WEITERE BEITRÄGE

DDS – Die Deutsche Schule
102. Jahrgang 2010, Heft 1, S. 60-69

Uwe Maier

Vergleichsarbeiten im Spannungsfeld zwischen formativer und summativer Leistungsmessung

Zusammenfassung
In diesem Beitrag werden Vergleichsarbeiten vor dem Hintergrund der Unterscheidung zwischen formativer und summativer Leistungsmessung analysiert. Obwohl dies intendiert ist, haben Vergleichsarbeiten in Deutschland nur wenig mit formativer Leistungsmessung gemeinsam. Datenbasierte Unterrichtsentwicklung wird damit eher blockiert. Der Beitrag schließt mit einem Ausblick auf Länder, in denen an einer sinnvollen Kombination externer Tests und formativer Leistungsmessung gearbeitet wird.
Schlüsselwörter: Vergleichsarbeiten, formative Leistungsmessung, summative Leistungsmessung, Evaluation, datenbasierte Unterrichtsentwicklung, Rückmeldungen

Mandatory Testing in the Field of Controversy Between Formative and Summative Assessment

Abstract
This paper aims at analyzing summative and formative functions of mandatory school performance tests. The analysis reveals that mandatory testing, which has been implemented by the German states in the last 5 years, can be attributed as summative assessment. As a consequence, mandatory testing provides only little fine-grained feedback information for data-based school improvement. The last section of this paper discusses international examples how to combine summative and formative functions of mandatory testing systems effectively.
Keywords: mandatory testing, formative assessment, summative assessment, evaluation, data based school improvement, school performance feedback

1. Funktionen von Vergleichsarbeiten

Bildungsstandards und Vergleichsarbeiten zielen darauf ab, das Schulleistungsniveau insgesamt zu heben und die soziale Selektivität des Schulsystems zu reduzieren. Vergleichsarbeiten und Testrückmeldungen werden dabei als wichtiger „Hebel" betrachtet. Sie sollen einerseits zur Qualitätssicherung auf Systemebene beitragen. Andererseits verspricht man sich konkrete Entwicklungsimpulse für den Unterricht und die datenbasierte Einzelschulentwicklung (vgl. Peek/Steffens/Köller 2006). Durch diese Doppelfunktion des Prüfens einerseits und des Innovierens andererseits besteht allerdings die Gefahr einer Funktionsüberfrachtung (vgl. Altrichter/Heinrich 2006; Kuper 2006; Kühle/Peek 2007; Posch 2009).

Besonders beeindruckend und ambitioniert ist beispielsweise die Liste an Zielen, die man mit den Lernstandserhebungen in NRW realisieren möchte: Standardüberprüfung und Qualitätssicherung; Feststellung von Lern- und Förderbedarf für Schülergruppen und eingeschränkt für einzelne Schülerinnen und Schüler; Stärkung der diagnostischen Kompetenz von Lehrkräften; Orientierungshilfe bei der Leistungsbewertung; Weiterentwicklung des Unterrichts in den Schulen; Unterstützung der Umsetzung der neuen Kernlehrpläne; Identifikation von Schulen mit möglicherweise unbefriedigender Wirksamkeit im Hinblick auf Unterstützungsnotwendigkeiten; Bereitstellung von ergänzenden Informationen für das Systemmonitoring.[1]

Die Spannbreite der Ziele reicht vom Systemmonitoring über die Evaluation der Wirksamkeit von Schulen (Prüfen) bis hin zur konkreten Diagnose von Schülerlernständen und der Ableitung von Förderbedarf (Innovieren). Auch in Vergleichsarbeitsprojekten mit einer weniger elaborierten Rückmeldestruktur, wie z.B. in Baden-Württemberg, findet man diese doppelte Zielsetzung:[2] „(1) Die Vergleichsarbeiten helfen zu überprüfen, inwieweit es den Schulen gelungen ist, die Erwartungen der baden-württembergischen Bildungsstandards zu erreichen (…). (2) Die Vergleichsarbeiten vermitteln den Lehrkräften, den Schülerinnen und Schülern sowie deren Eltern objektive Informationen über den Lernstand im Hinblick auf bestimmte Schwerpunktbereiche der Bildungsstandards. Die Informationen zu den Lernvoraussetzungen der Klasse und einzelner Schülerinnen und Schüler, die zu Beginn eines neuen Bildungsabschnittes festgestellt werden, können bei der gezielten Planung des weiteren Unterrichts helfen."

