zum Ziele des Anschubs einer Entwicklung in den Einzelschulen zurückgemeldet (Helmke und Jäger 2002). Im Land Brandenburg wurden 1999 im Rahmen der Qualitätsuntersuchung an Schulen zum Unterricht in Mathematik (QuaSUM) Stichproben von Schülerinnen und Schülern in ganzen Klassen der gesamten Jahrgänge 5 und 9 der gezogenen Schulen untersucht (Lehmann et al. 2000). Auch hier wurden den Lehrerinnen und Lehrern die klassenbezogenen Ergebnisse zur Verfügung gestellt. Eine zusätzliche, eigenständige, kombiniert qualitativ als auch quantitativ durchgeführte Untersuchung erhebt den Umgang des Kollegiums mit den rückgemeldeten Daten (Peek 2004; Nilshon 2004). Dennoch ist in den Konzepten dieser hier beschriebenen Studien eine Schulentwicklung nicht konkret vorgesehen. Ziel dieser Untersuchungen ist stärker die Information über das Bildungssystem und die Rückmeldung

über die Schülerleistungen an die Bildungsadministration als der Anstoss einer konkreten Schulentwicklung und Qualitätssteigerung in der Einzelschule. „Derzeit ist nicht bekannt, ob sich [...] im Regelfall eine konstruktive Zusammenarbeit zwischen Schulaufsicht, Schulleitungen und Lehrkräften im Sinne einer forcierten Schulentwicklungsarbeit herstellen lassen." (Lehmann 2001, 64).

Vergleichsarbeiten

Eines der Hauptziele der Vergleichsarbeiten, deren bekanntestes Beispiel in Deutschland die Studie VERA der Universität Koblenz-Landau ist, besteht hingegen in dem Forcieren des Schulentwicklungsprozesses und in der Steigerung der Qualität der Einzelschule. Anhand des innerschulischen Vergleichs von Parallelklassen und -schulen lassen sich damit relative Stärken und Schwächen der Schülerinnen und Schüler erkennen. Auf dieser Grundlage kann innerhalb der Schule versucht werden, die möglichen Gründe für die erzielten Leistungsstände zu analysieren. Im Jahr 2003 in einem Fach in Reinland-Pfalz eingeführt, werden die Vergleichsarbeiten inzwischen bereits in 7 Bundesländern und in den Fächern Mathematik und Deutsch geschrieben. Dennoch gibt es in Deutschland noch keine einheitlich geregelten Standards für Leistungsmessungen und die Verwendung ihrer Ergebnisse für die Schulentwicklung. Auch wenn in den letzten Jahren verstärkt eine Tendenz zur Zusammenarbeit und Einheitlichkeit zu erkennen ist, liegt die Entscheidung für oder gegen bestimmte Formen der Leistungsmessung in Schulen, sowie deren Verwendung für die Qualitätsentwicklung der Schulen allein bei den einzelnen Bundesländern.[3] Um trotz der föderalen Entscheidungsbefugnis der Länder eine gewisse Einheitlichkeit in der Bildung und Chancengleichheit über die Ländergrenzen hinweg zu gewährleisten, wird im Jahr 2004 das Institut zur Qualitätsentwicklung im Bildungswesen (IQB) in Berlin gegründet. Dieses Institut ist dafür zuständig, auf nationaler Ebene Aufgaben zu entwickeln, um die von der Kultusministerkonferenz beschlossenen Standards zu überprüfen und weiterzuentwickeln.

In den folgenden Abschnitten soll auf die Aufgaben des IQB sowie auf ausgewählte Modelle der Leistungsmessung näher eingegangen werden. Dazu werden exemplarisch zwei Modelle vorgestellt, in deren Konzeption die Qualitätsentwicklung der Einzelschulen vorgesehen ist. VERA steht für Vergleichsarbeiten, während die Lernstandserhebungen in Nordrhein-Westfalen mit einem ähnlichen Konzept die Jahrgangsstufe 9 näher beleuchten. Schliesslich wird die Konzeption des Instituts für Bildungsmonitoring in Hamburg vorgestellt, in dem zukünftig Leistungsmessungsdaten mit externen Schule-

[3] Im Anhang D.1 findet sich das Positionspapier des deutschen „Netzwerks Empiriegestützte Schulentwicklung" (EMSE) zu den zentralen standardisierten Lernstandserhebungen. Dieses Papier enthält auch eine Übersicht über die Formen der Lernstandserhebungen in den einzelnen Bundesländern.

valuationsdaten gemeinsam verarbeitet werden sollen.

3.1.1 Das IQB: Ein neues Institut zur Qualitätsentwicklung im Bildungswesen

Das IQB wurde im Jahr 2004 mit dem Ziel gegründet, die Länder in ihren Bemühungen um die Sicherung und Steigerung der Qualität schulischer Bildungsprozesse zu unterstützen. Das an der Berliner Humboldt-Universität angebundene Institut ist eine wissenschaftliche Einrichtung der Länder der Bundesrepublik Deutschland. Im Rahmen der oben genannten Zielstellung werden die von den Ländern festgelegten Bildungsstandards hier normiert, illustriert und weiterentwickelt. Ausserdem hat das IQB noch einen Forschungsauftrag, der sich auf die Gebiete der Lehr-Lernforschung, der pädagogischen Diagnostik und der Implementation und Evaluation der Bildungsstandards bezieht (IQB 2005).

Zielstellung

Im IQB werden Aufgaben und Beispiele konstruiert, die zum einen dazu geeignet sind, die Bildungsstandards im Unterricht zu implementieren und zum anderen dazu, die Kompetenzen der Schülerinnen und Schüler zu erfassen und damit den Grad, in dem die Standards erreicht werden, zu überprüfen.

Lernaufgaben

Ein Teil der entwickelten Aufgaben soll dementsprechend dazu dienen, die Leistungen zu veranschaulichen, die zur Erreichung der jeweiligen Standards notwendig sind. Diese Aufgabenbeispiele sind dazu gedacht, die Lehrpersonen bei ihrer Unterrichtsplanung und -gestaltung zu unterstützen. Sie sollten nicht als Testaufgaben, sondern vor allem zur Einführung und Vertiefung des Unterrichtsstoffes verwendet werden und somit eine an den Standards orientierte Unterrichtskultur unterstützen.

Testaufgaben

Der andere Teil der entwickelten Aufgaben dient hingegen zu Testzwecken. Diese Aufgaben sind häufig weniger ausführlich als die oben beschriebenen Beispielaufgaben, erlauben dafür aber eine relativ präzise Messung einzelner Kompetenzbereiche. Nach einer umfangreichen Testentwicklungsphase mit mehreren gross angelegten Feldtests können die Aufgaben zur Normierung der Standards verwendet werden. Normierung heisst, dass anhand einer mehrere tausend Schüler umfassenden Stichprobe ermittelt wird, wie viele Schüler die Standards erreicht haben, bzw. nicht erreicht haben. Daraus ergibt sich ein verlässlicher Vergleichsmassstab, der nationale Geltung hat. Die einzelnen Bundesländer können anschliessend ausgewählte Aufgaben für ihre

Vergleichsarbeiten verwenden und erhalten damit die Möglichkeit, die Schülerleistungen mit einer nationalen Norm in Beziehung zu setzen. Auch Einzelschulen sollen die Möglichkeit bekommen, die Testaufgaben zu verwenden, um sich konkret an den Bildungsstandards zu messen.

Aufgabenentwicklung

Die Aufgaben werden zunächst von Lehrpersonen in unterschiedlichen Arbeitsgruppen entworfen, diskutiert und in konkreten Unterrichtssituationen erprobt. Danach folgt eine formale Prüfung der Aufgaben durch eine Aufgabenbewertungsgruppe, die sich aus Fachdidaktikern und Bildungsforschern zusammensetzt. Dort werden die Aufgaben erneut diskutiert, bewertet und kommentiert. Nach einer darauf folgenden inhaltlichen Endredaktion werden die Aufgaben in die Aufgabendatenbank des IQB eingegeben. Das IQB führt dann, um die testtheoretischen Eigenschaften der Items zu prüfen, Vorstudien durch und bereitet den Feldtest vor. Die Durchführung des Feldtests und die Dateneingabe wird dabei gänzlich an ein externes Datenverarbeitungsinstitut vergeben, während die Auswertung der Daten wieder zusammen mit dem IQB erfolgt. Aus der Auswertung und Skalierung der Daten ergeben sich Aufgabenparameter, die für jede Aufgabe innerhalb eines Kompetenzbereiches die Schwierigkeit definieren und angeben, wie gut sie für die Testung der jeweiligen Kompetenz geeignet ist. Zusätzlich dazu geben die damit in Beziehung gesetzten Personenfähigkeiten Aufschluss darüber, wie viele Schülerinnen und Schüler Aufgaben einer bestimmten Schwierigkeitsstufe bewältigen können bzw. ob sie bestimmte Standards erreichen. Im letzten Schritt werden die überprüften und normierten Aufgaben in einer Testdatenbank gesammelt, auf die wiederum auch die Institute der Bundesländer Zugriff haben.

Parallel zu diesem Testentwicklungsprozess werden sowohl in der Entwicklungs- als auch in der Auswertungsphase zusätzliche Aufgaben zur Implementierung im Unterricht entwickelt und zurückgelegt.

Normierung

Im Grundschulbereich findet die Pilotierung der jeweils ca. 900 Items für die Fächer Deutsch und Mathematik im Rahmen der IEA-PIRLS-Studie 2006 statt. Die Hauptstudie zur Normierung der Items für die 3. und 4. Klasse wird im Frühjahr 2007 an die TIMSS-Grundschuluntersuchung angebunden. In der Sekundarstufe ist die Normierung der Items für Mathematik bis 2007 abgeschlossen, während in den Bereichen Deutsch und der 1. Fremdsprache in diesem Jahr der Aufgabenentwicklungsprozess beginnt. Die Normierung in diesen Fächern ist für 2008 avisiert, in den Naturwissenschaften wird erst im Jahr 2007 mit der Entwicklung der Aufgaben begonnen.

Verbindung zum Qualitätsmanagement

Die primäre Aufgabe des IQB ist also in erster Linie eine Normierung und Weiterentwicklung der nationalen Bildungsstandards, sowie eine Entwicklung von Aufgabensammlungen zur Unterstützung eines an Standards orientieren Unterrichts. Erst der Einsatz der Aufgaben im Unterricht oder im Rahmen von Vergleichsarbeiten bzw. schulinternen Leistungstests stellt eine grössere Nähe zur Einzelschule her. Erst damit kann eine Verbindung zu anderen externen oder internen Evaluations- und Qualitätsmanagementprozessen hergestellt werden. Die Herstellung dieser Verbindungen liegt jedoch nicht im Auftrag des IQB, sondern soll durch die föderalen, regionalen und lokalen Institutionen geschehen. Das IQB arbeitet deshalb eng mit den für externe Evaluation zuständigen Qualitätsagenturen der Bundesländer, und mit den Landesinstituten für Schule in den Bundesländern zusammen. Diese wiederum halten engen Kontakt zu den Einzelschulen und unterstützen durch Evaluationen und Beratung den Schulentwicklungsprozess.

3.1.2 VERgleichsArbeiten (VERA)

Im Mai 2002 beschliesst die Kultusministerkonferenz (KMK) „gemeinsame Standards für die Schulbildung“, deren Einhaltung in Verantwortung der Länder durch Orientierungs- und Vergleichsarbeiten zu überprüfen ist. Vergleichsarbeiten sind schriftliche Arbeiten, die durch den Einsatz derselben Aufgabensammlung an unterschiedlichen Standorten einen Vergleich der Schülerleistungen über die Klassen- bzw. Einzelschulebene hinaus ermöglichen. Im Gegensatz zu den in der Zentralschweiz durchgeführten Orientierungsarbeiten werden bei den Vergleichsarbeiten die Ergebnisse im Vergleich zu landesweiten Resultaten aus einer Normierungsstichprobe zurückgemeldet. Aufgrund der Ergebnisse der Vergleichsarbeiten ist also eine breitere Einordnung in den Kontext einer landesweiten Leistungserhebung und der Vergleich mit sozialstrukturell ähnlichen Klassen möglich.

Aufgrund des Beschlusses der KMK werden in Rheinland-Pfalz seit 2003 in allen 4. Klassen der Grundschulen Vergleichsarbeiten eingeführt. Das Projekt zu Vergleichsarbeiten (VERA) wird von der Universität Koblenz-Landau wissenschaftlich geleitet und ist zunächst auf 5 Jahre ausgerichtet. Inzwischen ist VERA auf weitere sechs Bundesländer ausgeweitet worden und es sind bereits auch über den 5-Jahreszeitraum hinaus Erhebungen geplant.

Zielstellung

VERA verfolgt neben der Standortbestimmung weitere wichtige Ziele zur Qualitätssteigerung in Schulen. Zur Entwicklung des Unterrichts liefern die Vergleichsarbeiten fachliche, fachdidaktische und pädagogisch-psychologische Impulse, um schulinterne Aktivitäten und Kooperationen zu fördern. Die Ergebnisse der Vergleichsarbeiten liefern darüber hinaus eine ergänzende Infor-

mation zur Beratung der Eltern. Ein weiteres Ziel von VERA ist die Förderung der Diagnosegenauigkeit der Lehrpersonen. Dafür werden die Lehrereinschätzungen zu den Lösungshäufigkeiten der Testaufgaben mit den tatsächlichen Lösungshäufigkeiten in den jeweiligen Klassen in Beziehung gesetzt und zurückgemeldet. Durch die Teilnahme an den Vergleichsarbeiten soll zusätzlich eine intensivere Auseinandersetzung mit den neuen Rahmenplänen in Rheinland-Pfalz und in Zukunft mit den Bildungsstandards erfolgen, womit die Hoffnung auf eine schnellere und einfachere Umsetzung des standardbasierten Steuerungsmodells im Unterrichtsalltag verbunden ist.

Aufgabenentwicklung

Auf der Basis des Rahmenplans für die Grundschule werden im Rahmen von VERA für die Fächer Mathematik und Deutsch Aufgabensammlungen entwickelt. Dies geschieht durch erfahrene Lehrpersonen, die von Fachdidaktikern und Curriculumexperten der Landesinstitute beraten und unterstützt werden. Durch die Verwendung eines Kompetenzrasters, in dem die Inhaltsbereiche und Tätigkeitsanforderungen klassifiziert sind, wird sichergestellt, dass alle Bereiche des Rahmenplans in hinreichender Breite abgedeckt sind. Die Aufgaben werden in einer zentralen Datenbank verwaltet. Bevor sie normiert werden, werden sie in einer Pilotierungsstudie in ca. 40 Klassen getestet. Dadurch werden teststatistische Vorinformationen über die Aufgaben und Informationen über typische Schülerfehler erlangt, so dass ggf. noch eine Optimierung der Testitems vorgenommen werden kann. Die zentral vorgegebenen Aufgaben werden nach Abschluss des Testdurchgangs im Internet veröffentlicht (www.uni-landau.de/vera) und müssen deshalb jeweils neu entwickelt werden.

Normierung

In einer Normierungsstudie werden die Aufgaben dann einer grossen sozial repräsentativen Gruppe von Schülern aus einer Zufallsstichprobe vorgelegt, um den Massstab für die Vergleichsarbeiten zu liefern. Zusätzlich zum Test werden einige Hintergrunddaten wie soziales Umfeld der Schule und Alter, Geschlecht und Herkunftssprache der Schüler erhoben. Die Lehrpersonen werden zusätzlich zum Testverfahren befragt, um daraus evtl. Verbesserungen ableiten zu können. Die Antworten der Schüler und Lehrpersonen werden in der Universität Koblenz-Landau zentral ausgewertet. Die Normierung für den Bereich Mathematik wird beginnend im Jahr 2002 in einem Zweijahresrythmus wiederholt. Für Deutsch startet die Normierung ein Jahr später und wiederholt sich ebenso im Zweijahresrythmus. Die Ergebnisse und auch typische Schülerfehler oder Fehlermuster werden in einer Datenbank gespeichert, die den Schulen für die Durchführung und Auswertung der regulären Vergleichsarbeiten zur Verfügung steht.

Durchführung der Datenerhebung

Mit der Auswertung der Normierungsstudien ist also der Massstab für die Vergleichsarbeiten gegeben. Die Vergleichsarbeiten im Rahmen des Projekts VERA in Rheinland-Pfalz beginnen für den Bereich Mathematik im Jahr 2003, für den Bereich Deutsch im Jahr 2004 und wiederholen sich jährlich. Seit 2005 werden in einem der beiden Fächer im jährlichen Wechsel nur einzelne Teilbereiche erhoben, während das jeweils andere Fach vollständig abgedeckt wird. Bis 2006 nehmen alle 4. Klassen an den Vergleichsarbeiten teil, danach werden sie auf das Ende der 3. Klassenstufe vorverlegt. Für die Durchführung und schulinterne Auswertung sind die Schulen selbst verantwortlich, wobei sie durch detaillierte Handreichungen und Instruktionen unterstützt werden. Eine Hälfte der Aufgaben in den Vergleichsarbeiten wurde bis 2005 jeweils zentral vorgegeben, die andere Hälfte konnte von den beteiligten Kollegien bzw. Fachgruppen selbst aus dem Aufgabenpool gewählt werden, um den Schwerpunktsetzungen der einzelnen Schulen besser gerecht werden zu können. Durch diese Wahlmöglichkeit in Verbindung mit der verwendeten Auswertungsmethode (klassische Testtheorie) war ein direkter Vergleich von Schule zu Schule oder ein Ranking nicht möglich. Innerhalb der Schule konnte hingegen ein Vergleich zwischen den Leistungen einzelner Parallelklassen stattfinden und sollte pädagogische Diskussionen anregen. Ab dem Jahr 2006 werden jedoch in beiden Fächern alle Aufgaben zentral vorgegeben.

