pedocs
Open Access Erziehungswissenschaften

DIPF

Maier, Uwe

Effekte testbasierter Rechenschaftslegung auf Schule und Unterricht. Ist die internationale Befundlage auf Vergleichsarbeiten im deutschsprachigen Raum übertragbar?

Zeitschrift für Pädagogik 56 (2010) 1, S. 112-128

Quellenangabe/ Reference:
Maier, Uwe: Effekte testbasierter Rechenschaftslegung auf Schule und Unterricht. Ist die internationale Befundlage auf Vergleichsarbeiten im deutschsprachigen Raum übertragbar? - In: Zeitschrift für Pädagogik 56 (2010) 1, S. 112-128 - URN: urn:nbn:de:0111-opus-71388 - DOI: 10.25656/01:7138

https://nbn-resolving.org/urn:nbn:de:0111-opus-71388
https://doi.org/10.25656/01:7138

in Kooperation mit / in cooperation with:

BELTZ JUVENTA

http://www.juventa.de

Nutzungsbedingungen

Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen: Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

We grant a non-exclusive, non-transferable, individual and limited right to using this document.
This document is solely intended for your personal, non-commercial use. Use of this document does not include any transfer of property rights and it is conditional to the following limitations: All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

pedocs
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation
Informationszentrum (IZ) Bildung
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Mitglied der Leibniz-Gemeinschaft

ZEITSCHRIFT FÜR PÄDAGOGIK

Jahrgang 56 – Heft 1
Januar/Februar 2010

Inhaltsverzeichnis

Thementeil: Bildung in der Demokratie

Allgemeiner Teil

Besprechungen

Dokumentation

Uwe Maier

Effekte testbasierter Rechenschaftslegung auf Schule und Unterricht

Ist die internationale Befundlage auf Vergleichsarbeiten im deutschsprachigen Raum übertragbar?

Zusammenfassung: Ausgehend von der Annahme, dass es ein internationales Grundmuster testbasierter Rechenschaftslegung gibt, werden in diesem Artikel internationale Befunde empirischer Forschung zu Effekten zentraler Tests zusammengefasst und geordnet. Die Literaturübersicht zeigt, dass extrem negative Konsequenzen vor allem mit der funktionellen Einbettung testbasierter Rechenschaftslegung in Ländern wie den USA oder England zusammenhängen. Darüber hinaus gibt es allerdings auch Studien, die auf die relative Bedeutungslosigkeit innovativer, zentraler Tests für die Unterrichtsentwicklung aufmerksam machen. Ein Forschungszweig, der im deutschsprachigen Raum noch zu wenig rezipiert wurde. Auch internationale Forschungsberichte zur Rezeption und Nutzung zentraler Leistungsrückmeldungen auf Schulebene könnten instruktiv für die Implementation von Vergleichsarbeiten sein. Abschließend wird erörtert, welche forschungsmethodologischen Implikationen sich für Wirkungs- und Rezeptionsstudien ergeben.

1. Fragestellung

Testbasierte Rechenschaftslegung ist die Kernkomponente einer weltweiten, bildungspolitischen Reformbewegung, die nun auch den deutschsprachigen Raum erreicht hat (vgl. Maier 2009). Sichtbar wurde dies an der Einführung von Bildungsstandards und zentralen Vergleichsarbeiten. Während Deutschland erst am Anfang dieser globalen Bewegung steht, befindet man sich in den USA mit der „No Child Left Behind"-Gesetzgebung (NCLB) bereits auf einem vorläufigen Höhepunkt. NCLB verpflichtet die US-Bundesstaaten zur konsequenten Einführung testbasierter Rechenschaftslegung für Einzelschulen und lokale Schuldistrikte. Zentrales Instrument sind länderspezifische, standardorientierte „high-stakes tests" mit empfindlichen Konsequenzen für Schulen, wenn mehrfach in Folge der festgelegte Leistungszuwachs nicht erreicht wird (Stecher 2002; Hursh 2005).

Die pädagogische Empörung über die Folgen einer radikal „output"-orientierten Bildungspolitik in den USA befindet sich ebenfalls auf dem Höhepunkt. Es gibt zahlreiche Autoren, die zum Teil auch sehr polemisch die Konsequenzen von „school accountability" und „high-stakes testing" vor Augen führen (z.B. Kohn 2000). Im Gegensatz zu dieser oft anekdotisch argumentierenden Literatur hat sich die empirische Bildungsforschung aus unterschiedlichen Perspektiven mit Effekten zentraler Tests auf Unterricht befasst (vgl. Stecher 2002; Herman 2004). Unabhängig davon entwickelte sich innerhalb der angewandten Linguistik die sogenannte „washback"-Forschung (Alderson/

Wall 1993; Cheng/Curtis 2004; Green 2007). Dort wird vor allem untersucht, inwiefern sich der Unterricht in sprachlichen Fächern durch standardisierte Zugangs- oder Abschlusstests beeinflussen lässt.