Ziel des Beitrags ist die Analyse dieser Doppelfunktion vor dem theoretischen Hintergrund der Unterscheidung zwischen formativen und summativen Aspekten

1 Vgl. URL: http://www.standardsicherung.schulministerium.nrw.de/lernstand8/; Zugriffsdatum: 24.11.2009.

2 Vgl. URL: http://www.schule-bw.de/entwicklung/dva/vadva/konzeption-dva/ziele; Zugriffsdatum: 22.12.2009.

von Leistungsmessungen.[3] Dabei werden keine Originaldaten präsentiert; vielmehr wird auf empirische Befunde zu formativer Leistungsmessung Bezug genommen. Der Beitrag möchte eine programmatische Diskussion darüber anregen, wie Vergleichsarbeiten als wichtige Instrumente einer datengestützten Schulentwicklung optimiert werden können.

2. Formative und summative Aspekte von Leistungsmessungen

Zunächst einmal ist zu fragen, worin der Unterschied zwischen formativer und summativer Leistungsmessung liegt. Wichtig für diese Differenzierung ist, dass es sich nicht um verschiedene Typen von Schulleistungstests handelt. Vielmehr ist die Verwendung der Testergebnisse von entscheidender Bedeutung (vgl. Harlen 2005). Es kommt darauf an, für welche Entscheidungen die Evaluationsinformationen genutzt werden. Wird ein bestimmtes Programm bewertet (summativ) oder steht die Optimierung eines noch laufenden Programms im Vordergrund (formativ)?

Eine weitere Differenz ist der Grad der Beteiligung verschiedener Akteure am Evaluationsprozess und an der Datennutzung (vgl. Black/Wiliam 1998b). Eine summative Leistungsmessung hat zum Ziel, anhand extern festgelegter Kriterien und Testaufgaben die Lernerträge nach bestimmten Bildungsabschnitten möglichst fair und vergleichend zu erfassen. Die Rückmeldungen werden in der Regel zur Kenntnis genommen und für weiterführende Ausbildungsabschnitte als Berechtigung genutzt. Dagegen können Rückmeldedaten aus formativen Leistungsmessungen nur dann sinnvoll für die Optimierung nachfolgender Lehr- und Lernprozesse genutzt werden, wenn Lehrer/innen, Schüler/innen und eventuell auch die Eltern in einem möglichst hohen Maße an der Zielsetzung, Planung und Auswertung der Leistungsmessung beteiligt werden. Nur wenn alle Beteiligten die Zusammenhänge zwischen Unterricht, Lernanstrengung, Testergebnis und möglichen Konsequenzen erkennen und auch akzeptieren können, ist eine Leistungsmessung formativ, d.h. kann den nachfolgenden Lehr-Lernprozess „informieren".

Die Unterscheidung zwischen formativer und summativer Leistungsmessung ist von besonderer Relevanz, weil in den letzten Jahrzehnten die überragende Bedeutung formativer Leistungsmessung für den Unterrichtserfolg vielfach empirisch bestätigt wurde (vgl. Wiliam u.a. 2005). Ein Meilenstein hierfür ist die immer wieder zitierte Metaanalyse zu Effekten formativer Leistungsmessung von Black und Wiliam (1998a). Beispielsweise wurden Lehrkräfte in speziellen Programmen geschult, wie sie Schüler/inne/n informative Rückmeldungen bereits nach kleinen Lernabschnitten geben können. Diese Implementation formativer Rückmeldestrukturen in den Unterricht

3 Im englischsprachigen Raum findet man analog hierzu die Konzepte „summative vs. formative assessment" bzw. „assessment of learning vs. assessment for learning".

führte zu substanziellen Lernzuwächsen im Vergleich zu Kontrollgruppen. Ebenso liefert die Lehr-Lernforschung eine Fülle von Hinweisen, dass Rückmeldungen im Unterricht in Kombination mit konkreten Lernhinweisen als eine der effektivsten Unterrichtsinterventionen überhaupt gelten (vgl. Fraser u.a. 1987; Hattie/Timperley 2007).