Rückmeldungen und Verarbeitung der Ergebnisse

Die von den Lehrpersonen vorgenommenen Auswertungen der Vergleichsarbeiten werden per Internet an die Universität Landau übermittelt. Den Schulen werden dann für ihre internen Auswertungsprozesse schul- und klassenspezifische Rückmeldungen zur Verfügung gestellt (siehe Anhang D.2.1). Im Rahmen eines auf ergebnisorientierte Schul- und Unterrichtsentwicklung gerichteten Projektes wie VERA stellen diese Rückmeldungen ein zentrales Element dar. Die Rückmeldungen beeinhalten unterschiedliche grafisch aufbereitete Ergebniszusammenfassungen auf Individual- Klassen- und Schulebene. Dazu gehören Leistungsverteilungen, Darstellungen von Fehlermustern und Lösungshäufigkeiten für einzelne Aufgaben sowie Rückmeldungen über die Diagnosegenauigkeit der Lehrperson.

Verbindung zum Qualitätsmanagement

Die Vergleiche mit den Ergebnissen aus der Normierungsstudie stellt die Einzelschule in einen breiteren Kontext und bietet einen übergreifenden Massstab im fairen Vergleich zu empirisch bestimmten Kontextgruppen (Schulen mit ähnlichen soziodemographischen Bedingungen werden verglichen). Aufgrund dieser Informationen soll, so die Hoffnung der Projektleitung, der

Entwicklungsprozess der Schule angeregt werden, so dass ein breiter Austausch über die Ergebnisse und mögliche Optimierungsmassnahmen aufgenommen wird, der letztendlich in einer tatsächlichen Qualitätsentwicklung für Schule und Unterricht mündet. Um diesen Prozess zu unterstützen, werden im Rahmen von VERA die beteiligten Lehrkräfte in den Gesamtprozess der Leistungsmessung (Aufgabenauswahl, Durchführung, Auswertung der Ergebnisse) einbezogen. Dennoch kann nicht automatisch davon ausgegangen werden, dass die Rückmeldung von Leistungsinformationen auch tatsächlich zu einem Bestreben führt, die Leistung zu steigern (vgl. Kohler und Schrader 2004). Aus der Reflexion der Ergebnisse der Leistungsuntersuchungen können, wie auch aus dem Beispiel „Check 5“ deutlich wird, die unterschiedlichsten Aktionen abgeleitet werden. Es kommt aber darauf an, dass ein Schulentwicklungsprozess in Gang gesetzt wird, der über eine oberflächliche Auseinandersetzung mit den Rückmeldungsinhalten hinausgeht und zu Aktionen führt, die die Qualität von Schule und Unterricht nachhaltig verbessern. Eine Lehrerbefragung, die kürzlich zu VERA durchgeführt wurde besagt, dass viele der Massnahmen, die aus den Ergebnisrückmeldungen abgeleitet wurden, relativ eng auf die eigene Klasse bezogen bleiben und dass die Möglichkeit der innerschulischen Kooperation in diesem Kontext noch zu wenig genutzt wird (Koch et al. 2006). Die Lehrpersonen scheinen teilweise das Bedürfnis nach einer stärkeren Anleitung im Schulentwicklungsprozess und in der Verarbeitung der Leistungsmessungsergebnisse zu haben, die in dieser Form im Konzept von VERA jedoch nicht vorgesehen ist. Aus- und Weiterbildungsangebote und Angebote zur stärkeren Begleitung von Einzelschulen sollen in den nächsten Jahren ausgebaut werden (Peek 2004).

3.1.3 Lernstandserhebungen in Nordrhein-Westfalen

Die Lernstandserhebungen in Nordrhein-Westfalen schliessen zum einen Vergleichsarbeiten in Klasse 4 durch das Projekt VERA ein und zum anderen werden seit Ende 2004 flächendeckend Lernstandserhebungen in den neunten Klassen in den Fächern Deutsch, Englisch und Mathematik geschrieben. Dabei wird nicht die gesamte Breite des Faches getestet sondern jährlich wechselnd der Schwerpunkt auf besondere Teilbereiche gelegt. Im Unterschied zu VERA werden zusätzlich die 9. Jahrgangsstufen (seit neuestem die 8. Jahrgangsstufen) getestet. Eine Besonderheit, die sich daraus ergibt ist die Berücksichtigung der jeweiligen Schulformen im dreigliedrigen Bildungssystem. Ansonsten sind die beiden Modelle in weiten Teilen vergleichbar.

Zielstellung

Durch das Design wird bereits deutlich, dass es bei den Lernstandserhebungen nicht primär um eine Feststellung des gesamten Fähigkeitsspektrums geht, sondern um einen gezielten und vertieften Einblick in einzelne Kom-

petenzbereiche. Die Funktion der Rechenschaftslegung bzw. des Systemmonitoring kann deshalb nur beschränkt ausgefüllt werden, der Schwerpunkt liegt klar auf der Unterrichtsentwicklung und Förderung in einzelnen Klassen (Burkard 2006). Dieser Argumentation folgt auch die Entscheidung ab dem Schuljahr 2006/2007 die Erhebungen in den achten, statt in den neunten Klassen durchzuführen. So bleibt in Zukunft etwas mehr Zeit, die Schülerinnen und Schüler auf die in Nordrhein-Westfalen neu eingeführten teilzentralen Abschlussprüfungen vorzubereiten. Auch ab dem Schuljahr 2006/2007 sollen die zentralen Lernstandserhebungen dann in Klasse 8 eine Klassenarbeit substituieren und so zur Leistungsbewertung herangezogen werden.

Im wesentlichen werden zwei Ziele mit der Auswertung der Daten aus den Lernstandserhebungen verfolgt: Erstens erhalten die Lehrpersonen eine computergestütze detaillierte Rückmeldung aus den Lernstandserhebungen mit dem Ziel die Auseinandersetzung mit den eigenen Ergebnissen und einen fachlichen Diskurs innerhalb der Schulen zu fördern. Zweitens liefert darüber hinaus eine landesweite, zentral ausgewertete Stichprobe Eckwerte über Lernstände in Teilkompetenzbereichen im Sinne eines eingeschränkten Systemmonitorings. Seit kurzem werden die 2% besten Schulen öffentlich ausgezeichnet.

Aufgabenentwicklung

Ähnlich wie im Projekt VERA sind die Aufgabensammlungen zunächst zentral durch Lehrpersonen mit der Unterstützung von Fachdidaktikern und anderen Experten des Ministeriums für Schule und Weiterbildung des Landes Nordrhein Westfalen entwickelt worden. Da die Aufgaben anschliessend veröffentlicht werden, ist es notwendig, jedes Jahr neue Aufgaben zu entwickeln. Ein direkter Vergleich der erfassten Kompetenzniveauwerte von Jahr zu Jahr ist dementsprechend nicht möglich.

Durchführung der Datenerhebung

Die Testdauer beträgt jeweils drei Stunden für Deutsch und Englisch und zwei Stunden für Mathematik und ist für alle Schülerinnen und Schüler der 9. Jahrgangsstufen (ca. 200.000) verpflichtend. Um die Tests den Lernvoraussetzungen der Schüler anzupassen, werden für unterschiedliche Schulformen zwei unterschiedliche Testformen eingesetzt. Diese enthalten zwar gleichermassen Aufgaben aller Schwierigkeitsstufen, jedoch legen sie in der einen Version einen Schwerpunkt auf grundlegendere und in der anderen Version auf anspruchsvollere Kompetenzen. Die Auswertung der Arbeiten erfolgt durch die Lehrpersonen, die hierfür ausführliche Auswertungsmanuale zur Verfügung gestellt bekommen. Die Manuale enthalten eine Darstellung der Auswertungskriterien, Kommentierungen der Aufgaben mit Bezügen zu den Lehrplänen und Hinweise auf besondere Anforderungsmerkmale. Zusätzlich

dazu wird eine Stichprobe von 250 Schulen angefordert und zweitkodiert um zu überprüfen, ob die Auswertungen in den Schulen in konsistenter Weise vorgenommen werden können. Eine vorbereitete Datenmaske, zu der die Lehrpersonen über Internet und persönlichem Passwort Zugang haben, ermöglicht die Dateneingabe und Übermittlung an das Ministerium für Schule und Weiterbildung. Die Datenauswertung erfolgt an der Universität Duisburg-Essen.

Fairer Vergleich

Wie auch im Rahmen des Projekts VERA werden in den Rückmeldungen die Leistungen der eigenen Klasse an den Leistungen anderer vergleichbarer Klassen in anderen Schulen gespiegelt. Es werden schulformspezifische Vergleichswerte geliefert. Um eine aussagekräftige Standortbestimmung vorzunehmen, werden den Schulen Referenzwerte zurückgemeldet, die nach ihren Rahmenbedingungen vergleichbar sind. Dazu werden im Vorfeld aufgrund von Angaben der Schulleiter zum Einzugsgebiet und zur Zusammensetzung der Schülerschaft so genannte Standorttypen gebildet, die unterschiedliche sozialökonomisch relevante Variablen wiederspiegeln. Im Rahmen dieses „fairen Vergleichs“ können die Leistungen aus Schulen verglichen werden, die unter ähnlichen Bedingungen arbeiten. Leistungsbeeinflussende Komponenten, die nicht mit Schule und Unterricht zusammenhängen, sondern eher mit einer selektiven Zusammensetzung der Schülerschaft, gehen auf diese Weise erheblich reduziert in die Auswertung und Interpretation der Ergebnisse ein.

Rückmeldungen und Verarbeitung der Ergebnisse

Die Schulen erhalten Handreichungen zum Umgang mit den Ergebnissen und Zugang zu einem Computerprogramm, mit dem die Ergebnisse einer Klasse im Vergleich zu Parallelklassen oder der Zentralstichprobe graphisch aufbereitet werden können. Es gibt mehrere Ebenen der inhaltlichen Auswertung. Zunächst gibt eine aufgabenbezogene graphisch aufbereitete Ergebnisrückmeldung einen Überblick über die Lösungshäufigkeiten bei einzelnen Aufgaben und zeigt den Lehrpersonen auf, an welchen Stellen spezifische Probleme aufgetreten sind, bzw. bei welchen Anforderungen die Leistung der Schüler den Erwartungen entspricht. Zu jeder Aufgabe werden Informationen bereitgestellt, die über Kompetenzerwartungen, mögliche Fehler und deren Ursachen Auskunft geben. Die Kompetenzniveaus der Schülerinnen und Schüler können weiterhin mit den inhaltlich definierten Anforderungsniveaus der Aufgaben in Beziehung gesetzt werden, so dass relativ präzise angegeben werden kann, welche Kompetenzen in einer Klasse vorhanden sind und in welcher Form sich der Lernbedarf für eine Verbesserung der Leistungen konkret darstellt.

Die Auseinandersetzung mit den Ergebnissen der Lernstandserhebungen

soll auf unterschiedlichen Ebenen erfolgen (zum idealtypischen Ablauf des Umgangs mit Ergebnissen siehe Anhang D.3.2). Auf der Ebene der Fachlehrpersonen soll eine individuelle Auswertung stattfinden, die Erkenntnisse über die Schwierigkeiten einzelner Schüler und einzelner Klassen zulässt. Auf der Ebene einer Fachgruppe oder Fachkonferenz soll eine Auseinandersetzung über die Durchführung und Rezeption der Erhebungen geführt werden und die Ergebnisse der Klassen als Teilergebnisse der Schule diskutiert werden. Ursachen für Defizite oder Erfolge im Hinblick auf ein schulinternes Konzept oder im Hinblick auf die gegebenen Rahmenbedingungen der Schule können dabei im Zentrum stehen. Schulsystemverantwortlichen wird die Möglichkeit gegeben, die Wirklichkeit an Problemschulen genauer zu analysieren oder Best-Practice Lösungen als Beispiele zu nutzen.

Auch die einzelnen Schülerinnen und Schüler bzw. ihre Eltern erhalten eine detaillierte Rückmeldung über die individuellen Lernstände (siehe Anhang D.3.1). Da dieses Instrument aber primär Aussagen über Lernstände in Klassen oder Schulen liefert und als Instrument für eine Individualdiagnostik nur beschränkt verwendbar ist, konnte diese Form der Rückmeldung noch nicht befriedigend umgesetzt werden. Schüler und Eltern hatten nach ersten Erkenntnissen erhebliche Schwierigkeiten, die Rückmeldungen zu lesen und zu verstehen, da das Material zu umfangreich war und die Ergebnisdarstellung zu kompliziert war.

Verbindung zum Qualitätsmanagement

Auch wenn eine Auseinandersetzung mit den Ergebnissen der Lernstandserhebungen im Rahmen eines innerschulischen Entwicklungsprozesses durchaus erwünscht ist, wird sie im Konzept dieses Projektes nicht explizit mitgedacht. Zwar gibt es neben der Ergebnisrückmeldung auch fachliche Erläuterungen zu den einzelnen Aufgaben und Hinweise zur Gestaltung einer Feedbackrunde mit Schülerinnen und Schülern, doch wie auch bei VERA schliesst dies keinen Automatismus ein, der in Richtung Schulentwicklung führt. Die Anstrengungen die in einzelnen Schulen in diesem Sinne gemacht werden, lassen sich vielmehr auf die Akzeptanz und das Engagement auf Seiten der verantwortlichen Lehrpersonen zurückführen. Je mehr Vertrauen die Lehrpersonen in die Notwendigkeit und die selbstverantwortete Möglichkeit einer Veränderung haben und je mehr sie sich für eine kritische Sichtweise über ihre eigene Schule und ihren eigenen Unterricht öffnen, desto eher kann ein Projekt wie dieses die erwünschte Wirkung erzielen. Derzeit gibt es nur wenige Rezeptionsstudien, die darüber berichten, wie mit den Ergebnissen von Leistungsmessungen in den Kollegien umgegangen wird. Vor allem aber stehen derzeit noch konkrete Planungen dazu aus, wie die Ergebnisse von Leistungsmessungen konkret und unabhängig vom Engagement der einzelnen Lehrpersonen in den Schulentwicklungsprozess einfliessen können. Seit kurzem führen Schulinspektoren in Nordrhein-Westfalen externe Eva-

luationen durch, und berücksichtigen dabei auch diese Frage. Zum jetzigen Zeitpunkt liegen allerdings noch keine Erfahrungsberichte zu einer Verbindung der Daten aus externen Schulevaluationen und Schülerleistungsdaten aus den Lernstandserhebungen vor.

3.1.4 Leistungsmessung und Inspektion in Hamburg

Die Behörde für Jugend und Sport in Hamburg richtet derzeit ein Institut für Bildungsmonitoring ein, in dem versucht werden soll, Schulinspektion und Leistungsmessung näher zusammenzubringen. Das Institut wird zum einen den Bereich Schulinspektion und zum anderen den Bereich Monitoring einschliessen. Der Bereich Monitoring ist wiederum in die Unterbereiche „Assessments und Lernstandserhebungen", „Qualitätskonzepte und Bildungsberichterstattung" sowie „Zentrale Prüfungen" unterteilt. Ähnlich wie in England soll also auch zukünftig (ab 2007) in Hamburg Leistungsmessung, Berichterstattung und Inspektion zentral unter einem Dach gesteuert werden, während die daraus resultierende Schulentwicklung den Einzelschulen und der Schulaufsicht obliegt.

Die Schulaufsicht ist deshalb auch explizit nicht in den Monitoring- oder Inspektionsprozess eingebunden. Sie erhält den Inspektionsbericht zeitgleich mit der Schule und berät die Einzelschulen hinsichtlich der Interpretationen der Ergebnisse und der Entwicklung von Massnahmen. Unterstützung erhalten die Einzelschulen bzw. die Lehrpersonen auch durch das Landesinstitut, das Qualifikationsseminare zum Umgang mit den Daten anbieten wird.

Schulinterne Qualitätsdaten, die für die Inspektion sowie für die Schulentwicklung relevant sind, sollen ähnlich wie in England auf einer Internetplattform „LUSD" für Schulen und Inspektion zur Verfügung gestellt werden. „LUSD" wird auch die Daten aus den zentralen Vergleichsarbeiten beinhalten, die, so die Planungen, zur Beurteilung der Schülerleistungen im Rahmen der Inspektion genutzt werden sollen. Weitere Daten, auf die sich die Inspektion abstützen wird, sind die früheren Inspektionsberichte sowie Daten aus Befragungen, Beobachtungen und Interviews. Ob auch interne Evaluationsdaten für die Inspektion nutzbar gemacht werden sollen, ist noch nicht geklärt. Eine schriftliche Befragung der Eltern, Schülerinnen und Schüler sowie der Lehrpersonen etwa 6 Wochen vor dem Inspektionsbesuch dient einer Vorbereitung des Teams auf die Gegebenheiten in der Schule.[4]

Zur Verbindung der Schülerleistungsdaten mit den Schulqualitätsdaten ist geplant, die Ergebnisse der zentralen Vergleichsarbeiten und der Abschlussprüfungen sowie die Quoten der Schulabbrechenden und Wiederholenden in der Inspektion zu berücksichtigen. Die zentralen Vergleichsarbeiten werden in den Klassenstufen 2, 4, 6 und 8 in Deutsch, Mathematik und

[4] Die Fragebögen sowie der Beobachtungsbogen für die Unterrichtshospitationen können unter www.schulinspektion.hamburg.de im Downloadbereich eingesehen werden.

der ersten Fremdsprache geschrieben. Für die 8. Klassenstufe gibt es zusätzlich eine Vergleichsarbeit in der zweiten Fremdsprache. Die Aufgaben sind lehrplanvalide und werden jedes Jahr neu von Testentwicklungsexperten zusammengestellt. Einmal im Jahr bearbeiten Hamburger Schülerinnen und Schüler der entsprechenden Jahrgangsstufen die Arbeiten unter Aufsicht ihrer Lehrpersonen im Klassenverband. Die Lehrpersonen sollen zukünftig auch die Codierungen und die Dateneingabe vornehmen (im Jahr 2007 werden die Daten extern verarbeitet). Die statistische Datenauswertung wird voraussichtlich durch das neue Institut für Bildungsmonitoring in der Abteilung für zentrale Prüfungen vorgenommen, die eine Rückmeldung an die Lehrpersonen und die Inspektion abgibt. Für die Auswertung ist geplant, Prozentwerte korrekter Aufgaben zu berechnen und Skalierungen nach der Item-Response-Theory[5] vorzunehmen. Die Vergleichsarbeiten können in der Regel auch wie Klassenarbeiten benotet werden.