3. Formative und summative Aspekte von Vergleichsarbeiten

Die empirisch gut belegten Erfolge formativer Leistungsmessung werfen die generelle Frage auf, wie im deutschen Schulsystem die Praxis der Leistungsmessung so reformiert werden kann, dass Schüler/innen, Lehrer/innen und auch Eltern mehr qualitativ hochwertige, informative Rückmeldungen über den Lernfortschritt erhalten. Wie bereits erläutert, ist ein wichtiges Ziel von Vergleichsarbeiten, eine informative Rückmeldekultur an den Schulen zu stärken bzw. zu etablieren. Doch inwiefern entsprechen Vergleichsarbeiten dem, was in der internationalen Literatur unter formativer Leistungsmessung verstanden wird? Thesenartig sollen im Folgenden einzelne Merkmale der in Deutschland implementierten Vergleichsarbeiten vor dem Hintergrund der Unterscheidung zwischen formativer und summativer Leistungsmessung analysiert werden.

Der summative Charakter der Vergleichsarbeiten wird durch folgende konzeptionelle Merkmale gestärkt:

- Vergleichsarbeiten werden im Rahmen der KMK-Gesamtstrategie zum Bildungsmonitoring entwickelt. Damit stehen sie in einer Reihe mit weiteren bilanzierenden Rückmeldeinstrumenten, wie dem Bildungsbericht oder den internationalen *Large Scale*-Studien.
- Die Testadministration in den einzelnen Bundesländern ist zentral und Teil der schulbürokratischen Strukturen. Damit haben Vergleichsarbeiten ihren Ursprung nicht in der professionellen Diskussion, sondern gelten an Schulen als bildungspolitischer Reflex. Trotz Deregulierungs- und Autonomierhetorik in den letzten beiden Jahrzehnten werden Vergleichsarbeiten genauso wie beispielsweise zentrale Abiturprüfungen vorgegeben und durchgeführt.
- Die Testzeitpunkte von Vergleichsarbeiten sind festgelegt und die Leistungsmessung umfasst einen mehrjährigen Bildungsabschnitt. Damit geht es eindeutig um Bilanzierung. Selbst wenn Lehrkräfte gewillt sind, aus den Befunden etwas zu lernen, stellt sich die Frage nach dem Zeitpunkt der Umsetzung möglicher Konsequenzen. In einigen Bundesländern wurde aufgrund dieser Erwägungen der Testzeitpunkt verschoben: in NRW von Lernstand 9 auf 8; in Baden-Württemberg von Ende Klasse 6 und 8 auf Anfang Klasse 7 und 9; bei VERA von Klasse 4 auf 3; in Thüringen von Schuljahresende auf Mai. Es bleibt allerdings bei dem Versuch, den formativen Charakter der zentralen Tests zu stärken. Lehrkräfte können bei gutem Willen auf die Rückmeldedaten reagieren. Allerdings bleibt es bei der

Testung größerer Bildungsabschnitte, grundlegender Kompetenzen und der administrativen Vorgabe des Testzeitpunkts.

- Grundlage für die Aufgabenentwicklung sind die Standards der Länder und mit der Übernahme der Aufgabenentwicklung durch das Institut zur Qualitätsentwicklung im Bildungswesen (IQB) jetzt ganz dezidiert die nationalen Bildungsstandards der KMK. Es geht um Standardsicherung, d.h. summative, externe Qualitätsprüfung. Die Kluft zwischen diesen Standards und standardisierten Tests und dem, was Lehrkräfte tatsächlich unterrichten, ist groß. Damit wird es für Lehrkräfte sehr schwer sein, die Testresultate mit dem gehaltenen Unterricht und den eigenen Schülerbewertungen in Verbindung zu setzen.
- Die Aufgabenformate der Vergleichsarbeiten sind in der Regel geschlossen. Bei halboffenen oder offenen Aufgaben müssen klare Bewertungskriterien vorgegeben werden. Vergleichsarbeitsaufgaben müssen überwiegend testtheoretischen Kriterien genügen. Es wird zwar immer betont, dass man gerade mit zentralen Tests auch fachdidaktische Innovationen im Bereich Aufgabenkultur transportieren möchte. Man bleibt jedoch weit hinter den realen Möglichkeiten zurück.
- Die Auswertung von Vergleichsarbeiten ist produktorientiert. Die Lösungsprozesse sind weitgehend irrelevant. Gerade Lösungsversuche oder Lösungsprozesse sind jedoch für Lehrkräfte ein wichtiger Hinweis, wo Defizite bei den Schüler/inne/n liegen und an welcher Stelle Förderung ansetzen muss. In einigen Vergleichsarbeitsprojekten werden Aufgabenkommentare mit Fehlerschwerpunkten zur Verfügung gestellt. Diese Materialien erhöhen den Informationsgehalt der Rückmeldung für Lehrkräfte und Schüler/innen (z.B. Büchter/Leuders 2005).
- Die Bezugsnorm ist in den wissenschaftlich führenden Vergleichsarbeitsprojekten (VERA, Lernstand, Kompetenztests etc.) eine kriteriale. Es wird nach Kompetenzbereichen differenziert, und die Schülerleistungen werden in Kompetenzstufen angegeben. Dies ist ein wichtiger Schritt in Richtung formative Leistungsmessung. Gleichzeitig wird aber die soziale Bezugsnorm gepflegt und durch die Berechnung von „fairen Vergleichen" technisch hochgerüstet. Es besteht die Gefahr, dass diese sozialen Referenzwerte die Interpretation der Daten dominieren. In Kollegien wird dann zuerst auf den „adjustierten Landeswert" und erst danach auf den kriterialen Bezugswert geschaut.
- Durch die Methoden der psychometrischen Testkonstruktion werden instruktionssensitive Aufgaben systematisch ausgeschlossen. Testkonstrukteur/inn/e/n erhöhen die Reliabilität durch die Auswahl von Aufgaben, die gut zwischen den Testpersonen diskriminieren. Aufgaben, die alle Schüler/innen korrekt beantworten oder nicht beantworten, werden ausgeschlossen. Es ist jedoch sehr wahrscheinlich, dass gerade diese Aufgaben messen, was im Unterricht behandelt bzw. nicht behandelt wurde.

Die Analyse der Gestaltungselemente führt zu der These, dass der summative Charakter von Vergleichsarbeiten überwiegt. Eine Rückwirkung auf die Optimierung von Lehr-Lernprozessen wäre damit stark in Frage gestellt. Die deutsche Variante

testbasierter Rechenschaftslegung scheint so konzipiert zu sein, dass die in den Zielen proklamierten Vorteile einer formativen Nutzung von Rückmeldedaten praktisch kaum realisierbar sind. Dies wird sich auch mit der Umstellung auf VERA 8 in den meisten Bundesländern nicht grundsätzlich ändern. Lediglich Aufgabenentwicklung und Pilotierung werden vom IQB übernommen. Die Länder bleiben in der Regel bei den bisher entwickelten und implementierten Konzepten der Testdurchführung und Ergebnisrückmeldung. Diese Argumente lassen sich auch durch erste Befunde der neueren Rezeptionsforschung bestätigen. In diesen Studien gibt es Hinweise, dass Vergleichsarbeiten nur in einem eher bescheidenen Umfang für die objektive Diagnose von Schülerleistungen und die Reflexion über Unterricht genutzt werden (vgl. z.B. Sill/Sikora 2007; Nachtigall/Jantwoski 2007; Kuper/Hartung 2007; Maier 2007; Maier 2008).

Erschwerend kommt hinzu, dass viele Lehrkräfte und Schulen nicht auf die Nutzung von Leistungsrückmeldungen vorbereitet werden. Posch (2009) argumentiert beispielsweise, dass Rückmeldungen aus Lernstandserhebungen nicht zu einer Veränderung von Unterricht führen, weil Evaluation und datenbasierte Schulentwicklung nicht zur Arbeitskultur von Lehrkräften gehören. Testrückmeldungen werden dann als Kontrolle, d.h. weitestgehend summativ interpretiert.