Basierend auf den Ergebnissen aus den Vergleichsarbeiten, den Abschlussprüfungen und den Abbrecher- sowie Wiederholerquoten erarbeitet die Inspektion dann eine Einschätzung zur schülerleistungsbezogenen Schulqualität. Wie die Auswertungen in diesem Zusammenhang konkret durchgeführt werden und welchen Stellenwert dieser Aspekt im Inspektionsbericht haben wird, ist jedoch derzeit noch nicht klar.

3.1.5 Zusammenfassung und Beurteilung

Mit der Einsetzung des Instituts für Qualitätsentwicklung im Bildungswesen (IQB) ist in Deutschland ein grosser Schritt in die Richtung der Vereinheitlichung von Leistungsstandards und Leistungsmessungen getätigt worden. Die dort entwickelten Aufgaben orientieren sich an den nationalen Bildungsstandards und sind vielfältig für unterschiedliche nationale, bundesländerweite oder regionale Tests einsetzbar. Zusätzlich dazu liefern die entwickelten Trainingsaufgaben den Lehrpersonen die Möglichkeit, die Standards besser im Unterricht zu implementieren. Mit Projekten wie VERA oder den Lernstandserhebungen in 9. Klassen in Nordrhein-Westfalen können die Leistungen der Schülerinnen und Schüler gezielt auf einem einheitlichen Massstab verglichen werden. Durch die Bereitstellung der Testaufgaben wird den Lehrpersonen die Möglichkeit eröffnet, die Ergebnisse gezielt und detailliert für ihre eigene Klasse und die eigene Schule zu analysieren. Dennoch ist die Beschäftigung mit den Ergebnissen aus den Leistungstests absolut abhängig vom Engagement der einzelnen Lehrpersonen. Überlegungen zur Einbindung spezifischer schulischer Aktionen in die Konzepte der Leistungsmessung, wie beispielsweise im Rahmen von Check 5, sind bisher nicht vorgesehen. Die in Deutschland eingesetzten Leistungsmessungen enthalten gleichwohl ein fachbezogenes diagnostisches Potential, sie geben Hinweise über die Stärken

[5] siehe Erklärung im Anhang A

und Schwächen individueller Schülerleistungen oder Lernstände von Klassen und Schulen. Daraus können Rückschlüsse über den Unterricht und evtl. künftige Akzentsetzungen geschlossen werden und es stehen unterschiedliche Materialien bereit, die den Lehrpersonen beim Umgang mit den Ergebnissen in unterschiedlichen Teilleistungsbereichen helfen. Neuere Überlegungen schliessen allerdings auch die Zusammenführung von Ergebnissen aus Lernstandserhebungen und diagnostischen Prozessbeobachtungen (vgl. Dobbelstein und Peek 2005) oder die Verknüpfung von Schulinspektion und Leistungsmessung ein.

3.2 Beispiele aus England

In England hat sich in den 1980er und 1990er Jahren eine spürbare Entwicklung in Richtung auf einen verstärkten Wettbewerb zwischen Einzelschulen bemerkbar gemacht. Die 1988 von der Thatcher-Regierung eingeleitete Bildungsreform bringt eine Reihe von Veränderungen mit sich, die sich auf die Vorstellung stützen, dass Qualität von Schule und ökonomischer Erfolg für das Land durch einen Wettbewerb zwischen den Schulen vorangebracht werden können (van Ackeren 2003). Obwohl im Zuge der Reform 1988 ein nationales Curriculum eingeführt wird, das die Lernziele für Schülerinnen und Schüler ab dem dritten Lebensjahr einheitlich vorgibt und für die Erreichung dieser Lernziele einheitliche Kriterien definiert, wird damit nicht automatisch eine Einheitlichkeit der Schulen, sondern ebenso auch eine höhere Konkurrenz zwischen ihnen erreicht. Der Grund dafür ist, dass die Kriterien verstärkt dafür eingesetzt werden, das Erreichen der vorgegebenen Standards durch Tests und durch Schulinspektionskontrollen zu überprüfen. Anders als in in der Schweiz werden in England die auf Schulebene kumulierten Testergebnisse öffentlich in der Tagespresse dargestellt und können über das Internet abgerufen werden. Auch die Berichte der Inspektoren über die einzelnen Schulen sind öffentlich verfügbar. Zusammen mit einer höheren finanziellen Autonomie der Einzelschulen bei gleichzeitiger Pro-Kopf-Ressourcenzuteilung nach Anzahl der Schülerinnen und Schüler und einer freien Schulwahl der Eltern führt dies zu einem verstärkten Druck auf die Schulen, die nun noch mehr bemüht sind, die Leistungen, sei es durch Förderung und Forderung, sei es durch Selektion zu steigern.

Das nationale Curriculum sieht für Kinder im Alter von drei bis fünf Jahren in der „Foundation Stage“ ein spielorientiertes Lernen vor. Für die Altersstufen von 5 bis 16 Jahren spielen Inhalte und obligatorisch abzudeckende Studienprogramme eine Rolle. Dazu werden sogenannte „Attainment targets“ festgelegt, Standards, die am Ende unterschiedlicher Bildungsabschnitte erreicht werden sollen. Für Kinder und Jugendliche ab dem Alter von fünf Jahren ist das nationale Curriculum in Form von vier Lernblöcken formuliert, die sich an dem Alter und der Entwicklungsstufe der Schülerinnen und Schüler orientieren. Diese vier Lernblöcke werden „key stages“ genannt.

National Curriculum Assessments

Um sicherzustellen, ob die im nationalen Curriculum festgeschriebenen Standards in den einzelnen Schulen erreicht werden und um zu überprüfen, inwiefern individuelle Lernfortschritte zu verzeichnen sind, stehen am Ende jedes dieser Lernblöcke formale Evaluationen der Leistungen. Die Ergebnisse dieser „national curriculum assessments" dienen zum einen der Rechenschaftslegung der Schulen gegenüber der Regierung und den Eltern. Andererseits liefern sie Informationen an die Schulen und die lokalen Schulaufsichtsbehörden „local authorities" (LA). Für Schülerinnen und Schüler selbst ergibt sich keine Rechenschaftspflicht. Eltern und Lehrern wird mit den Ergebnissen eine Hilfestellung zur Identifizierung der Stärken und Schwächen individueller Schüler an die Hand gegeben. Es wird damit überprüft, ob die „national attainment targets" erreicht wurden und die Lernentwicklung in den Einzelschulen erwartungsgemäss verlaufen ist. Schulen, die mit grösseren Schwierigkeiten konfrontiert sind, wird gegebenenfalls Unterstützung angeboten. Die „national curriculum assessments" bestehen zum einen aus den „national curriculum tests" und zum anderen aus einer Einschätzung der Lehrpersonen. Diese Einschätzung soll, basierend auf den Beobachtungen der Lehrperson, ein weiteres Bild über die Arbeit eines individuellen Schülers im Laufe einer „key stage" geben. Die Einschätzung orientiert sich an den Beschreibungen im Nationalen Curriculum zu Leistungsniveaustufen. Beispiele zur Beurteilung finden sich ebenso wie die Niveaubeschreibungen auf einer Internetseite der „Qualification and Curriculum Authority" (www.ncaction.org.uk/). Schriftliche, mündliche und praktische Leistungen, Schul- und Hausarbeiten während der Periode der „key stage" werden dafür zugrunde gelegt. Die Einschätzungen der Lehrpersonen müssen jährlich zu einem fest vorgegebenen Zeitpunkt an die NAA geschickt werden. Zusammen mit den Ergebnissen der Tests werden sie im Internet für die Schulen, die LAs und die Schulinspektion zur Verfügung gestellt.

Seit Mitte der 1990er Jahre werden in England regelmässig „national curriculum tests" durchgeführt. Sie sind für alle staatlichen Schulen verpflichtend und finden zu den vier unterschiedlichen „key stages" in der obligatorischen Bildungslaufbahn der Schülerinnen und Schüler statt (nach der „key stage 4" folgt die standardisierte Abschlussprüfung, die Teil der „national curriculum tests" ist). In England bearbeiten jährlich 1,25 Millionen Schülerinnen und Schüler in 26.000 Schulen 7 Millionen Tests (zu den Assessments in den einzelnen „key stages" siehe Anhang E.1).

Insgesamt sind 8 Niveaustufen festgelegt, an denen die Leistung der Schülerinnen und Schüler festgemacht werden kann. In der Regel soll es innerhalb von zwei Jahren einen Leistungszuwachs von einer Niveaustufe geben. Zur Orientierung der Eltern, der Schülerinnen und Schüler und der Allgemeinheit werden die Resultate der Schulen aus den Tests öffentlich bekannt gegeben.

Veröffentlichung der Testergebnisse

Die auf Schulebene aggregierten Ergebnisse der „national curriculum tests“ auf den „key stages“ 2 bis 4 werden vom Bildungsministerium, dem Department for Education and Skills (DfES) auf ihrer Webseite in so genannten „Attainment and Achievement Tables“ veröffentlicht (für ein Beispiel siehe Anhang E.2). Dabei werden die Schulen nach Distrikten und in alphabetischer Anordnung gelistet. Jedoch wird diese Information regelmässig von den öffentlichen Medien genutzt, um daraus Rankingtabellen herzustellen. Zusätzlich zu den „Attainment and Achievement Tables“, die in erster Linie der Information einer breiten Öffentlichkeit dienen, gibt das DfES speziell für Schulen und deren Partner im Schulentwicklungsprozess eine Software zur vertiefenden Bearbeitung der Testergebnisse sowie der Lehrpersoneneinschätzungen, den so genannten „Pupil Achievement Tracker“, heraus. Diese Software wird im Jahr 2007 durch die weiterentwickelte, interaktive Internetplattform „RAISEonline“ ersetzt werden (für Beispielscreenshots aus Anwendungen von RAISEonline siehe Anhang E.3).

Administrative Verantwortung

Das nationale Curriculum und die „national curriculum tests“ stehen unter der Verantwortung der „Qualifications and Curriculum Authority“ (QCA), einer öffentlichen, vom Ministerium unabhängigen Organisation. Die QCA erhält den Auftrag für diese Aufgabe durch das DfES. Die der QCA untergeordnete Abteilung „national assessment agency“ (NAA) wurde im Jahre 2004 vom DfES eingesetzt und ist für die Überprüfung und Aktualisierung der Examina und Tests verantwortlich. Dazu arbeitet sie direkt mit Schulen und Testentwicklungsinstituten sowie mit den „local authorities“ zusammen. Sie vergibt Aufträge zur Entwicklung der Tests an unabhängige Forschungseinrichtungen oder Universitäten und ist für die Organisation der Kodierung der Tests durch speziell dafür eingestellte Lehrpersonen (so genannte „marker“) verantwortlich. Da die Tests vielfach so konzipiert sind, dass sie eine Reihe von offenen Antwortformaten beinhalten, wird dies zu einer hoch komplexen, logistisch anspruchsvollen Aufgabe. An der Kodierung der Hauptuntersuchungen der „national curriculum tests“ für die „key stages“ 2 und 3 sind ca. 12.500 „marker“ beteiligt. Die „marker“ werden für ihre Arbeit finanziell entschädigt, zusätzlich wird auf der Internetseite der NAA der hohe Lerneffekt in Bezug auf Lehr- und Diagnosekompetenzen herausgestellt.

Weitere Leistungsmessungen und Inspektionen

Neben diesen regelmässigen obligatorischen und schullaufbahnbegleitenden Tests werden in England noch eine Reihe weiterer Schulleistungsstudien durchgeführt. Die National Foundation for Educational Research (NFER)

spielt bei der Entwicklung und Durchführung solcher Schulleistungsstudien eine herausragende Rolle. Auch die Aufträge für die „national curriculum tests“ werden zu einem erheblichen Teil an diese Nicht-Regierungs-Organisation vergeben. Als weitere wichtige „test development agencies“, die an der Entwicklung von Assessment-Aufgaben beteiligt sind, sind die Universitäten von Cambridge, Leeds und Liverpool zu nennen.

Neben den „national curriculum tests“ werden zur Überprüfung der Qualität von Schulen regelmässig Inspektionen durchgeführt. Verantwortlich für diese Inspektionen ist das „Office for Standards in Education“ (Ofsted). Die Berichte des Ofsted sind ebenso wie die Testresultate auf Schulebene öffentlich zugänglich.

Verarbeitung der Ergebnisse

Auch wenn die Rechenschaftslegung der Schulen und die Information der Eltern sowie das Aufzeigen nationaler Trends als Funktionen offensichtlich im Vordergrund stehen, gibt es Ansätze dazu, die Testresultate und Inspektionsberichte zur Qualitätsentwicklung der Schulen zu nutzen. Dazu werden vom DfES und von Ofsted Informationen zusammengestellt, auf deren Grundlage faire Vergleiche zwischen Schulen mit ähnlichen Bedingungen angestellt werden können. Zukünftig werden diese Informationen auf der Internet-Plattform RAISEonline zusammengeführt (siehe Anhang E.3). Die Schulen sind dazu angehalten, inhaltliche Schlüsse daraus abzuleiten und Massnahmen zur Qualitätssteigerung zu entwickeln.

Das „target setting“ ist eine weitere Massnahme, die die Verantwortlichen der Einzelschulen zur intensiven Auseinandersetzung mit den Ergebnissen von Tests und Inspektion anhalten soll. Im Rahmen dieser Massnahme sollen Ziele für einzelne individuelle Schülerinnen und Schüler sowie für die gesamte Schule definiert werden. Der Entwicklungsprozess der Schulen auf der Basis der Ergebnisse der Tests und Inspektionen wird durch die Local Authorities (LA) und insbesondere deren abgeordnete „School improvement partners“ (SIP) unterstützt. Dafür findet ein jährliches Treffen der Verantwortlichen der Einzelschulen mit den SIP statt, bei dem die Entwicklung der Schule vor dem Hintergrund ihrer Zielvereinbarung beurteilt wird. Die SIP unterstützen und fördern den Entwicklungsprozess der Schulen in der Rolle „kritischer Freunde“ und helfen dabei, Selbstevaluationen durchzuführen und neue Entwicklungspläne aufzustellen. Die Selbstevaluationen der Schulen liegen ihrerseits wiederum im Fokus der Inspektion durch Ofsted.

Im folgenden wird näher auf die in England eingesetzten „national curriculum tests“ sowie auf das Vorgehen des Inspektorats und die Verbindung der beiden Ansätze eingegangen.

3.2.1 National Curriculum Tests

Die „national curriculum tests“ dienen dazu, Informationen über die Leistungen der einzelnen Schülerinnen und Schüler in den unterschiedlichen „key stages“ zu erheben, und diese für die Eltern, die Schüler selbst, die Schulen und die übergeordneten bildungspolitischen Institutionen zur Verfügung zu stellen. Es sollen dadurch auch die Vergleichbarkeit zwischen den Schulen ermöglicht und nationale Trends aufgezeigt werden.

Aufgaben- und Testentwicklung

Die Tests werden in einem ca. ein bis eineinhalbjährigen Prozess von den „test development agencies“ entwickelt, wobei zunächst die Testaufgaben in Anlehnung an das nationale Curriculum entworfen und anschliessend durch Curriculumexperten und Fachdidaktiker validiert werden. In einem ersten Pretest werten speziell trainierte erfahrende Lehrpersonen („marker“) anhand vorgegebener Lösungshilfen die Antworten der Schülerinnen und Schüler aus der Pilotierungsstichprobe aus. Mit den daraus gewonnenen Daten wird die Angemessenheit der Aufgaben für die entsprechenden Zielpopulationen (keystage 1 bis 3) überprüft und der Schwierigkeitsgrad der Aufgaben analysiert. Zusätzlich werden eine Reihe von Teststatistiken berechnet die Auskunft darüber geben, wie trennscharf die einzelnen Testaufgaben geeignet sind, die Kompetenzen der Schülerinnen und Schüler zu messen. Es werden zur Beurteilung der Aufgabenqualität ausserdem noch Einschätzungen der Lehrpersonen zum Testmaterial und zu den Leistungen der Schülerinnen und Schüler herangezogen. Etwa die Hälfte der Aufgaben wird daraufhin für den nächsten Analyseschritt ausgewählt. Auch die Lösungshilfen für die „marker“ werden entsprechend der Korrekturvorschläge aus dem ersten Einsatz überarbeitet. In einem zweiten Pretest wird dann mit einer relativ kleinen Stichprobe überprüft, ob der Test hinsichtlich der gemessenen Leistung mit anderen Tests aus den Vorjahren vergleichbar ist. Dazu wird ein so genannter Ankertest verwendet, der jedes Jahr in derselben Form verwendet wird. Anhand des Vergleichs der Ergebnisse dieses Ankertests mit dem neu entwickelten Instrument ist es möglich, die relative Schwierigkeit des neuen Tests abzugleichen. So kann eine Einheitlichkeit der Testschwierigkeit über die Zeit abgesichert werden. Weiterhin wird auch der zweite Pretest nochmals für eine weitere Optimierung der Lösungshilfen für die „marker“ genutzt. Mit der endgültigen Zusammenstellung der Items für den Haupttest und der Formulierung der endgültigen Lösungshilfen wird der Entwicklungsprozess abgeschlossen. Im Vorfeld der Hauptuntersuchung werden die ca. 12.500 beteiligten „marker“ noch einmal speziell im Hinblick auf die Bewertung der Haupttestversion trainiert. Ausserdem legen die Vertreter der NAA und der zuständigen „test development agencies“ die zugehörigen Leistungsniveaus fest. Sie bedienen sich dazu statistischer Verfahren anhand von Daten aus

dem zweiten Pretest und einer Einschätzung von besonders erfahrenen Lehrpersonen zu den Wahrscheinlichkeiten, mit denen Schülerinnen und Schüler unterschiedlicher Leistungsniveaus die einzelnen Aufgaben lösen können. Die daraus resultierende Festlegung der Leistungsniveaus wird im Anschluss des Haupttests noch einmal anhand einer ersten Auswahl erhobener Daten im Rahmen des „level confirmation exercise“ (LCE) überprüft. Daraus resultiert eine endgültige Festlegung der Leistungsniveaus anhand derer alle Tests bewertet werden.