4. Möglichkeiten der Verknüpfung von Rechenschaftslegung und datenbasierter Unterrichtsentwicklung

Um die formative Nutzung von Rückmeldedaten für die Unterrichtsentwicklung und die Förderung einzelner Schüler/innen zu verbessern, sind theoretisch verschiedene Strategien denkbar. Ein Ansatz wäre, die Bereitschaft zu kritischer Evaluation und Veränderung bei Schulen und Lehrer/inne/n zu fördern. Posch (2009) weist auf strukturelle (Regulationen und Ressourcen) und individuelle (Werthaltungen und Kompetenzen) Voraussetzungen für die Nutzung von Evaluationsdaten hin. Wichtige Maßnahmen zur Erhöhung der Bereitschaft, externe Evaluationsdaten zu nutzen, sind nach Posch: Erweiterung der Schulautonomie, Verpflichtung der Schulen zur Evaluation und Dokumentation von Entwicklungsschritten, Aufbau professioneller Lerngemeinschaften oder die Einrichtung einer Unterstützungsstruktur auf regionaler Ebene. Eine zweite Strategie wäre, die Vergleichsarbeitssysteme so zu konzipieren, dass Lehrkräften und Schulen deutlich wird, dass Testrückmeldungen einen evaluativ-formativen Charakter haben und zu Maßnahmen der Optimierung von Unterricht und der Förderung einzelner Schüler/innen führen können. In diesem Abschnitt sollen Hinweise gegeben werden, welche Richtung eine Weiterentwicklung von Testsystemen im deutschsprachigen Raum nehmen könnte.

Interessanterweise werden gerade in den Ländern, die unter den Konsequenzen von *high-stakes testing* zu leiden haben, Reformmodelle erprobt, die auch für Deutschland wertvolle Entwicklungsimpulse liefern könnten. Diese Beispiele zeigen, dass eine sinnvolle Verknüpfung von formativer und summativer Leistungsmessung im Rahmen testbasierter Rechenschaftslegung durchaus möglich ist. Allerdings erfordert dies ein grundlegendes Umdenken gemäß den folgenden Prinzipien zur Konstruktion instruktionssensitiver Leistungsmessungen (vgl. z.B. Sloane/Kelly 2003; Herman 2004; Harlen 2005; Birenbaum u.a. 2006):

- Die Testzeitpunkte müssen so gewählt werden, dass Lehrer/innen, Schüler/innen und Eltern mit den Rückmeldedaten weiterarbeiten können. Idealerweise gibt es mehrere Testzeitpunkte, vor allem zu Beginn eines Bildungsabschnitts. Dies schließt einen summativen Test am Ende eines Bildungsabschnitts nicht aus.
- Die Rückmeldeformate müssen so konzipiert sein, dass auf Klassen- und Schülerebene didaktisch relevante Informationen zur Verfügung stehen. Gleichzeitig können auf höheren Ebenen aggregierte Daten die Leistungsfähigkeit von Schulen oder Schuldistrikten nach außen demonstrieren. Negative Effekte können jedoch nur durch eine faire, mehrperspektivische und auf Entwicklung bezogene (ipsative) Form der Leistungsevaluation verhindert werden.
- Lehrer/innen müssen an der Auswahl bestimmter Testkomponenten beteiligt werden. Nur damit lässt sich ein auf die Unterrichtssituation abgestimmtes Testverfahren realisieren.
- Schüler/innen und Eltern können nur dann von zentralen Testsystemen profitieren, wenn sie in die Planung und Interpretation der Testergebnisse mit einbezogen werden. Ziel dieses Prozesses sollte die Stärkung der Selbsteinschätzungsfähigkeit bei Schüler/inne/n sein.
- Leistungsindikatoren müssen mehrperspektivisch angelegt sein. Es muss möglich sein, dass durch zentrale, externe Tests ein differenziertes Bild der Schülerleistungen gezeichnet wird, z.B. durch Kombination von performanzorientierten Tests, Portfolios und Leistungsbewertung durch Lehrkräfte.
- Die mit zentraler Leistungsmessung verknüpften Sanktionen müssen auf ein Mindestmaß reduziert werden.