Um die Tests jährlich weiter zu entwickeln und zu optimieren, werden nach der Durchführung noch einige Evaluationen zum Testverfahren durchgeführt. Dazu gehen Fragebögen an die Schulen und die „marker“ und zusätzlich werden teilweise telefonische Befragungen an den Schulen durchgeführt. Die Testunterlagen werden ebenso nachträglich noch einmal einer vertiefenden Evaluation hinsichtlich der Validität der verwendeten Texte, Graphiken und Fragestellungen unterworfen. Der abschliessende Bericht dieser Evaluation bildet ein Instrument zur Weiterentwicklung der Tests in den Folgejahren. Weiterentwicklung heisst in diesem Zusammenhang das Aufnehmen und Weiterverarbeiten von Besonderheiten im Prozess der Entwicklung. Die Aufgaben werden nicht mehrfach verwendet, es stehen jedes Jahr neu entwickelte Tests zur Verfügung.

Durchführung der Datenerhebung

Beim „key stage 1“ wird mit Rücksicht auf die jüngeren Kinder etwas mehr Flexibilität hinsichtlich des Testzeitpunkts gewährt als in den anderen Altersstufen. Auch die Bewertung erfolgt in dieser Altersstufe durch die die eigenen Lehrpersonen. Die Schülerinnen und Schüler auf den „key stages“ 2 und 3 werden jährlich im Mai an einem national einheitlichen Tag getestet. Dazu verteilen die Lehrpersonen die vorher in verschlossenen Umschlägen versandten Testmaterialien selbst in ihrer Klasse und lassen den Test nach den vorgegebenen schriftlichen Anweisungen durchführen. Durch die kurze Zeitspanne zwischen der Zusendung der Hefte und der Durchführung der Tests soll eine Manipulation weitgehend verhindert werden, auch wenn dies nicht in allen Fällen gelingt.

Die Schulleitungen sind dazu verpflichtet, die Testunterlagen mit Namen, Geburtsdatum und Geschlecht der Jugendlichen zusammen mit Angaben zur Schule innerhalb von zwei Tagen an die zuständigen „marker“ zu schicken. Diese bewerten die Tests und schicken die Unterlagen mit einer Kopie eines Benotungsbogens bis Anfang Juli zurück an die Schulen.

Rückmeldung und Verarbeitung der Ergebnisse

Die Bewertungen werden ebenfalls an die QCA verschickt, die weitere detaillierte Analysen vornimmt und einen zusammenfassenden Bericht (standards

report) mit Hinweisen auf Lehr- und Lernstrategien erstellt. Diese Form der Rückmeldung kann von den Schulen für ihre Schulentwicklungsarbeit genutzt werden. Auch wenn die „standards reports“ keine Rückschlüsse auf einzelne Schulen oder sogar Schüler zulassen, können sie den Lehrpersonen dabei helfen, die Ergebnisse besser zu verstehen und zu bewerten. Die „standards reports“ wenden sich im wesentlichen an Schulleitungen und Lehrpersonen, sind jedoch auch für die Öffentlichkeit zugänglich. Es werden des weiteren Informationsveranstaltungen zu den Resultaten der Tests angeboten.

In Ergänzung dazu werden jedes Jahr vom DfES die bereits genannten „Attainment and Achievement Tables“ herausgegeben, in denen nationale Ergebnisse aus den Tests für alle „key stages“ sowie für die Abschlussprüfungen mit Verweis auf die Einzelschulen dargestellt werden. Den Berechnungen der „Attainment and Achievement Tables“ liegen Informationen aus Hintergrund- und Kontextvariablen zugrunde, die faire Vergleiche zwischen Schulen mit ähnlichen Bedingungen zulassen. Zusätzlich sind „value added“ Informationen enthalten, die Aufschluss über die Differenzen zwischen den Altersstufen geben. Mit Hilfe eines interaktiven Programms, dem so genannten „Pupil Achievement Tracker“ (ab 2007 RAISEonline) können die Schulen auch Daten der eigenen Schule einfügen und eigene gezielte Berechnungen zum Vergleich mit den nationalen Testergebnissen durchführen.

Die Primar- und Sekundarschulen erhalten darüber hinaus von Ofsted einen „Performance and Assessment Report“ (PANDA) mit Hilfe dessen sich die Inspektoren auf ihren Schulbesuch vorbereiten. Ab 2007 wird der PANDA ebenfalls durch RAISEonline abgelöst. Des Weiteren sind zusätzliche Materialien von der Stiftung „Fischer Family Trust“ erhältlich, die durch die Bereitstellung von detaillierten Analysen die Selbstevaluation in den Schulen unterstützen möchte.

Das englische Modell der Schulqualitätsbeurteilung zeichnet sich durch eine sehr enge Zusammenarbeit zwischen den Systemen der externen Schulevaluation und der Leistungsmessung aus. So werden die Ergebnisse aus den „national curriculum assessments“ (Tests und Lehrpersoneneinschätzungen) an die nationale Schulaufsicht (Ofsted) gemeldet, so dass diese sich über den Leistungsstand der inspizierten Schulen informieren kann. Ebenso werden das Ministerium und die regionalen Bildungsbehörden („local authorities“ LA) informiert.

3.2.2 Inspektionen durch das Office for Standards in Education (Ofsted)

England hat ähnlich wie die Niederlande eine Tradition der Schulinspektion, die bis in das 19. Jahrhundert zurückreicht. Das erklärt zum Teil die öffentliche Bedeutung und die finanziellen Zuwendungen, die das „Office for Standards in Education“ (Ofsted) als zentrale Inspektoratsinstanz erfährt und die Akzeptanz die ihm durch die inspizierten Schulen entgegengebracht wird

(vgl. Grubb 1999). Seit 1993 ist das vom Ministerium unabhängige staatliche Ofsted für die Planung und Durchführung dieser externen Evaluationen verantwortlich. Regelmässig werden alle ca. 26.000 öffentlich oder teilöffentlich finanzierten Schulen in England inspiziert. 60 Million Pfund sind im Budget von Ofsted für die Schulinspektion vorgesehen.

Ziel der Inspektionen ist es, eine unabhängige externe Evaluation der Schulen bereitzustellen und damit die Selbstevaluation der Schulen zu ergänzen. Neben den Verantwortlichen in den Schulen werden vor allem die Eltern und das DfES über die Qualität und die erreichten Standards der Schulen informiert. Mit der Inspektion ist die Hoffnung verbunden, dass die Identifikation der Stärken und Schwächen einer Schule den Schulentwicklungsprozess anschieben.

Im Rahmen der Inspektion wird besonderer Wert auf die Qualität der Bildungsangebote gelegt und darauf, inwiefern sie den Bedürfnissen der Schülerinnen und Schüler entsprechen. Auch die Schülerleistungen spielen dabei eine Rolle. Als weiterer zentraler Punkt ist die Qualität der Schulleitung und des Managements zu nennen. Ausserdem werden Aspekte der geistigen, moralischen, sozialen und kulturellen Entwicklung sowie Massnahmen zur Steigerung des Wohlbefindens der Schülerinnen und Schülern in den Blick genommen.[6] Die Einzelheiten des Ablaufs der Inspektion sind im „Every-child-matters"-Rahmenprogramm zur Inspektion festgeschrieben. Mit diesem neuen Rahmenprogramm vom September 2005 werden die Inspektionen verschlankt. Während vorher ein sehr grosser Schwerpunkt auf das Unterrichtsgeschehen gelegt wurde und ca. 60 Prozent der Inspektionszeit, mit Unterrichtshospitationen ausgefüllt wurde, konzentrieren sich die Inspektoren seit neuestem mehr auf den gemessenen Leistungsoutput bzw. auf die Lernentwicklung, die über die Ergebnisberichte des DfES, für jeden Schüler und jede Schülerin am Ende jeder „key stage" zur Verfügung stehen. Eine wichtige Rolle bei der Inspektion spielt die Betrachtung der Selbstevaluation der Schule. In diesem Zusammenhang wird auch geprüft, wie die Schule auf die Ergebnisrückmeldungen aus den „national curriculum assessments" und den vorangegangenen Inspektionen reagiert hat.

Auf der Grundlage der Selbstevaluationen und der Leistungsdaten aus den „national curriculum assessments", die von Ofsted speziell für die Schulen und für die Inspektoren in den „PANDA-Reports" (in Zukunft auf der Plattform RAISEonline) zusammengestellt werden, bereiten sich die Inspektoren vor. Die Besonderheit der Inspektion durch Ofsted ist, dass die Daten aus der Leistungsmessung und der Inspektion sehr eng miteinander verknüpft und auch im Inspektionsbericht zueinander in Beziehung gesetzt werden (siehe auch Beispielbericht im Anhang E.4).

Jede Schule wird in einem Zeitraum von drei Jahren einmal inspiziert,

[6] Ein anonymisiertes Beispiel eines Inspektionsberichts findet sich im Anhang E.4. Darin ist auch eine Tabelle der Inspektionskriterien enthalten.

Schulen, die Schwächen aufweisen, sogar noch öfter. Dabei orientiert sich das Inspektionsteam aus bis zu 5 Inspektoren im Wesentlichen an den Vorgaben des Rahmenprogramms. In welche Bereiche jedoch die Schwerpunkte der Inspektion gelegt werden, hängt von der Besonderheit der Einzelschule und ihrer Leistung ab. In der Vergangenheit wurde die intensive Vorbereitung auf die Inspektionen als besonderer Stressfaktor für die Lehrpersonen und als Störfaktor bei der Beurteilung angesehen. Deshalb wird seit September 2005 der Zeitpunkt erst 2 bis 5 Tage im Voraus mitgeteilt und die Inspektion wurde zudem zeitlich verkürzt.

Die Ergebnisse der Inspektion werden der Schulleitung zunächst mündlich dargelegt. Innerhalb von drei Wochen folgt dann der ca. 6-seitige schriftliche Inspektionsbericht zusammen mit einem Erklärungsbrief an die Schülerinnen und Schüler. Der Bericht darf keine Urteile enthalten, die nicht in der mündlichen Fassung bereits genannt wären. Die Schule schickt innerhalb von fünf Tagen nach Erhalt des Berichts, eine Kopie an die Eltern. Ebenso wird der Bericht auf der Webseite von Ofsted unter Angabe des vollständigen Namens für jede Schule veröffentlicht. Den Eltern soll damit eine Entscheidungshilfe für Schulwahl oder -wechsel an die Hand gegeben werden. Auch die Schulaufsichtsbehörden, die „local authorities“, erhalten eine Kopie des Berichts.

Erfüllt eine Schule die erwarteten Kriterien nicht, so können unterschiedliche Massnahmen greifen. In der Regel wird die Schule in die Kategorie „special measures“ eingeordnet, was bedeutet, dass sie ihren Schülern keinen akzeptablen Standard anbieten kann und dass das Management der Schule nicht die Kompetenzen zeigt, die für eine Verbesserung der Situation notwendig wäre. Schulen in dieser Kategorie werden nach zwei Jahren ein weiteres Mal inspiziert. Einige Schulen bekommen auch ein „notice to improve“, da sie erhebliche Mängel bei den für die Inspektion relevanten Aspekten aufweisen. Diese Schulen werden nach einem Jahr nochmals inspiziert.

Um die Qualität der Inspektionen sicherzustellen, wird die Arbeit der Inspektoren regelmässig intern überprüft. Zusätzlich liefert eine Befragung der Verantwortlichen Personen in den Schulen Informationen darüber, in welchen Bereichen die Inspektion verbessert werden kann. Kürzlich ist vom NFER eine Evaluation der Schulinspektion bei Ofsted (Pilotstudie) durchgeführt worden (Atkinson et al.). Dazu wurde ein Fragebogen an 200 Schulen versandt und zusätzlich wurden 36 Schulleiter befragt, in deren Schulen im letzten Jahr eine Inspektion stattgefunden hat. Vor allem die Konzentration der Inspektion auf die Selbstevaluation wird laut dieser Studie sehr positiv wahrgenommen und als Antriebsmotor für Schulentwicklung gesehen. Das Gesamtfeedback, das die Schulen sowohl mündlich als auch in Form eines Reports erhalten, wird dieser Studie zufolge von der Mehrheit der beteiligten Schulen als fair und hilfreich angesehen. Dabei ist auffällig, dass der Grad in dem die Inspektion als hilfreich angesehen wird, sinkt, je kritischer der Report ausfällt. Teilweise wird moniert, dass die Inspektion zu datenorientiert

ist, was sicherlich auch damit zusammenhängt, dass die Leistungsdaten der Schülerinnen und Schüler bei der Planung der Inspektion eine grosse Rolle spielen. Die Wirkung der Inspektion auf die Schulentwicklung wird von den meisten Befragten wahrgenommen und mehr als drei Viertel geben an, dass bereits positive Effekte sichtbar sind.

3.2.3 Zusammenfassung und Beurteilung

Die Entwicklung der Tests und die Inspektion hat in England Wirkung gezeigt. Tatsächlich sind die Leistungen der Schülerinnen und Schüler seit der Einführung dieses umfassenden Monitoringsystems gestiegen. Dies kann an den Ergebnissen der jährlich neuen aber einheitlich geeichten Tests abgelesen werden. Die Erklärungen dazu sind sehr vielfältig. Es ist durchaus kein Konsens, dass diese Steigerung der Leistungen ausschliesslich aus einer verstärkten Schul- und Unterrichtsentwicklung in Folge der Tests und Inspektionen resultiert. Vorgebracht werden auch andere Erklärungen, wie das „Teaching to the test“, verordnete Krankheit für schlechte Schüler am Testtag, Hilfestellung der Lehrpersonen während des Tests oder schlicht eine Gewöhnung der Schüler an die Testsituation.

Wie die Entwicklung auch immer erklärt wird, es scheint die Verbindung zwischen der Inspektion, der institutionalisierten Leistungsmessung und der Unterstützungssysteme wie LAs zu sein, die in Englands Schulen tatsächlich eine Entwicklung in Richtung auf eine höhere Verantwortlichkeit der Einzelschulen angestossen hat. Auch wenn die Test und Inspektionen auch eine Belastung für die Schulen darstellen, scheint trotzdem eine recht hohe Akzeptanz dafür zu bestehen (vgl. Atkinson et al.). Dennoch zeigen sich teilweise Probleme der Kompatibilität der einzelnen Funktionen. Einerseits sollen die Schulen einer Kontrolle unterzogen werden und andererseits bei ihrer eigenen Entwicklung unterstützt werden. Wie die Studie von Atkinson et al. zeigt, sind Schulen, die gut bewertet werden, auch stärker an einer Weiterentwicklung interessiert, während Schulen, die einen weniger guten Bericht erhalten, häufiger mit Resignation oder Ablehnung reagieren. Die Wirkung der Entwicklungsfunktion scheint somit relativ stark von den Kontrollresultaten der externen Evaluation abzuhängen.

Selbstverständlich ist ein solch umfassendes und komplexes Monitoringsystem nicht ohne eine solide Finanzierungsbasis umsetzbar. Die Kritik an diesem System begründet sich dementsprechend häufig auf der Sorge, einen zu hohen finanziellen Aufwand zu betreiben und dabei zu wenig Effekt zu erzielen (van Ackeren 2003).

3.3 Beispiele aus den Niederlanden

Die Niederlande zeichnet sich im internationalen Vergleich durch eine grosse Autonomie in wesentlichen Bereichen der Schulorganisation aus. Die bereits

1917 verfassungsmässig festgeschriebene Freiheit zur Gründung, Richtung und Einrichtung von Schulen ist dabei eines der auffälligsten Charakteristika. Das niederländische Schulwesen wird über eine Pauschalfinanzierung, die sich nach der Anzahl der Schülerinnen und Schüler an der jeweiligen Schule richtet, vollständig vom Staat finanziert. Für Lernende mit besonderem Förderbedarf werden zusätzliche Mittel bereitgestellt. Der Staat greift aber nicht in die pädagogischen oder organisatorischen Belange der Schulen ein. So sind in den Niederlanden sowohl die privaten als auch die öffentlichen Schulen hinsichtlich der Gestaltung der Unterrichtsprozesse und der Bestimmung ideologischer, religiöser oder pädagogischer Prinzipien relativ wenigen Beschränkungen unterworfen. Statt staatlicher Lehrpläne werden vom Ministerium verbindliche Kernziele formuliert und für die verbindlichen Fächer werden Mindestunterrichtszeiten vorgegeben die 75% des Gesamtstundenvolumens ausmachen. Die Gestaltung des übrigen Unterrichts liegt in der Verantwortung der Einzelschule. Jede Schule hat die Aufgabe, ein eigenes Profil zu entwickeln, das sich nur in den strukturellen Rahmen eines zweistufigen Systems aus Basisschulen (achtjährige gemeinsame Grundschule) und Sekundarschulen einpassen muss. Die Verantwortung für die Qualität des pädagogischen Angebots liegt in wesentlichen Teilen also bei der Einzelschule. Im Profil sind unter anderem Aspekte der Unterrichts-, Personal- und Qualitätsentwicklung sowie der Evaluation enthalten. Unterstützung bei der Organisation dieser Arbeit erhalten die Schulen bei den sogenannten Schulbegleitungsdiensten.