Ein Alternativkonzept, das diesen Forderungen genügen möchte, wird von Birenbaum und Kollegen (2006) vorgeschlagen. Die Autoren des Positionspapiers fordern einen grundlegenden Paradigmenwechsel im Schulsystem von *assessment of learning* zu *assessment for learning*. Testbasierte Schulreformen müssten hierzu auf der Basis von *Integrated Assessment Systems* (IAS) weiterentwickelt werden. Kernstück von IAS ist eine Integration summativer und formativer Aspekte zentraler Tests mit Synergieeffekten. IAS sollten so konzipiert sein, dass für Lehrer/innen und Schüler/innen der formative Aspekt zentraler Leistungsmessungen dominiert. Damit können sich Lehrer/innen und Schüler/innen wieder auf das Curriculum und den Unterricht konzentrieren anstatt auf den Test. Die für das Bildungsmonitoring relevanten Daten werden auf institutioneller Ebene aggregiert und können nicht mehr direkt mit ein-

zelnen Lehrkräften in Verbindung gebracht werden. Das Fundament von IAS ist ein komplexer Lernbegriff. Dies impliziert alternative Formen der Leistungsmessung, eine mehrperspektivische Erfassung von Schulleistung, eine Beteiligung der Lehrer/innen bei der Planung von Leistungsmessungen und eine konsequente Orientierung an den Schülerlernvoraussetzungen. Nur so können wertvolle und feinkörnige Informationen zurückgemeldet werden, die sich direkt auf den Lernfortschritt des individuellen Lerners bzw. der individuellen Lernerin beziehen.

Beispiele für diese Art der intelligenten Rechenschaftslegung findet man in Ländern, die auch bei PISA bisher überdurchschnittlich gut abgeschnitten haben. Carless (2005) beschreibt die Einführung von Praktiken formativer Leistungsmessung in Hongkong, das sowohl durch die konfuzianische Tradition als auch durch die britische Kolonialzeit mehr als ausreichend Erfahrung mit externen Schulleistungstests aufweisen kann. Ziel der Reform ist eine bessere Vorbereitung der Schüler/innen auf lebenslanges Lernen. Deshalb soll durch staatliche Testreformen *assessment for learning* gefördert und gleichzeitig *high-stakes testing* reduziert werden.

Auch beim „PISA-Primus" Finnland wird seit den 1980er-Jahren eine Form der intelligenten Rechenschaftslegung entwickelt, in deren Zentrum das Vertrauen in die professionellen Akteure vor Ort steht (vgl. Sahlberg 2007). Die erste summative Leistungsmessung in der Biographie von finnischen Schüler/inne/n ist das Abitur. Alle weiteren Leistungsmessungen, seien sie zentral oder dezentral (von Lehrkräften erstellt), haben überwiegend einen formativen Charakter und dienen als differenzierte Rückmeldungen für Schüler/innen, Eltern und Lehrer/innen. Die Grundschule bleibt eine komplett testfreie Zone.

Zum Abschluss dieses Beitrags soll noch etwas ausführlicher auf Entwicklungen in Schottland hingewiesen werden (vgl. Hutchinson/Hayward 2005; Hayward 2007). Das schottische Schulministerium hat bereits Anfang der 1990er-Jahre eine Handreichung erstellt, in der die Bedeutung von formativen Leistungsmessungen für die Lern- und Leistungsentwicklung hervorgehoben wurde. Dieses Programm wurde zunächst nur halbherzig umgesetzt. Zur selben Zeit wurden in Großbritannien zudem nationale Testsysteme zur Überprüfung der Schülerleistungen eingeführt (*Education Reform Act*). Schottland folgte den bildungspolitischen Vorgaben in England, Wales und Nordirland.

Ende der 1990er-Jahre wurden die Ansätze einer formativen Leistungsmessung vor dem Hintergrund empirischer Belege der Effektivität dieser Methode wieder aufgenommen. Ebenfalls gab es eine bildungspolitisch vorangetriebene Neujustierung der bisherigen Praxis der Rechenschaftslegung: von zentralen Tests hin zu einer Selbstevaluation auf Schul- und Schulbezirksebene. Lehrer- und Elternverbände konnten zudem durchsetzen, dass ein „lower-stake"-Testsystem für Primar- und Sekundarstufe I eingeführt wurde, das Unterricht und Leistungsmessung der Lehrkräfte unterstützen kann.

Diese Bemühungen mündeten direkt in die Einführung des Programms „Assessment is for Learning“ (AifL) Anfang der 2000er-Jahre (vgl. Hayward 2007). Leistungsmessung wird in AifL als ein zentraler Bestandteil von Lehren und Lernen betrachtet und soll direkt zur Verbesserung von Lernprozessen und Lernergebnissen führen. Die bisherige Praxis der Leistungsmessung in schottischen Schulen soll durch AifL in kleinen Schritten verändert werden. Im Zentrum des Programms stehen die Schülerbeurteilungen durch Lehrkräfte (*professional judgements*). Gleichzeitig werden externe Leistungsmessungen zusammen mit den formativen Leistungsmessungen auf Schulebene in ein Bildungsmonitoring-System integriert, das der Administration als Rückmeldeinstrument dient. Auf diese Weise können summative und formative Aspekte der Tests klar voneinander getrennt werden. Überdies wird zwischen einer nach innen und nach außen gerichteten Nutzung der Daten unterschieden. Dies erhöht die Transparenz und führt dazu, dass jede Ebene informative und passgenaue Rückmeldedaten erhält.

Dieser kurze Blick über die Grenze zeigt, dass man in Deutschland nicht unbedingt die Fehler anderer Länder wiederholen müsste. Selbstverständlich werden Vergleichsarbeiten wohl nie zu so schädlichen Konsequenzen führen wie beispielsweise *high-stakes testing* in den USA. Allerdings zeigt eine Analyse vor dem Hintergrund des Konzepts der formativen Leistungsmessung, dass die in Deutschland implementierten Testsysteme noch stark entwicklungsbedürftig sind, wenn sie einen signifikanten Schritt in Richtung datenbasierter Schulentwicklung darstellen sollen.

Literatur

Altrichter, H./Heinrich, M. (2006): Evaluation als Steuerungsinstrument im Rahmen eines „neuen Steuerungsmodells“ im Schulwesen. In: Böttcher, W./Holtappels, H.-G./Brohm, M. (Hrsg.): Evaluation im Bildungswesen. Eine Einführung in Grundlagen und Praxisbeispiele. Weinheim/München: Juventa, S. 52-64.

Birenbaum, M./Breuer, K./Cascallar, E./Dochy, F./Dori, Y./Ridgway, J./Wiesemes, R./Nickmans, G. (2006): A Learning Integrated Assessment System. In: Educational Research Review 76, H. 1, S. 61-67.

Black, P./William, D. (1998a): Assessment and Classroom Learning. In: Assessment in Education 5, H. 1, S. 7-74.

Black, P./William, D. (1998b): Inside the Black Box: Raising Standards through Classroom Assessment. In: Phi Delta Kappan 80, H. 2, S. 139-148.

Büchter, A/Leuders, T. (2005): From Students' Achievement to the Development of Teaching: Requirements for Feedback in Comparative Tests. In: Zentralblatt für Didaktik der Mathematik 37, H. 4, S. 324-334.

Carless, D. (2005): Prospects for the Implementation of Assessment for Learning. In: Assessment in Education 12, H. 1, S. 39–54.

Fraser, B.J./Walberg, H.J./Welch, W.W./Hattie, J.A. (1987): Syntheses of Educational Productivity Research. In: International Journal of Educational Research 11, H. 2, S. 147-252.