Zu der erweiterten Autonomie der Einzelschule kommt noch die Besonderheit der freien Wahl der Schule, die in den Niederlanden eine lange Tradition hat. Eltern können, von einzelnen regionalen Ausnahmen abgesehen, die Schulen für ihre Kinder frei und unabhängig vom Wohnort wählen. Deshalb sind die Schulen dazu verpflichtet, regelmässige Selbstevaluationen durchzuführen und einen „Schulführer“ herauszugeben, in dem sie über ihre Ziele und Errungenschaften berichten.

Leistungsmessung und Inspektion

Bedingt durch die erweiterte Autonomie kommen der Leistungsmessung und der Inspektion in niederländischen Schulen eine besondere Bedeutung zu. Sie stellen ein wichtiges Instrument zur Vereinheitlichung schulischer Inhalte und Anforderungsniveaus dar.

Es bestehen unterschiedliche Formen der Leistungsmessung (vgl. van Ackeren 2003). Zum einen gibt es freiwillige Tests am Ende der Primarstufe, die jedoch einen quasi obligatorischen Charakter dadurch erhalten, dass sie seit 1994 von den abnehmenden Sekundarschulen als Eingangsvoraussetzung angesehen werden. Am verbreitetsten ist der Eintoets Basisonderwijs, der von ca. 80 Prozent der Primarschulen eingesetzt wird und vom Centaal Instituut voor Toetsontwikkeling (CITO) jedes Jahr neu entwickelt und

erprobt wird.

Das CITO ist ein zentrales Institut für Testentwicklung, das seit 1968 besteht und 1999 vollständig privatisiert wurde. Die Hauptaufgabe des CITO besteht darin, professionelle Tests, zentrale Endexamina und Leistungsstudien zu entwickeln, einzusetzen und auszuwerten. Für den Bereich der Primarschule werden von CITO zusätzlich zu dem oben genannten Eintoets Basisonderwijs auch diagnostische Tests zur Verfügung gestellt, die von den Lehrpersonen zur Evaluation und zur Schulentwicklung eingesetzt werden können. Dazu gehören Lernzuwachstests mit externen, nationalen Vergleichsmassstäben und das Leerlingvolgsysteem, das eher auf eine Ursachenbestimmung individueller Leistungen einzelner Schüler abgestimmt ist. Am Ende der Sekundarstufe stehen dann die ebenfalls von CITO entwickelten obligatorischen zentralen Abschlussprüfungen, deren Ergebnisse zu einem nicht unerheblichem Teil die Abschlussnote der Schülerinnen und Schüler mitbestimmen.

Neben diesen schullaufbahnbegleitenden Tests und den Abschlusstests werden Schulleistungsstudien zum Zweck des Systemmonitoring durchgeführt. Dazu gehört der ebenfalls von CITO durchgeführte PPON (Periodieke Peiling van het Onderwijsniveau) und der PRIMA (Landelijk Cohorten-Onderzoek Primair Onderwijs) der Katholischen Universität Nijmegen und der Universität Amsterdam.

Das Inspektorat, heute durch die nationale Schulaufsicht (Inspectie van het Onderwijs) durchgeführt, hat in den Niederlanden eine lang gewachsene Tradition seit über 200 Jahren. Die Hauptaufgaben des Inspektorats sind in der Kontrolle und Evaluation der Schulen sowie der Förderung von Qualitätsentwicklung und der öffentlichen Berichterstattung zu sehen. Die Berücksichtigung der Ergebnisse der CITO-Tests im Rahmen der Schulinspektion stellt eine wichtige Verbindung zwischen der Leistungsmessung und der externen Schulevaluation dar.

Im folgenden werden ausgewählte Modelle der Leistungsmessung in den Niederlanden sowie das Vorgehen des Inspektorats näher beleuchtet, wobei ein besonderes Augenmerk auf die Verbindung der beiden Ansätze gelegt werden soll.

3.3.1 Eintoets Basisonderwijs (CITO-Test)

Der Eintoets Basisonderwijs der auch CITO-Test genannt wird, dient in erster Linie einer Orientierung für die abnehmenden Sekundarschulen. Dadurch, dass die Sekundarschulen vor der Aufnahme einen Abschlusstest der Primarschule verlangen, bekommen diese Tests einen quasi obligatorischen Charakter für die Primarschulen und einen quasi selektiven Charakter für die Schülerinnen und Schüler die sich am Übergang zur Sekundarschule mit den Testergebnissen präsentieren müssen. Allerdings bleibt es der Primarschule freigestellt, welchen Test sie einsetzen will. Der Eintoets Basisonderwijs ist

mit ca. 85 Prozent inzwischen der am häufigsten verwendete Primarabschlusstest in den Niederlanden. Er zeichnet sich dadurch aus, dass er ein relativ breites Spektrum an Leistungen in Niederländisch, Mathematik, Problemlösen und Weltkunde abprüft und abgestützt auf eine wissenschaftliche Begleitforschung präzise Vorhersagen über die Entwicklungen der Schülerinnen und Schüler in unterschiedlichen Sekundarschultypen zulässt. Diese Information dient auch den Eltern und Primarschullehrpersonen zur Orientierung über die mögliche weitere Schullaufbahn des Schülers oder der Schülerin. Eine Rückmeldung darüber, wie die Einzelschule allgemein im Vergleich zum nationalen Durchschnitt und zu anderen vergleichbaren Schulen abgeschnitten hat, soll zusätzliche Impulse für die Schulentwicklung liefern. Schliesslich sollen die Testergebnisse auch dem Systemmonitoring dienlich sein. Es wird dem CITO-Test also ein relativ breites Spektrum an Funktionen zugeschrieben.

Die Einordnung der Ergebnisse des CITO-Tests orientiert sich in erster Linie an der sozialen Bezugsnorm, also dem Vergleich mit anderen Schulen sowie mit dem nationalen Durchschnitt. Eine Veröffentlichung aggregierter Ergebnisse war ursprünglich nicht vorgesehen, bis eine Zeitung unter Berufung auf das Recht auf Informationsfreiheit die Publikation von Test- und Prüfungsergebnissen auf Schulebene einforderte und dieses Recht gerichtlich zugesprochen bekam. Das Inspektorat stimmte daraufhin 1997 einer regelmässigen Veröffentlichung der Testergebnisse zu.

Die Aufgaben für den CITO-Test werden jedes Jahr neu von Primarschullehrpersonen und Fachdidaktikern unter der Leitung von CITO-Mitarbeitern entwickelt und in einem Pilottest erprobt. Es handelt sich ausschliesslich um Multiple-Choice-Aufgaben. Im Rahmen der Hauptuntersuchung wird der Test unter der Aufsicht der Lehrpersonen in den eigenen Klassen durchgeführt. Dies geschieht inzwischen bereits zu 10 bis 15 Prozent am Computer, in den meisten Fällen werden jedoch noch Papier-und-Bleistift-Tests eingesetzt. Die Antwortbögen werden anschliessend an CITO zurückgeschickt und dort ausgewertet. Drei Wochen später werden die individuellen und die kumulierten Ergebnisse an die Schulen zurückgemeldet (für die Individualrückmeldung siehe Anhang F.1).

Durch die Veröffentlichung der Ergebnisse des Tests ist die Bedeutung, die dieser Test für das Renommee der Einzelschule hat, erheblich gestiegen. So kommt es teilweise dazu, dass versucht wird, das Resultat zu manipulieren, indem überdurchschnittlich viele schwächere Schüler an den Testtagen krank geschrieben werden oder Lehrpersonen bei der Bearbeitung der Aufgaben helfen. Diese Probleme zeigen sehr deutlich auf, welche Konsequenzen eine Kumulation von Funktionen eines Tests haben kann.

3.3.2 Leerlingvolgsysteem

Mit dem Leerlingvolgsysteem hat CITO ein System längschnittlicher Messungen von Schülerleistungen geschaffen, das nicht nur punktuelle Aussagen über den Lernstand der Schülerinnen und Schüler zulässt, sondern auch die Entwicklung der Leistung in die nähere Betrachtung zieht. Leistungsmessungen über die gesamte Primarschulzeit, die in den Niederlanden 8 Jahre beträgt, können so in ein schulinternes Monitoringsystem aufgenommen werden. Das Leerlingvolgsysteem stellt Tests aus den Bereichen Sprache/Lesen, Rechnen/Mathematik und Weltorientierung sowie spezielle Tests für Kinder zwischen 4 und 6 Jahren zur Verfügung.

Die Besonderheit des Leerlingvolgsysteem ist seine spezielle Funktion als Instrument der Unterrichtsentwicklung, in der es ein konkretes Hilfsmittel für die Lehrpersonen zur systematischen Überwachung von Lernfortschritten und Unterrichtsqualität darstellt. Das Leerlingvolgsysteem wird von den Einzelschulen erworben und intern eingesetzt und in der Regel mittels einer speziell dazu ausgelegten Software ausgewertet. Das System beinhaltet unterschiedliche Tests mit steigenden Schwierigkeitsgraden, die in halbjährlichen Abständen die Schülerleistungen messen. Die Tests können als Papier-und-Bleistift-Tests oder am Computer durchgeführt werden. In der neuesten Version stehen auch adaptive Tests zur Verfügung. Mit Hilfe von probabilistischen Testskalierungsverfahren (Item-Response-Theory)[7] bei der Auswertung der Daten kann sichergestellt werden, dass die gemessene Leistung aus den unterschiedlichen Tests auf einer gleichen Metrik abgebildet werden kann. Dadurch wird es möglich, Lernfortschritte von einem Testzeitpunkt zum nächsten und über die gesamte Primarschulzeit hinweg zu beobachten und zueinander in Beziehung zu setzen. Die Analysen können unter Verwendung der speziellen Software problemlos schulintern durchgeführt werden.

Aufgrund einer vorausgegangenen Normierungsstudie können die Ergebnisse der Schülerinnen und Schüler präzise in einen nationalen Vergleichsrahmen mit 5 unterschiedlichen Niveaustufen eingeordnet werden. Für jede der 5 Niveaustufen wird aufgrund der Ergebnisse der Normierungsstudie eine bestimmte Lernentwicklung angenommen. Entwickelt sich ein Schüler oder eine Schülerin nun erwartungswidrig und fällt im Laufe der Primarschulzeit von einer höheren auf eine tiefere Niveaustufe ab, so kann die Lehrperson die Art der Probleme genauer analysieren und ermitteln, ob zusätzliche Unterstützungsangebote notwendig sind und inwiefern Lehrstoff und Unterricht stärker auf die Bedürfnisse der Schülerinnen und Schüler abgestimmt werden müssen. Für diese Schritte stellt das Leerlingvolgsysteem neben den Testmaterialien für die einzelnen Jahrgangsstufen umfangreiches Analysematerial sowie didaktische Anweisungen und Hilfsmittel zur Verfügung. Mithilfe der Software können die Daten sehr umfangreich analysiert und graphisch aufbereitet werden.

[7] siehe Erklärung im Anhang A

Es handelt sich bei dem Leerlingvolgsysteem nicht ausschliesslich um ein Testinstrumentarium zur Überprüfung von Schülerleistungen und Lernentwicklung, sondern konkret um ein Entwicklungsinstrument, in dem die Unterrichtsentwicklung im Konzept bereits mitgedacht wurde. Dieses Konzept sieht es vor, dass die Ergebnisse der Tests regelmässig schulintern besprochen und unter Hinzuziehen anderer Informationen über einzelne Schülerinnen und Schüler analysiert werden. Ferner sieht das Konzept des Leerlingvolgsysteem es vor, dass auf der Grundlage der Analysen konkrete Handlungspläne für individuelle Schülerinnen und Schüler oder für ganze Klassen erstellt und ausgeführt werden, deren Resultate sich dann evtl. in den Lernfortschritten der Schülerinnen und Schüler messbar widerspiegeln. Die Ergebnisse aus diesen Tests werden ausschliesslich intern verwendet. Das Inspektorat kann nur dann Einsicht in die Daten bekommen, wenn dies explizit von der Schule gewünscht wird, etwa weil sie sich dadurch eine weitere Hilfestellung bei der Umsetzung der Schulentwicklungspläne verspricht.

3.3.3 Inspektorat der nationalen Schulaufsicht

Die nationale Schulaufsicht führt regelmässig Inspektionen in allen Schulen in den Niederlanden durch, um die Qualität des Bildungsangebotes zu überprüfen. Es soll damit erreicht werden, dass für alle Schulen, unabhängig von ihrer Schwerpunktsetzung ein einheitlicher Standard hinsichtlich vorgegebener Qualitätsaspekte gewährleistet werden kann. Die Leitung der Inspektion setzt sich zusammen aus dem Generalinspekteur, drei Hauptinspekteuren sowie einem Direktor des Managements. Von den ungefähr 550 im Inspektorat beschäftigten Personen, sind ca. 250 direkt mit der Inspektion befasst. Die Inspektion arbeitet trotz der Finanzierung durch das Ministerium für Bildung und Kultur weitgehend unabhängig.

Der im September 2002 in Kraft getretene Bildungsinspektionserlass (Wet op het onderwijstoezicht, WOT) regelt die Verantwortlichkeit der Einzelschule für die Qualität des Unterrichts sowie für die Qualität der Leistungsmessungen und Evaluationen. Die Inspektion verwendet die Resultate der schulinternen Qualitätsüberprüfungen, um spezieller auf die Bedürfnisse einzelner Schulen eingehen zu können und um die Zeit der Inspektion möglichst gering zu halten. Als Grundlage für die Inspektionen dient der Supervisionsrahmen „Toezichtkader Primair Onderwijs 2005“ (Inspectie van het onderwijs 2005), der im Zuge des neuen Inspektionserlasses formuliert wurde. Eine im Dezember 2005 herausgegebene Version löst die Vorgängerversion vom Januar 2003 ab. Der Supervisionsrahmen soll die Einheitlichkeit und Präzision sowie die Transparenz der Inspektionen gewährleisten. Im Vergleich zu den früheren Inspektionen wird demnach heute mehr Wert auf die Erfolge und Resultate einer Schule gelegt als auf deren konkrete Arbeitsweise. Der Autonomie der Schulen hinsichtlich eigenständiger Planung und Gestaltung ihres Unterrichts soll damit Rechnung getragen werden. Deshalb werden

verstärkt die internen Qualitätsüberprüfungen und Evaluationen einbezogen. Die gegenwärtige Praxis der Inspektion in den Niederlanden sieht seit dem neuen Inspektionserlass eine sogenannte proportionale Supervision vor. Danach hängt die Häufigkeit und die Form der Inspektion von der Qualität der Schule und dem Risiko eines Qualitätsverlustes ab. Diese Aspekte werden mit Hilfe eines Kriterienkataloges eingeschätzt. Ausserdem werden die Ergebnisse der Selbstevaluation so einbezogen, dass Qualitätsaspekte, die intern bereits ausreichend analysiert wurden, nicht mehr Gegenstand einer weiteren Untersuchung der Inspektion werden. Es besteht auch die Möglichkeit in Ausnahmefällen und in Abstimmung mit der Schule weitere Aspekte einzubeziehen, sofern die Inspektoren eine Notwendigkeit dafür sehen, oder die Schule es wünscht.

Qualitätsindikatoren

Der Kriterienkatalog (siehe Anhang F.2) umfasst 27 Indikatoren, für die die Inspektoren auf einer 4-stufigen Skala eine Einschätzung vornehmen. Die Basisindikatoren beinhalten Aspekte zum Lehrstoffangebot, zur Lern- und Unterrichtszeit, zum pädagogischen und didaktischen Handeln, zur Abstimmung auf die Bedürfnisse der Lernenden, zu selbständigen Aktivitäten der Lernenden, zum Schulklima, zur Begleitung und Förderung und zum Lernstand am Ende der Schulzeit sowie zur Lernentwicklung der Schülerinnen und Schüler. Zusätzlich gibt es Indikatoren, die die interne Qualitätskontrolle der Schule beschreiben. Lernstand und Lernentwicklung wird auf der Grundlage der Abschlusstests eingeschätzt, die die Schulen obligatorisch durchführen. Die Verwendung von Tests mit repräsentativer Referenzstichprobe ermöglicht es, die Schule auch im Vergleich und unter Berücksichtigung der Kontextfaktoren zu betrachten und die tatsächlichen Ergebnisse mit Erwartungswerten abzugleichen.

Formen der Inspektion

Aufgrund der Einschätzung der Inspektoren über die Leistungen der Schülerinnen und Schüler und über das Angebot der Schule wird entschieden, ob die Qualität der Schule ausreichend ist. Davon hängt dann die Form und die Häufigkeit zukünftiger externer Evaluationen ab. In den Niederlanden gibt es fünf unterschiedliche Typen von Inspektionen.

1. Jährliche Inspektionen,
2. Periodische Inspektionen (PQI),
3. Weitere Inspektionen,
4. Qualitätssteigerungsinspektionen,
5. unregelmässige Inspektionen.

In der Regel werden Schulen einmal in vier Jahren im Rahmen einer

periodischen Qualitätsinspektion (PQI) (2) evaluiert. Dabei werden nach Berücksichtigung von Selbstevaluationen, Eindrücken aus Unterrichtshospitationen, Gesprächen mit Leitungspersonen, Lehrpersonen und Lernenden und der Betrachtung der Ergebnisse aus den Leistungstests Einschätzungen zu den Indikatoren abgegeben. In den darauf folgenden zwei Jahren werden in jährlichen Inspektionen (1) Fragebögen zu den Leistungen der Schule ausgewertet sowie Dokumentenanalysen (beispielsweise zu Selbstevaluationsberichten) durchgeführt. Das Qualitätsprofil der Schule wird gegebenenfalls dementsprechend angepasst. Anhand einer Risikoanalyse wird jedes Jahr neu eingeschätzt, ob der vorgegebene Inspektionszyklus ausreichend ist, oder ob weitere Massnahmen ergriffen werden müssen. Auch im dritten Jahr wird eine jährliche Untersuchung durchgeführt, bei der allerdings zusätzlich ein Schulbesuch stattfindet. Aufgrund der in diesem Rahmen geführten Gespräche werden Verabredungen zur Schwerpunktsetzung für die im folgenden Jahr durchzuführende periodische Qualitätsinspektion (PQI) getroffen. Wenn die Ergebnisse der PQI den Verdacht auf mangelnde Qualität aufbringen, wird innerhalb von 6 Monaten eine weitere Inspektion (3) vorgenommen. Sie orientiert sich im wesentlichen an den Punkten, die während der PQI als problematisch identifiziert wurden. Wenn auch diese weitere Inspektion den ursprünglichen Verdacht auf mangelnde Qualität der Schule bestätigt, wird eine Qualitätssteigerungsinspektion (4) zur Supervision eingesetzt. Beratungsaufgaben werden explizit jedoch nicht von der Inspektion übernommen. Die Leitung der Schule muss dem Inspektorat eine Stellungnahme zur Einschätzung der Mangelsituation und zu den diesbezüglich geplanten Massnahmen zustellen. Die Massnahmen werden dann danach überprüft, ob sie zielgerichtet, konkret, realistisch und erfolgversprechend sind und die Schule erhält eine Rückmeldung darüber. Nach spätestens zwei Jahren werden die durchgeführten Massnahmen und die Qualitätskriterien im Rahmen einer erneuten Inspektion evaluiert. Wenn diese Inspektion weiterhin die Qualitätsmängel bestätigt, wird die Schule verwarnt. In Fällen, in denen mit einer Verbesserung der Situation innerhalb eines Jahres gerechnet werden kann, kann die Supervisionszeit um ein weiteres Jahr verlängert werden. Wenn auch dies nicht zu einer Verbesserung führt, wird das Ministerium benachrichtigt, um weitere administrative Massnahmen (wie z.B. Bestellen eines externen Experten zur Unterstützung der Schulleitung, erweiterte finanzielle Ressourcen etc.) vorbereiten zu können. Weiterhin kann das Inspektorat auch unregelmässige Evaluationen (5) durchführen. Dies ist in der Regel der Fall, wenn spezifische Beschwerden gegenüber einer Schule geäussert werden.

Darstellung der Ergebnisse

Die Ergebnisse der Inspektionen (ausser der unregelmässigen Inspektionen) werden für jede Schule im Internet veröffentlicht. Die sogenannten Quali-

tätskarten (siehe Anhang F.4) geben Auskunft über die Einschätzung der Inspektion zu den 27 Indikatoren. Bei Wunsch kann die Schule eine Stellungnahme zu der Qualitätskarte formulieren, auf der eigenen Internetseite zur Verfügung stellen und mit der Qualitätskarte direkt verlinken. Ausserdem wird eine etwas ausführlichere Darstellung in einem etwa 15-seitigen Inspektionsbericht (für ein Beispiel siehe Anhang F.3) festgehalten und ebenfalls im Internet zur Verfügung gestellt. Als Teilbereich der Qualitätsindikatoren stellt die Beurteilung der Ergebnisse der Leistungsmessung einen festen Bestandteil des Inspektionsberichts und der Qualitätskarten dar.

3.3.4 Zusammenfassung und Beurteilung

In den Niederlanden besteht ähnlich wie in England eine enge Verbindung zwischen der externen Evaluation durch das Inspektorat und der schulischen Leistungsmessung. Drei der 27 Basisindikatoren zur Beurteilung von Schulen im Rahmen der periodischen Qualitätsinspektion hängen explizit mit den Ergebnissen der Leistungsmessung zusammen (siehe Anhang F.2, Indikatoren 12.1, 13.1 und 13.2). Die Ergebnisse des CITO-Tests gehen direkt in die Überlegungen zur Vorbereitung und Durchführung der Inspektion ein und spielen auch in der internen Evaluation der Schulen eine Rolle. Die Erkenntnisse aus dem Leerlingsvolgsysteem können den Inspektoren offengelegt werden, wenn dies von der Schule gewünscht wird. Dadurch dass die Leistung der Lernenden einen wesentlichen Aspekt der Inspektion ausmacht findet sie sich auch explizit im Inspektionsbericht wieder. Dass die Auswahl des geeigneten Tests für die Leistungsmessung am Ende der Grundschulzeit bei der Schule selbst liegt, überrascht nicht bei dem Ausmass an Autonomie, das in den Niederlanden generell den Schulen zugesprochen wird. Dieser Autonomie ist es eventuell auch zuzuschreiben, dass die Qualitätsentwicklung selbst ein ausserordentlich wichtiger Aspekt im Rahmen der Inspektion ist. Schulen sind dazu angehalten, Selbstevaluationen durchzuführen und diese Leistung wird dadurch anerkannt, dass die Ergebnisse der Selbstevaluation die Untersuchungen durch das Inspektorat ersetzen können. Die Autonomieunterstützung beschränkt sich aber nicht nur auf die Aspekte der Leistungsmessung und der Inspektion, sondern spiegelt sich auch in den Angeboten zur Qualitätsentwicklung wider. Das Leerlingvolgsysteem ist ein gutes Beispiel dafür, welche Möglichkeiten Schulen in den Niederlanden haben, aus eigener Initiative heraus näheren Einblick in die Situation der Schule zu nehmen und den Schulentwicklungsprozess darauf abzustützen. Die Tests selbst werden sehr professionell und auf hohem qualitativen Niveau entwickelt und geeicht und die Ergebnisse werden mit Hilfe modernster Verfahren ausgewertet. Durch den hohen Grad an Standardisierung sind die Tests jedoch auf spezifische Bereiche messbarer Kompetenzen beschränkt.

Die Autonomie der Schulen bringt natürlicherweise ein höheres Mass an Verantwortlichkeit und die Notwendigkeit öffentlicher Rechenschaftslegung

mit sich. Deshalb kommt den Ergebnissen aus der Inspektion und der Leistungsmessung eine so hohe Bedeutung zu. Schülerinnen und Schüler sowie Eltern erhalten, in einer Situation, in der sie mit einer enormen Vielfalt von schulischen Angeboten und Schwerpunktsetzungen konfrontiert sind, Einschätzungen zu einheitlich gehaltenen Variablen. Auch wenn die Konsequenzen ungünstiger Beurteilungen eher die Form von Hilfestellungen als die Form von Sanktionen annehmen, sorgt die Öffentlichkeit letztendlich dafür, dass die Inspektionen und die Leistungsmessungen mit dem notwendigen Ernst betrachten werden.

4 Zusammenfassung und Ausblick

In allen der betrachteten vier Länder bestehen umfangreiche Modelle der Leistungsmessung und der externen Evaluation, die freilich unterschiedliche Funktionen erfüllen und in unterschiedlicher Weise auf die spezifischen politischen Anforderungen der einzelnen Länder reagieren. Eine konsequente Verknüpfung von Leistungsmessung und Schulevaluation kann derzeit jedoch nur in England und in den Niederlanden beobachtet werden, auch wenn sich beide Modelle hinsichtlich der Aussagekraft über Schulqualität ganz erheblich voneinander unterscheiden.

Aufgabe der externen Evaluation ist die Bestandsaufnahme der Qualität von Schule nach einem System von schulübergreifend vergleichbaren Indikatoren. Dazu gehört neben einer Reihe von institutionellen und organisatorischen Merkmalen auch die Qualität der Lehre.

Testeigenschaften

Häufig werden Erhebungen des Lernstandes von Schülerinnen und Schülern zur Beschreibung der Qualität von Lehre herangezogen. Um ein konsistentes Abbild der Schülerleistung zu bekommen, ist es notwendig, dass die Ergebnisse schulübergreifend vergleichbar sind und Kompetenzen und Fähigkeiten erfassen, die für Schulleistung als relevant angesehen werden. Ohne die Festlegung eines gemeinsamen Kriterienkatalogs für Schülerleistungen ist eine Beurteilung der Ergebnisse eines Tests relativ willkürlich. Wenn aber ein Bereich der Qualität von Schule gemessen werden soll, so sollte auch definiert sein, was unter dieser Qualität zu verstehen ist. Die Festlegung von Kompetenzmodellen in den Kernlehrfächern ist der Schritt in diese Richtung und standardorientierte Tests beginnen, einen Vergleich mit einem vorgegebenen Kriterium zu ermöglichen.

Dabei muss sehr sorgfältig und unter Zuhilfenahme der technischen Möglichkeiten darauf geachtet werden, dass die Kompetenzen, die die Tests vorgeben zu messen, auch tatsächlich gemessen werden. Um dies zu gewährleisten, sind, wie das Beispiel der englischen „National-Curriculum-Tests“ zeigt, teilweise extrem aufwendige Testentwicklungs- und Auswertungsprozeduren

notwendig. Das Abwägen von Nutzen und Effektivität bzw. die Frage nach der Wirtschaftlichkeit dieser aufwendigen Testverfahren ist sicherlich ein gewichtiger Faktor, der für die unterschiedlich komplexen Ausprägungen der Modelle in den einzelnen Ländern verantwortlich ist.

Wer wird informiert, wer ist verantwortlich?

Häufig werden aufgrund dieser Wirtschaftlichkeitsüberlegungen den Leistungsmessungen mehrere Funktionen zugesprochen, die sich teilweise gegenseitig behindern, da die Verantwortung für die Testergebnisse geteilt wird. Je höher der Rechenschaftsdruck ist, desto höher ist in der Regel auch die Tendenz zur Manipulation. Wenn aber die Ergebnisse mit Hinblick auf eine Funktion manipuliert werden, so stellen sie sich auch für die andere Funktion nicht mehr korrekt dar. Die Gefahr der Manipulation steigt ebenfalls mit der Anzahl der mit Rechenschaftsdruck verbundenen Funktionen.

Die Frage danach, wer auf welcher Ebene für die Ergebnisse der Leistungsmessungen gegenüber wem und in welcher Weise verantwortlich ist, ist deshalb entscheidend bei der Abwägung einer sinnvollen Verwendung von Testergebnissen für Qualitätsaussagen. Dabei ist zum einen zu berücksichtigen, welche Personen Einblick in die Daten haben und zum anderen, wer für die Ergebnisse zur Verantwortung gezogen wird. Dies gilt jeweils genauso auf Individualebene wie auf Klassen-, Schul- oder Systemebene.

Berücksichtigung weiterer Variablen

Nun ist, wie in der Einleitung dieses Berichtes ausgeführt, der Lernstand der Schülerinnen und Schüler für sich genommen, noch kein Indikator für die Qualität des Unterrichts in einer Schule. Zu viele andere Faktoren können dabei eine Rolle spielen. Ebenso wie ein hoher Leistungsstand durch eine hohe Unterrichtsqualität erklärt werden könnte, könnte er auch gänzlich ein Effekt der Eingangsselektion sein. Wenn keine zusätzlichen Variablen betrachtet werden, ist der Leistungsstand der Schülerinnen und Schüler für eine Aussage über die Qualität von Schule nicht brauchbar.

Für die in Schulleistungserhebungen involvierten Personen ist dies seit längerem evident. In der Regel werden die Leistungstests deshalb mit umfangreichen Fragebögen zu Hintergundvariablen begleitet. Auf diese Weise können bei der Beurteilung der Schülerleistungen Variablen berücksichtigt werden, von denen ein Einfluss auf die Leistung zu erwarten ist. Der Sozialindex einer Schule oder die Anzahl der Schülerinnen und Schüler mit fremdsprachlichem Hintergrund sind prominente Beispiele für berücksichtigte Variablen. Es können auf diese Weise im Vergleich mit Schulen ähnlicher Sozialstruktur Erwartungswerte formuliert werden, die dabei helfen, die Leistungen der Schülerinnen und Schüler einer Schule besser in einem „fairen Vergleich“ einordnen zu können.

Lernentwicklungsmessung

Dennoch lässt auch ein „fairer Vergleich“ im Grunde keine Aussagen über die Qualität des Unterrichts zu, da auch bei der Berücksichtigung von relevanten Hintergrundvariablen der Einfluss einer Eingangsselektion nicht auszuschliessen ist. Lediglich längsschnittliche Untersuchungen, aus denen die Lernentwicklung des einzelnen Schülers über eine längere Zeitspanne abzulesen ist, korrigieren den Effekt der Eingangsselektion. Zu berücksichtigen ist allerdings dabei, dass auch die Lernentwicklung nicht ausschliesslich dem Unterricht, sondern zusätzlich einer Reihe weiterer Faktoren aus Persönlichkeit und alltäglichen Umfeld des Schülers zuzuschreiben ist. Eine längsschnittliche Erhebung von Leistungsdaten in Verbindung mit der Berücksichtigung von Hintergrundvariablen ist deshalb derzeit die beste Annäherung an realistische Aussagen über leistungsbezogene Schulqualität.

Verbindung von Leistungsmessung und externer Schulevaluation

Um mehr über die Lernentwicklung zu erfahren und dadurch eine bessere Grundlage für den „fairen Vergleich“ zu bekommen, ist es notwendig, die Verbindung der externen Schulevaluation mit einer umfassenden Leistungsmessung, bzw. Lernentwicklungsmessung zu suchen. Die Erhebung weiterer relevanter Informationen im Rahmen der externen Schulevaluation stellt dabei eine besondere Chance dar, die Ergebnisse der Leistungsmessungen in einem breiteren Kontext zu interpretieren. Wenn eine möglichst realistische Aussage über eine auf Schülerleistung bezogene Schulqualität getroffen werden soll, so ist es unerlässlich, dass die Ergebnisse aus den Leistungserhebungen und der externen Schulevaluation in gemeinsamer Berücksichtigung verarbeitet werden.

Aus den hier dargestellten Bedingungen zur Darstellung von schülerleistungsbezogener Schulqualität lassen sich die zehn in der Einleitung dieses Berichts genannten Fragen ableiten. Im folgenden wird diskutiert, inwieweit sich die Modelle in den einzelnen Ländern zur Abbildung von Schulqualität in diesem Sinne eignen. Die Besonderheiten der Modelle sind zusätzlich im letzten Kapitel 5 aufgeführt, strukturiert nach den Leitfragen.

Schweiz

In der Schweiz bestehen unterschiedliche Modelle der Leistungsmessung, die sehr unterschiedliche Funktionen abdecken. Grösstenteils stützen sie sich auf Lehrplanvorgaben und sind speziell für den Einsatz in einzelnen Kantonen entwickelt, auch wenn sie teilweise kantonsübergreifend adaptiert werden. Eine Verbindung von Leistungsmessung mit externer Schulevaluation hat bisher in der Schweiz noch nicht stattgefunden, sieht man einmal von den ersten Versuchen der Einbindung der Check8-Ergebnisse bei einzelnen Evaluationen ab. Wie sich die Leistungsmessungen in der Schweiz momentan

darstellen, ist auch fraglich, ob sie dazu geeignet sind, Fragen zur Qualität von Schule beantworten zu können. Für die Orientierungsarbeiten und den Check5 ist dies aus Gründen der Funktionsüberlappung auch explizit nicht vorgesehen. Der Stellwerktest, der probeweise in die externe Evaluation eingebunden wurde, bringt in diesem Zusammenhang andere Probleme mit sich. In der Funktion als Test für ein Systemmonitoring kann er einen schnellen und groben Überblick über die Leistungsstände der Lernenden geben. Um aber eine Qualitätsbeurteilung auf der Grundlage dieses Tests vornehmen zu können, wären mehr Informationen zu spezifischen Leistungsbereichen, bzw. die Möglichkeit der Reflexion über die Testaufgaben notwendig. Die Ergebnisse sind schwierig nachzuvollziehen und es werden kaum Hintergrundvariablen zur vertiefenden Betrachtung und zum Vergleich mit Kontextgruppen bereitgestellt. Wie auch die anderen Leistungstests in der Schweiz kann der Stellwerktest bisher keine Aussagen über die Lernentwicklung machen. Dazu kommt, dass er mit einer grossen Anzahl von Funktionen belastet ist, die Rechenschaftspflicht aber zukünftig im wesentlichen bei den Schülerinnen und Schülern allein liegt. Wenn die Rechenschaftspflicht auf die Schulen ausgeweitet wird, kann aufgrund der Möglichkeit der Manipulation die Funktion der Individualselektion gleichzeitig in Frage gestellt werden.

Deutschland

In Deutschland stellt sich die Situation in einer ähnlichen Weise dar. Zwar bestehen unterschiedliche Formen der Leistungsmessung ebenso wie Institutionen, die eine externe Schulevaluation durchführen. Meistens haben die einzelnen Bundesländer sich aber entweder für die Leistungsmessung oder für die Schulevaluation entschieden. Verknüpfungen zwischen den zwei Bereichen gibt es bisher kaum. Erste Versuche eine Verbindung herzustellen laufen in Nordrhein-Westfalen und auch Hamburg hat konkrete Planungen für die Umsetzung eines solchen Modells. Erfahrungsberichte liegen noch nicht dazu vor, doch die Überlegungen hinsichtlich der Probleme bei der Überschneidung von Funktionen gelten ähnlich wie auch für die Schweiz. Es ist in diesem Zusammenhang zu überlegen, welche der Leistungstests sich für die Einbindung in die Evaluation eignen. In Hamburg ist beispielsweise geplant, die Ergebnisse der bisher verwendeten Vergleichsarbeiten einzubinden, während für die bundesländerübergreifenden Lernstandserhebungen (VERA und Lernstandserhebungen für die 8. Klasse) zunächst noch keine Verwendung in der externen Schulevaluation geplant ist. Dies wird damit begründet, dass die Vergleichsarbeiten sich spezifischer an Hamburgs Lehrplan orientieren. Man möchte man auf diese Weise zumindest einen direkten Bezug zum schulischen Lehrplan und damit eine Annäherung an Schulqualitätsaspekte erreichen. Ohne längsschnittliche Untersuchungen muss diese Annäherung allerdings noch recht vage bleiben. Zusätzlich soll dadurch eine Kumulation von Funktionen der Leistungstests vermieden werden. Dabei spielt die Überlegung eine

Rolle, dass eine mit der Verwendung im Rahmen der Inspektion verbundene Rechenschaftspflicht möglicherweise zu Manipulationen führt und damit die Lernstandserhebungen beeinträchtigt. In Nordrhein-Westfalen hingegen geht auch ein Bericht über die Lernstandserhebungen in das Schulportfolio ein, das für die Inspektion zur Verfügung gestellt werden soll. Erfahrungsberichte darüber, wie dieser Bericht im Rahmen der Inspektion verarbeitet wird, liegen jedoch derzeit noch nicht vor.