Harlen, W. (2005): Teachers' Summative Practices and Assessment for Learning – Tensions and Synergies. In: The Curriculum Journal 16, H. 2, S. 207-233.
Hattie, J.A./Timperley, H. (2007): The Power of Feedback. In: Review of Educational Research 77, H. 1, S. 81-112.
Herman, J.L. (2004): The Effects of Testing on Instruction. In: Fuhrman, S.H./Elmore, R.F. (Hrsg.): Redesigning Accountability Systems for Education. New York/London: Teachers College Press, S. 141-166.
Haywards, E.L. (2007): Curriculum, Pedagogies and Assessment in Scotland: the Quest for Social Justice. 'Ah kent yir faither'. In: Assessment in Education 14, H. 2, S. 251-268.
Hutchinson, C./Hayward, L. (2005): The Journey so Far: Assessment for Learning in Scotland. In: The Curriculum Journal 16, H. 2, S. 225-248.
Kühle, B./Peek, R. (2007): Lernstandserhebungen in Nordrhein-Westfalen. Evaluationsbefunde zur Rezeption und zum Umgang mit Ergebnisrückmeldungen in Schulen. In: Empirische Pädagogik 21, H. 4, S. 428-447.
Kuper, H. (2006): Rückmeldung und Rezeption – zwei Seiten der Verwendung wissenschaftlichen Wissens im Bildungssystem. In: Kuper, H./Schneewind, J. (Hrsg.): Rückmeldung und Rezeption von Forschungsergebnissen. Münster: Waxmann, S. 7-16.
Kuper, H./Hartung, V. (2007): Überzeugungen zur Verwendung des Wissens aus Lernstandserhebungen. In: Zeitschrift für Erziehungswissenschaft 10, H. 2, S. 214-229.
Maier, U. (2007): Welche Konsequenzen ziehen Mathematiklehrkräfte aus verpflichtenden Diagnose- und Vergleichsarbeiten? In: mathematica didacta 30, H. 2, S. 5-31.
Maier, U. (2008): Rezeption und Nutzung von Vergleichsarbeiten – Ergebnisse einer Lehrerbefragung in Baden-Württemberg. In: Zeitschrift für Pädagogik 54, H. 1, S. 95-117.
Nachtigall, C./Jantowski, A. (2007): Die Thüringer Kompetenztests unter besonderer Berücksichtigung der Evaluationsergebnisse zum Rezeptionsverhalten. In: Empirische Pädagogik 21, H. 4, S. 401-410.
Peek, R./Steffens, U./Köller, O. (2006): Positionspapier des Netzwerks Empiriegestützte Schulentwicklung (EMSE) zu: Zentrale standardisierte Lernstandserhebungen. 5. EMSE-Tagung, Berlin, 08.12.2006.
Posch, P. (2009): Zur schulpraktischen Nutzung von Daten: Konzepte, Strategien, Erfahrungen. In: Die Deutsche Schule 101, H. 2, S. 119-135.
Sahlberg, P. (2007): Education Policies for Raising Student Learning: The Finnish Approach. In: Journal of Education Policy 22, H. 2, S. 147-171.
Sill, H.-D./Sikora, C. (2007): Leistungserhebungen im Mathematikunterricht – Theoretische und empirische Studien. Hildesheim: Franzbecker.
Sloane, F.C./Kelly, A.E. (2003): Issues in High-Stakes Testing Programs. In: Theory Into Practice 42, H. 1, S. 12-17.
Wiliam, D./Lee, C./Harrison, C/Black, P. (2005): Teachers Developing Assessment for Learning: Impact on Student Achievement. In: Assessment in Education 11, H. 1, S. 49-65.

Uwe Maier, PD Dr., Jg. 1971, Akademischer Oberrat im Institut für Erziehungswissenschaft der Pädagogischen Hochschule Schwäbisch Gmünd, momentan Professurvertretung an der Universität Erfurt; Arbeitsschwerpunkte: Nutzung von Vergleichsarbeiten, Aufgabenkultur, datenbasierte Schulentwicklung.

Anschrift: Pädagogische Hochschule Schwäbisch Gmünd, Institut für Erziehungswissenschaft, Oberbettringerstraße 200, 73525 Schwäbisch Gmünd
E-Mail: uwe.maier@ph-gmuend.de