England

Das englische Modell zeigt sehr deutlich auf, welche Möglichkeiten es bereits mit den derzeit zur Verfügung stehenden Mitteln gibt, um Leistungen und Lernentwicklungen in grossem Umfang zu messen. Gleichfalls gelingt in England eine enge Verknüpfung von Leistungsmessungsergebnissen und Inspektion. Die „National-Curriculum-Tests“ werden jedes Jahr eigens dafür entwickelt, die Leistungen der Schüler am Ende jeder „key stage“ zu überprüfen, um damit Erkenntnisse über die Qualität einzelner Schulen zu erhalten. Die Rechenschaftspflicht liegt damit im wesentlichen bei der Einzelschule, die sich mit der Veröffentlichung ihrer Ergebnisse konfrontiert sieht.

Der Testentwicklungsprozess ist enorm aufwendig, da eine hohe Genauigkeit auch bei den Aussagen zur individuellen und kumulierten Lernentwicklung der Schülerinnen und Schüler erreicht werden soll. Dazu ist es nötig, die Tests so in ihrem Niveau anzupassen, dass in jedem Jahr eine vergleichbare Messung erreicht werden kann. In England geschieht dieses Angleichen ohne Verwendung der Verfahren der Item-Response-Theory[8]. Die Ergebnisse der Tests sind dadurch für Laien etwas leichter zu erfassen, die Verfahren, die verwendet werden müssen, um eine Vergleichbarkeit der Tests über mehrere Jahre herzustellen, sind hingegen sehr viel aufwendiger. Zusätzlich basieren die Tests nicht auf Multiple-Choice-Aufgaben sondern werden von externen „markern“ codiert. Das Verfahren zur Entwicklung und Auswertung der Tests impliziert also äusserst komplexe Prozesse und ist verglichen mit den in anderen Ländern verwendeten Tests sehr kostenintensiv. Im Gegenzug dazu eignen die Tests sich aber auch besonders gut für eine erweiterte Nutzung im Rahmen der Inspektion. Auch mit Hilfe der zur Verfügung gestellten Software RAISEonline können umfangreiche Analysen durchgeführt werden, die im Prozess der externen Evaluation zur Vorbereitung der Inspektionen und zur Bewertung der Schulen hinsichtlich ihrer schülerleistungsbezogenen Qualität herangezogen werden. Die Inspektion geht dabei nach festgelegten Indikatoren vor, die Leistungsstand und Lernentwicklung für die Gesamtschülerschaft sowie für einzelne spezifische Gruppen berücksichtigen.

[8] siehe Erklärung im Anhang A

Niederlande

Auch in den Niederlanden ist die Verbindung zwischen Leistungsmessung und externer Schulevaluation gelungen. Im Gegensatz zum englischen Modell ist der Aufwand, der damit betrieben wird, jedoch etwas geringer. Der CITO-Test setzt sich ausschliesslich aus Multiple-Choice-Aufgaben zusammen, die zwar eine recht komplexe Entwicklungsprozedur verlangen, die Auswertungen jedoch stark vereinfachen. Teilweise werden auch bereits computerbasierte und adaptive Tests verwendet und die Skalierungen werden nach modernen statistischen Verfahren (IRT[9]) vorgenommen. Der Vorteil dieses Vorgehens liegt eindeutig in der Vereinfachung der Prozesse zur Leistungsmessung. Allerdings haben sowohl Multiple-Choice-Aufgaben als auch computerbasierte Tests relativ eng gefasste Formatvorgaben, und können damit auch die Möglichkeiten einschränken, die zur Messung von Kompetenzen und damit auch zur Verarbeitung im Rahmen der externen Evaluation zur Verfügung stehen. Durch das Fehlen längsschnittlicher Testergebnisse, ist trotz der Einbindung von Kontextvariablen auch die Aussagekraft über die Qualität der Schule eingeschränkt. Dazu kommt, dass dem CITO-Test gleichzeitig mehrere Funktionen zugeordnet sind. Einerseits dient er als Übergangszertifikat für die abnehmenden Sekundarschulen, andererseits aber auch als veröffentlichtes Qualitätsmerkmal der Schule. Weiterhin sollen die Ergebnisse des CITO-Tests für ein Systemmonitoring zur Verfügung stehen, Impulse für die Schulentwicklung liefern und Eltern, Lehrpersonen und Schülern als Orientierung für die weitere Leistungsentwicklung dienen. Es ist denkbar, dass der Test nicht allen dieser Funktionen in gleicher Weise gerecht werden kann.

In den Niederlanden besteht aber darüber hinaus die Möglichkeit der Nutzung des Lehrlingvolgsystems, dass explizit die Funktion der Schulentwicklung in den Vordergrund stellt. Auch individuelle Lernentwicklungen können damit nachvollzogen werden. Die Schulinspektion bekommt allerdings in diese Daten nur Einsicht, wenn die Schule dies wünscht.

Ausblick

Die dargestellten Modelle zeigen ein breites Spektrum an Möglichkeiten zum Einsatz von Leistungsmessungen im Rahmen externer Schulevaluationen. Dabei wird deutlich, dass präzise Aussagen über Qualität von Schule im hier betrachteten Sinne mit sehr aufwendigen und kostenspieligen Verfahren erkauft werden müssen. Wenn Testentwicklung, Durchführung und Auswertungsverfahren spezifisch darauf ausgerichtet sind, den auf Schülerleistung ausgerichteten Aspekt der Qualität von Schule zu beleuchten, dann sind die Ergebnisse dieser Leistungsmessung sinnvoll in den Rahmen der externen Evaluation integrierbar. Wenn jedoch die Leistungsmessungen hauptsächlich für einen anderen Kontext entwickelt werden und nur im Sinne eines

[9] siehe Erklärung im Anhang A

Nebenproduktes in die externe Schulevaluation eingehen, dann muss sehr sorgfältig abgewogen werden, welche Aussagen sie überhaupt über die Qualität von Schule zulassen.

5 Besonderheiten der Modelle strukturiert nach Leitfragen

5.1 Schweiz

	Check 5
Was wird gemessen?	Jahrgangsstufe 5: Mathematik, Deutsch, kooperatives Problemlösen, selbstreguliertes Lernen.
Wie wird gemessen?	Papier-und-Bleistift-Tests, gemessen wird Leistung und Geschwindigkeit durch begrenzte Testzeit.
Wer misst?	Lehrpersonen, Didaktiker und Bildungsforscher entwickeln die Aufgaben. Die Lehrpersonen führen die Tests nach Anweisung durch. Codierung und Dateneingabe sowie Datenauswertung an der Universität Zürich
Wie läuft die Durchführung ab?	Einmal pro Jahr im Klassenverband, freiwillige Teilnahme, Lehrpersonen melden ihre Klassen dazu an.
Wie werden die Daten analysiert?	Prozentwerte der gelösten Aufgaben pro Person, Durchschnitt Klasse, Durchschnitt aller Klassen für die Rückmeldung. Zusätzlich Modelle der Item-Response-Theory für die wissenschaftlichen Auswertungen.
Wer erhält Einblick in die Daten und Ergebnisse?	Rückmeldung der Ergebnisse auf Individualebene an die Lehrpersonen, Schülerinnen und Schüler, Bericht an die Eltern. Verantwortliche im BKS und an der Universität Zürich haben Einsicht in den Datensatz.
Bei wem liegt die Rechenschaftspflicht?	Eine Rechenschaftspflicht ist nicht explizit vorgesehen. Der Test ist freiwillig und anonym. Die Ergebnisse werden nicht veröffentlicht.
Wie werden Verknüpfungen von Schülerleistungsdaten zu Schul- und Unterrichtsqualitätsdaten vorgenommen?	Ist nicht vorgesehen, allenfalls bezogen auf Unterrichtsqualität in internen Diskussionen oder individuellen Überlegungen einzelner Lehrpersonen.
Wie fliessen die Daten in den externen Evaluationsprozess bzw. ins schulinterne Qualitätsmanagement ein?	Ist nicht explizit vorgesehen. Keine Verknüpfung mit der bestehenden externen Schulevaluation.
Wie erscheinen Leistungsdaten im Evaluationsbericht?	Ist nicht vorgesehen.

	Check 8
Was wird gemessen?	Jahrgangsstufe 8: Deutsch, Mathematik, Französisch, Englisch.
Wie wird gemessen?	Computerbasierter adaptiver Test, gemessen wird Leistung, keine Geschwindigkeitskomponente.
Wer misst?	Lehrpersonen, Didaktiker und Bildungsforscher entwickeln die Items. Stellwerk organisiert die Auswertung (Automatische Codierung, Dateneingabe und Rückmeldung an Lehrpersonen und Schülerinnen und Schüler). Wissenschaftliche Begleitung durch Universität Zürich
Wie läuft die Durchführung ab?	Einmal pro Jahr im Klassenverband oder auch individuell, z.Z. in Erprobung: Schulen und Klassen melden sich dazu an, ab 2009 voraussichtlich obligatorische Teilnahme für alle Klassen des Jahrgangs im Kanton Aargau.
Wie werden die Daten analysiert?	Modelle der Item-Response-Theory, Schätzung der Personenparameter während des Tests und dementsprechende Anpassung des Tests auf die Personenfähigkeit (adaptives Testen), teilweise selbstkalibrierende Aufgaben. Einzelne Aufgaben können nicht im Klassenverband ausgewertet werden. Die Aufgaben sind nicht öffentlich zugänglich.
Wer erhält Einblick in die Daten und Ergebnisse?	Rückmeldung der Ergebnisse auf Individualebene an die Lehrpersonen, Schülerinnen und Schüler, Bericht an die Eltern. Verantwortliche im BKS und an der Universität Zürich haben Einsicht in die Daten (zukünftig auch offen für die externe Schulevaluation), ab 2009 dient der Test voraussichtlich einer offiziellen Zertifizierung. Die Individualleistung einzelner Schüler wird damit auch für potentielle Arbeitgeber einsichtig.
Bei wem liegt die Rechenschaftspflicht?	Rechenschaftspflicht vor allem beim Schüler/ bei der Schülerin, da die Leistungen zukünftig zertifiziert werden.
Wie werden Verknüpfungen von Schülerleistungsdaten zu Schul- und Unterrichtsqualitätsdaten vorgenommen?	Ist zwar als Funktion vorgesehen aber nicht näher konzeptionell ausgeführt. Durch die relativ abstrakte Rückmeldung schwierig zu verwirklichen. Der Austausch über die Ergebnisse ist intern möglich, aber nicht explizit vorgesehen. Inhaltliche Diskussionen nur über Interpretationshilfen. Es können keine konkreten Testaufgaben analysiert werden. Hohe Abstraktion führt vermutlich zu einer Konzentration auf Parameter statt auf Inhalte.
Wie fliessen die Daten in den externen Evaluationsprozess bzw. ins schulinterne Qualitätsmanagement ein?	Datenfluss zur externen Schulevaluation ist geplant aber bisher noch in den Anfängen (freiwillige Teilnahme). Daten werden vorher für die externe Schulevaluation bereitgestellt und analysiert. Im Rahmen der Evaluation werden die Ergebnisse mit den Lehrpersonen diskutiert und interpretiert.
Wie erscheinen Leistungsdaten im Evaluationsbericht?	Bisher ein kleiner Abschnitt im Anhang des Berichts.

	Orientierungsarbeiten Zentralschweiz
Was wird gemessen?	Jahrgangsstufen 2-9: Deutsch, Mathematik, Mensch und Umwelt, Musik, bildnerisches Gestalten, technisches Gestalten, Naturlehre, Geographie, Hauswirtschaft, ab 2007 Geschichte und Politik, ab 2009 Lebenskunde (einzelne Fachbereiche stehen nur für eine Auswahl von Jahrgangsstufen zur Verfügung.
Wie wird gemessen?	Papier-und-Bleistift-Tests, gemessen wird Leistung. Keine Geschwindigkeitskomponente.
Wer misst?	Lehrpersonen, Didaktiker und Bildungsforscher entwickeln die Items. Die Lehrpersonen (bzw. die Schule) organisieren die Durchführung und Auswertung (Codierung, Bewertung und Rückmeldung an die Schülerinnen und Schüler). Keine wissenschaftliche Begleitung.
Wie läuft die Durchführung ab?	Evtl. mehrmals pro Jahr vor, während oder nach einer Lerneinheit (freiwillige Teilnahme durch Entscheidung der Lehrperson), bzw. im Übertrittsverfahren (obligatorisch). Im Klassenverband oder individuell.
Wie werden die Daten analysiert?	Auszählen der richtigen Antworten. Abgleich mit vorgegebenen Kriterienraster zu den einzelnen Aufgaben.
Wer erhält Einblick in die Daten und Ergebnisse?	Lehrpersonen, Schülerinnen und Schüler, Eltern, evtl. Schulleitung
Bei wem liegt die Rechenschaftspflicht?	Rechenschaftspflicht liegt bei den Schülerinnen und Schülern, da die Leistungen teilweise benotet werden können oder zur Beurteilung der Fähigkeiten am Übertritt herangezogen werden.
Wie werden Verknüpfungen von Schülerleistungsdaten zu Schul- und Unterrichtsqualitätsdaten vorgenommen?	Ist in der Regel nicht vorgesehen, allenfalls bezogen auf Unterrichtsqualität in internen Diskussionen oder individuellen Überlegungen einzelner Lehrpersonen. Wenn die Arbeiten zur Qualitätssicherung der gesamten Schule verwendet werden, können über die Reflexion der Ergebnisse auf Schulebene Verknüpfungen zur Schulqualität hergestellt werden (kein angeleitetes Verfahren).
Wie fliessen die Daten in den externen Evaluationsprozess bzw. ins schulinterne Qualitätsmanagement ein?	Ist nicht vorgesehen.
Wie erscheinen Leistungsdaten im Evaluationsbericht?	Ist nicht vorgesehen.

5.2 Deutschland

	VERA
Was wird gemessen?	Jahrgangsstufe 3 (bisher Jahrgangsstufe 4) Mathematik, Deutsch.
Wie wird gemessen?	Papier-und-Bleistift-Tests, gemessen wird Leistung und Geschwindigkeit durch begrenzte Testzeit.
Wer misst?	Lehrpersonen, Didaktiker und Bildungsforscher entwickeln die Aufgaben. Die Lehrpersonen führen die Tests nach Anweisung durch. Codierung und Dateneingabe durch die Lehrpersonen. Datenauswertung an der Universität Koblenz.
Wie läuft die Durchführung ab?	Einmal pro Jahr im Klassenverband, alle Klassen des Jahrgangs der beteiligten Bundesländer nehmen teil.
Wie werden die Daten analysiert?	Mittelwerte, Prozentwerte gelöster Aufgaben und Vergleich mit Erwartungswerten. Aufgaben werden veröffentlicht und können in der Schule konkret ausgewertet werden.
Wer erhält Einblick in die Daten und Ergebnisse?	Rückmeldung der Ergebnisse auf Individualebene an Schulleitung, Lehrpersonen und Fachkonferenzen. Berichtspflicht der Schulen an die Eltern und Schüler. Verantwortliche im Ministerium und an der Universität Landau haben Einsicht in die Daten.
Bei wem liegt die Rechenschaftspflicht?	Eine Rechenschaftspflicht ist nicht explizit vorgesehen. Die Ergebnisse werden nicht unkumuliert veröffentlicht.
Wie werden Verknüpfungen von Schülerleistungsdaten zu Schul- und Unterrichtsqualitätsdaten vorgenommen?	Ist nicht vorgesehen, allenfalls bezogen auf Unterrichtsqualität in internen Diskussionen oder individuellen Überlegungen einzelner Lehrpersonen.
Wie fliessen die Daten in den externen Evaluationsprozess bzw. ins schulinterne Qualitätsmanagement ein?	Ist nicht explizit vorgesehen.
Wie erscheinen Leistungsdaten im Evaluationsbericht?	Ist nicht vorgesehen.

	Lernstandserhebung in Nordrhein-Westfalen
Was wird gemessen?	Jahrgangsstufe 9: Mathematik, Deutsch, Englisch (jeweils wechselnde Teilkompetenzbereiche).
Wie wird gemessen?	Papier-und-Bleistift-Tests, gemessen wird Leistung und Geschwindigkeit durch begrenzte Testzeit.
Wer misst?	Lehrpersonen, Didaktiker und Bildungsforscher entwickeln die Aufgaben. Die Lehrpersonen führen die Tests nach Anweisung durch. Codierung und Dateneingabe durch die Lehrpersonen. Datenauswertung am Ministerium.
Wie läuft die Durchführung ab?	Einmal pro Jahr im Klassenverband, alle Klassen des Jahrgangs in Nordrhein Westfalen nehmen teil.
Wie werden die Daten analysiert?	Mittelwerte, Prozentwerte gelöster Aufgaben und Vergleich mit Erwartungswerten. Aufgaben werden veröffentlicht und können in der Schule konkret ausgewertet werden.
Wer erhält Einblick in die Daten und Ergebnisse?	Rückmeldung der Ergebnisse auf Individualebene an Schulleitung, Lehrpersonen und Fachkonferenzen. Berichtspflicht der Schulen an die Eltern und Schüler. Verantwortliche im Ministerium und an der Universität Essen haben Einsicht in die Daten.
Bei wem liegt die Rechenschaftspflicht?	Wenn ein Schüler zwischen zwei Noten steht, kann das Ergebnis der Lernstandserhebung neben anderen Kriterien zur Notenfindung herangezogen werden. Schulen haben im positiven Sinn Rechenschaftspflicht, neuerdings werden die 2% besten Schulen öffentlich ausgezeichnet. Die Namen der Schulen werden im Internet bekannt gegeben.
Wie werden Verknüpfungen von Schülerleistungsdaten zu Schul- und Unterrichtsqualitätsdaten vorgenommen?	Ist zunächst nicht vorgesehen, allenfalls bezogen auf Unterrichtsqualität in internen Diskussionen oder individuellen Überlegungen einzelner Lehrpersonen. Für die Verwendung im Rahmen der Inspektion liegen noch keine Erfahrungen vor.
Wie fliessen die Daten in den externen Evaluationsprozess bzw. ins schulinterne Qualitätsmanagement ein?	Ist zunächst nicht explizit vorgesehen. Für die Verwendung im Rahmen der Inspektion liegen noch keine Erfahrungen vor.
Wie erscheinen Leistungsdaten im Evaluationsbericht?	Ist zunächst nicht vorgesehen. Für die Verwendung im Rahmen der Inspektion liegen noch keine Erfahrungen vor.

	Leistungsmessungen und Inspektion, Hamburg
Was wird gemessen?	Vergleichsarbeiten in Jahrgangsstufe 2, 4, 6, 8: Mathematik, Deutsch, 1 Fremdsprache (in Klasse 8 auch 2. Fremdsprache). Wiederholerquoten, Abbrecherquoten und Ergebnisse der Abschlussprüfungen.
Wie wird gemessen?	Für die Vergleichsarbeiten: Papier-und-Bleistift-Tests, gemessen wird Leistung und Geschwindigkeit durch begrenzte Testzeit.
Wer misst?	Für die Vergleichsarbeiten: Lehrpersonen, Didaktiker und Bildungsforscher entwickeln die Aufgaben. Die Lehrpersonen führen die Tests nach Anweisung durch. Codierung und Dateneingabe durch die Lehrpersonen (im Jahr 2007 noch externe Codierung und Dateneingabe). Datenauswertung voraussichtlich in der Behörde für Bildung und Sport.
Wie läuft die Durchführung ab?	Für die Vergleichsarbeiten: Einmal pro Jahr im Klassenverband, alle Klassen der entsprechenden Jahrgänge in Hamburg nehmen teil.
Wie werden die Daten analysiert?	Für die Vergleichsarbeiten: Mittelwerte, Prozentwerte gelöster Aufgaben und Vergleich mit Erwartungswerten aus Vergleichsschulen. Skalierungen nach Item-Response-Theory sind geplant. Aufgaben werden jedes Jahr neu entwickelt. Die alten Arbeiten können in der Schule konkret ausgewertet werden.
Wer erhält Einblick in die Daten und Ergebnisse?	Rückmeldung der Ergebnisse auf Individualebene an Lehrpersonen und Inspektion. Berichtspflicht der Schulen an die Eltern und Schüler. Verantwortliche in der Behörde für Jugend und Sport haben Einsicht in die Daten.
Bei wem liegt die Rechenschaftspflicht?	Die Vergleichsarbeiten werden wie Klassenarbeiten benotet. Die Rechenschaftspflicht liegt bei den Schülerinnen und Schülern. Auf der Ebene der Inspektion werden die Leistungsdaten erneut beurteilt. Der Bericht der Inspektion geht parallel an die Schulen und die Schulaufsicht. Auf dieser Ebene trägt die Schule Rechenschaft gegenüber der Schulaufsicht.
Wie werden Verknüpfungen von Schülerleistungsdaten zu Schul- und Unterrichtsqualitätsdaten vorgenommen?	Durch die Einbindung von Vergleichsarbeiten, Abschlussprüfungen, Abbrecher- und Wiederholerquoten im Rahmen der Inspektion. Das genauere Vorgehen ist noch nicht bekannt.
Wie fliessen die Daten in den externen Evaluationsprozess bzw. ins schulinterne Qualitätsmanagement ein?	Anschliessend an die Inspektion soll im Austausch zwischen den Lehrpersonen und unter Beratung durch die Schulaufsicht eine inhaltliche Analyse der Aufgaben vorgenommen werden und spezifische didaktische Materialien eingesetzt werden. Gewünscht wird, dass im Zuge der stärkeren Autonomie der Einzelschule die Initiative von den Schulen unter Anleitung und Hilfestellung der Schulaufsicht ausgeht.
Wie erscheinen Leistungsdaten im Evaluationsbericht?	Die Leistungsdaten werden einen Platz im Bericht haben, explizite Beispiele liegen jedoch noch nicht vor.

5.3 England

	National Curriculum Assessments in England
Was wird gemessen?	Am Ende jeder "key stage". (Alterstufen 7, 11, 14 und 16). Englisch, Mathematik, Naturwissenschaften. In "key stage" 3 zusätzliche Fächer wie Kunst, Design, Geographie, ab 2008 voraussichtlich auch ICT.
Wie wird gemessen?	In Mathematik und Englisch sowie Naturwissenschaften (nur "key stages" 2 und 3) Papier-und-Bleistift-Tests, gemessen wird Leistung und Geschwindigkeit durch begrenzte Testzeit. Zusätzlich Einschätzungen der Lehrpersonen über die Periode des jeweiligen "key stage" in allen genannten Fächern.
Wer misst?	Lehrpersonen, Didaktiker und Bildungsforscher entwickeln die Aufgaben. Die Lehrpersonen führen die Tests bzw. nach Anweisung durch, bzw. nehmen Einschätzungen vor. Codierung und Dateneingabe der Tests durch professionelle "marker". Datenauswertung durch das Ministerium (DfES) sowie durch die Inspektion (Ofsted). Weitere Auswertungen werden vom Fischer Familiy Trust durchgeführt.
Wie läuft die Durchführung ab?	Im Mai jeden Jahres für die Schüler am Ende der "key stage". Alle Schüler dieser Jahrgänge nehmen im Klassenverband daran teil. Eine Ausnahme stellt der GCSE am Ende der "key stage" 4 dar. Die Teilnahme ist freiwillig, obwohl alle Schülerinnen und Schüler angehalten sind, daran teilzunehmen.
Wie werden die Daten analysiert?	Für die unterschiedlichen Antworten in den Tests werden Punkte vergeben und ausgezählt. Berichtet werden Mittelwerte, Prozentwerte gelöster Aufgaben und Vergleich mit Erwartungswerten sowie Value-Added-Werte, die die Differenzen zwischen den Altersstufen angeben. Lehrpersoneneinschätzungen werden auf der Grundlage von Niveaubeschreibungen vorgenommen. Den Lernenden wird dementsprechend eine Niveaustufe zugeteilt.
Wer erhält Einblick in die Daten und Ergebnisse?	In "key stage" 1 werden die Daten von den Lehrpersonen an die "Local Authorities" gesendet, die diese wiederum an das Ministerium weiterleiten. Auch Ofsted und der Fischer Family Trust können die Daten weiterverarbeiten. Zusätzlich Bericht über Ergebnisse auf Individualebene an Schülerinnen und Schüler und Eltern. Auf Schulebene kumulierte Ergebnisse werden veröffentlicht. In "key stage" 2 und 3 erhalten die Schulen die Rückmeldung von den "markern", sonst wie "key stage" 1.
Bei wem liegt die Rechenschaftspflicht?	Die Rechenschaftspflicht liegt bei den Schulen. Die auf Schulebene kumulierten Daten werden öffentlich zur Verfügung gestellt. Davon wird vor allem die Schulwahl durch Schüler und Eltern beeinflusst. Darüber hinaus gehen die Leistungsdaten in die Beurteilung der Inspektion ein, die ihre Berichte ebenfalls veröffentlicht und spezifische Massnahmen für mangelhaft bewertete Schulen verordnen kann.
Wie werden Verknüpfungen von Schülerleistungsdaten zu Schul- und Unterrichtsqualitätsdaten vorgenommen?	Sowohl im Rahmen der Inspektion als auch im Rahmen der Selbstevaluation werden bei der Analyse der Ursachen und Auswirkungen von Schülerleistungen sowie der getroffenen Massnahmen die Leistungen mit anderen Schulqualitätsdaten in Beziehung gesetzt.

Wie fliessen die Daten in den externen Evaluationsprozess bzw. ins schulinterne Qualitätsmanagement ein?	Die Inspektion bereitet sich auf der Grundlage der Ergebnisse aus den National Curriculum Assessments vor und richtet die Inspektionsbesuche danach aus. Die Leistungen der Schülerinnen und Schüler sind ein wichtiges Qualitätskriterium dabei. Ergebnisse aus den Leistungsmessungen liegen darüber hinaus im Fokus in der internen Selbstevaluation der Schule, die wiederum durch die externe Inspektion evaluiert wird.
Wie erscheinen Leistungsdaten im Evaluationsbericht?	Die Leistungen der Schülerinnen und Schüler sind insbesondere im Abschnitt "Achievement and Standards" des Inspektionsberichts behandelt. Darüber hinaus finden sich Hinweise auf Leistungsaspekte auch in anderen Abschnitten, in denen eine Beziehung zwischen Schulqualität und Schülerleistung hergestellt wird. Speziell im Abschnitt "Overall Effectiveness" werden Hinweise auf die Schülerleistung gegeben.

5.4 Niederlande

	CITO-Test
Was wird gemessen?	Am Ende der Primarschule (Jahrgangsstufe 8) für Niederländisch, Mathematik, Problemlösen, Weltkunde.
Wie wird gemessen?	In der Regel Papier-und-Bleistift-Tests, 10-15% der Schülerinnen und Schüler bearbeiten ihren Test bereits am Computer. Gemessen wird Leistung und Geschwindigkeit durch begrenzte Testzeit.
Wer misst?	Lehrpersonen, Didaktiker und Bildungsforscher entwickeln die Aufgaben. Die Lehrpersonen führen die Tests nach Anweisung durch. Codierung und Dateneingabe sowie Datenauswertung am CITO.
Wie läuft die Durchführung ab?	Im Februar jeden Jahres für Schüler am Ende der Primarstufe, deren Schulen sich für den CITO-Test entschieden haben (ca. 85 Prozent). Der Test wird im Klassenverband durchgeführt.
Wie werden die Daten analysiert?	Die Multiple Choice-Aufgaben werden automatisch codiert und mit Hilfe der Item-Response-Theory skaliert. Berichtet werden Anzahl und Prozentwerte gelöster Aufgaben sowie der Skalenwert (individuell und kumulativ). Zusätzlich wird in der individuellen Rückmeldung die hypothetische Position im Prozentrangband der einzelnen weiterführenden Schulformen berichtet.
Wer erhält Einblick in die Daten und Ergebnisse?	CITO wertet die Daten aus und nutzt sie wissenschaftlich. Rückmeldung der Ergebnisse auf Individualebene an die Lehrpersonen, Schülerinnen und Schüler, Bericht an die Eltern. Das Inspektorat hat Einsicht in die Ergebnisse. Kumulierte Ergebnisse der einzelnen Schulen werden regelmässig in der Presse veröffentlicht.
Bei wem liegt die Rechenschaftspflicht?	Die Rechenschaftspflicht liegt teilweise bei den Schulen, da die Ergebnisse veröffentlicht werden, und die Eltern und Schüler sich in ihrer Schulwahl daran orientieren. Weiterhin orientiert sich auch die Inspektion daran. Teils liegt sie auch bei den Schülerinnen und Schülern, die aus den Ergebnissen eine Orientierung für die Wahl der weiterführenden Schulform erhalten.
Wie werden Verknüpfungen von Schülerleistungsdaten zu Schul- und Unterrichtsqualitätsdaten vorgenommen?	Im Rahmen der Inspektion werden die Leistungsdaten als Teilbereich der Qualitätsindikatoren betrachtet.
Wie fliessen die Daten in den externen Evaluationsprozess bzw. ins schulinterne Qualitätsmanagement ein?	Die Inspektion bereitet sich aufgrund der Testergebnisse auf den Besuch vor und richtet ihn danach aus. Die Leistungen der Schülerinnen und Schüler sind ein wichtiges Qualitätskriterium dabei (Der Umfang der Inspektion hängt davon ab).
Wie erscheinen Leistungsdaten im Evaluationsbericht?	Die Leistungen der Schülerinnen und Schüler sind insbesondere im Zusammenhang mit den Indikatoren zu Lernergebnissen (12.1) und Lernentwicklung (13.1 und 13.2) behandelt. Darüber hinaus finden sich Hinweise auf Leistungsaspekte auch in anderen Abschnitten, in denen eine Beziehung zwischen Schulqualität und Schülerleistung hergestellt wird.

	Leerlingsvolgsystem
Was wird gemessen?	Jedes halbe Jahr in der Primarschule in Sprache, Lesen, Rechnen, Mathematik und Weltorientierung.
Wie wird gemessen?	In der Regel Papier-und-Bleistift-Tests, teilweise am Computer. Gemessen wird Leistung und Geschwindigkeit durch begrenzte Testzeit.Teilweise Anwendung adaptiver Verfahren.
Wer misst?	Lehrpersonen, Didaktiker und Bildungsforscher entwickeln die Aufgaben. Die Lehrpersonen führen die Tests nach Anweisung durch. Codierung und Dateneingabe an der Schule, Datenauswertung automatisch durch spezifische Software.
Wie läuft die Durchführung ab?	im Klassenverband
Wie werden die Daten analysiert?	Automatische Codierung, Skalierung mit Item-Response-Theory. Vergleiche mit anderen Testzeitpunkten sind auf der gleichen Metrik möglich. Auch landesweite Vergleiche. Inhaltliche Interpretation möglich, da die Tests den Schulen direkt zur Verfügung stehen.
Wer erhält Einblick in die Daten und Ergebnisse?	Die Schulen werten die Daten mit Hilfe der Software selbst aus und nutzen sie für die Unterrichtsentwicklung. Rückmeldung der Ergebnisse auf Individualebene an die Schülerinnen und Schüler. Das Inspektorat hat in der Regel keine Einsicht in die Ergebnisse. Die Ergebnisse werden nicht veröffentlich sondern verbleiben in der Schule.
Bei wem liegt die Rechenschaftspflicht?	Eine Rechenschaftspflicht ist nicht explizit vorgesehen. Der Test ist freiwillig und anonym. Die Ergebnisse werden nicht veröffentlicht.
Wie werden Verknüpfungen von Schülerleistungsdaten zu Schul- und Unterrichtsqualitätsdaten vorgenommen?	Intern soll im Sinne der Qualitätsindikatoren für Schulen eine Verbindung hergestellt werden, die sich in Schul und Unterrichtsentwicklung niederschlägt. Extern wird keine Verbindung hergestellt.
Wie fliessen die Daten in den externen Evaluationsprozess bzw. ins schulinterne Qualitätsmanagement ein?	Ist nicht vorgesehen.
Wie erscheinen Leistungsdaten im Evaluationsbericht?	Ist nicht vorgesehen.

Literatur

[1] van Ackeren, I. (2003) Evaluation, Rückmeldung und Schulentwicklung: Erfahrungen mit zentralen Tests, Prüfungen und Inspektionen in England, Frankreich und den Niederlanden. Münster: Waxmann.

[2] Atkinson et al. (in Vorbereitung). Impact of Section 5 Inspections: Maintained Schools in England.

[3] Burkard C. *Zentrale Lernstandserhebungen in Klasse 9. Landesweite Ergebnisse 2005*
http://learnline.de/angebote/lernstand9/download/ergebn_05/lse-ergebnisse_2005.pdf.

[4] Dobbelstein P. & Peek, R. *Von der Bestandsaufnahme zur Förderung: Diagnostische Potenziale von Lernstandserhebungen und die Verbindung zur gezielten Förderung von Schülerinnen und Schülern.* Forum Schule, (1), 24-25.

[5] Flaugher R. *Item Pools* H. Wainer (Hrsg.) Computerized Adaptive Testing: A Primer. Mahwah, New Jersey: Lawrence Erlbaum. 37-59

[6] Grubb, N. W. (1999) *Improvement or control? A US view of english inspection.* C. Cullingford (Hrsg.) An inspector calls. Ofsted and its effect on school standards. London: Routledge Falmer. 70-96

[7] Inspectie van het Onderwijs (2005). 2005 Supervisory Framework for Primary Education: Content and working method of the inspection supervision.
www.onderwijsinspectie.nl/
Documents/pdf/EngelsToezichtkaderpo05.pdf

[8] Institut für Qualitätsentwicklung im Bildungswesen (2005) Perspektiven und Visionen. Die Normierung und Präzisierung der Bildungsstandards in den Ländern der Bundesrepublik Deutschland. Das IQB stellt sich vor. Berlin: IQB.

[9] Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E. & Vollmer, H.J. (2003) Zur Entwicklung nationaler Bildungsstandards: Eine Expertise. Berlin: Bundesministerium für Bildung und Forschung.

[10] Koch, U., Groß Ophoff, J., Hosenfeld, I. & Helmke, A. (2006). *Von der Evaluation zur Schul- und Unterrichtsentwicklung - Ergebnisse der Lehrerbefragungen zur Auseinandersetzung mit den VERA-Rückmeldungen.* F. Eder, A. Gastager und F. Hofmann (Hrsg.) Qualitäts durch Standards? Tagungsband zur 68. Tagung der Arbeitsgruppe der Empirischen Bildungsforschung (AEPF), Salzburg.

[11] Kohler, B. & Schrader, F.-W. (2004). *Ergebnisrückmeldung und Rezeption: Von der externen Evaluation zur Entwicklung von Schule und Unterricht.* Empirische Pädagogik, 18 (1), 3-17.

[12] Lehmann, R (2001). *Systembeobachtung: Lernausgangslage und Lernentwicklung in der Sekundarstufe I.* Klaus-Jürgen Tillmann und Witlof Vollstädt (Hrsg.): Politikberatung durch Bildungsforschung. Das Beispiel: Schulentwicklung in Hamburg. Opladen.

[13] Peek, R. *Qualitätsuntersuchung an Schulen zum Unterricht in Mathematik (QuaSum)- Klassenbezogene Ergebnisrückmeldung und ihre Rezeption in Brandenburger Schulen.* Empirische Pädagogik, 18 (1), 82-114.