Auch in Deutschland bildete sich ein entsprechender Forschungszweig heraus (z.B. Sill/Sikora 2007; Hosenfeld/Groß Ophoff 2007; Tresch 2007; Nachtigall/Jantowski 2007). Forschungsmethodisch wird vor allem mit Lehrerbefragungen gearbeitet und in der Regel wird die Rezeptionsstudie von dem Testentwicklungsteam durchgeführt. Weiter fällt auf, dass in diesen Studien die Anbindung an internationale Ergebnisse zu Effekten zentraler Tests nicht oder allenfalls sporadisch geschieht. Dies kann zum einen daran liegen, dass die länderspezifischen Voraussetzungen für Standards und zentrale Tests sowie deren Ziele und Konsequenzen als sehr heterogen angesehen werden und somit eine Rezeption der internationalen Befunde gar nicht erst in Erwägung gezogen wurde. Ein weiterer Unsicherheitsfaktor sind unterschiedliche theoretische Referenzrahmen, die in der bisherigen Diskussion über Folgen von Standards und zentralen Tests angeboten werden. Klieme (2004) macht auf diese Problematik aufmerksam, empfiehlt dann aber doch sehr deutlich eine Orientierung am internationalen, empirischen Forschungsstand zu „high-stakes testing“ und „school accountability“.

In diesem Artikel soll der Frage nachgegangen werden, inwiefern die Befunde der internationalen, empirischen Forschung zu Effekten zentraler Tests auf die Situation im deutschsprachigen Raum übertragbar sind und sich damit auch als Referenzrahmen für die Weiterentwicklung von Wirkungs- und Rezeptionsstudien eignen. Die Basisannahme dieser Untersuchung ist ein internationales Grundmuster der testbasierten Rechenschaftslegung, das selbst bei vielfältigen, länderspezifischen Spielarten die Richtung der Effekte vorgibt. Zentrale, standardbasierte Tests sind innerhalb eines Schulsystems flächendeckend durchgeführte Leistungsmessungen in bestimmten Jahrgangsstufen und Kernfächern. Im deutschsprachigen Raum werden hierzu vor allem Begriffe wie Vergleichsarbeiten oder Lernstandserhebungen verwendet.

Was die Generalisierung und damit auch die Übertragbarkeit internationaler Befunde grundsätzlich erschwert, sind funktionelle Differenzen der Testsysteme zwischen den Ländern aber auch Funktionsüberfrachtungen innerhalb einzelner Länder (O‘Day 2004; Linn 2004). Einerseits werden Schulergebnisse veröffentlicht, um Eltern eine Informationsgrundlage für die freie Schulwahl zu bieten. Andererseits möchte man Schulen valide und vertrauenswürdige Leistungsdaten für die datenbasierte, interne Schulentwicklung zur Verfügung stellen. Weiter verschärft wird dieses Problem durch die Nutzung von „high-stakes tests“ für die schülerbezogene Zertifizierung. In Deutschland dagegen wird die Nutzung standardisierter Tests für wichtige Laufbahnentscheidungen noch abgelehnt.

Die an Leistungsrückmeldungen geknüpften Konsequenzen sind eine zweite Differenzlinie. Während im anglo-amerikanischen Raum zum Teil empfindliche Konsequenzen für Schüler, Lehrer und Schulen folgen können, wird in Deutschland vor allem an die schulinterne Nutzung der Daten gedacht. Allerdings kann man auch hier argumentieren, dass Deutschland erst am Beginn einer Entwicklung steht und Bildungspolitiker bald auf die Idee kommen könnten, wirkungsmächtigere Anreize und Sanktionen einzuführen.

Inwiefern die hier herausgestellten Differenzen eine Übertragbarkeit internationaler Befunde zu Effekten zentraler Tests relativieren, muss im Anschluss an die folgende Literaturübersicht diskutiert werden.

2. Literaturrecherche und Systematik

Um die internationale, empirische Forschung zu Wirkungen testbasierter Rechenschaftslegung beschreiben zu können, wurde eine systematische Literaturrecherche durchgeführt. Ausgangspunkt waren einschlägige Übersichtsartikel in Fachzeitschriften (z.B. Alderson/Wall 1993; Cheng/Curtis 2004; Watanabe 2004; Rea-Dickins/Scott 2007) und Buchkapitel (Stecher 2002; Herman 2004; O'Day 2004). Zusätzlich wurde eine Datenbankrecherche in ERIC, Science Direct und PsycInfo durchgeführt.

Die Suche wurde eingegrenzt auf begutachtete, englischsprachige Zeitschriftenartikel, die sich auf „Elementary and Secondary Education" beziehen und in den Jahren 2000 bis 2009 publiziert wurden. Als Suchbegriffe für testbasierte Rechenschaftslegung wurden vorgegeben: „high-stakes testing", „mandatory testing", „state mandated testing" oder „accountability". Mindestens einer dieser Suchbegriffe sollte im Titel vorkommen. Mit einer Und-Verknüpfung wurden weitere Suchbegriffe zu Effekten oder Wirkungen auf Unterricht und Schülerleistungen vorgegeben: „achievement", „performance", „impact", „effects", „consequences", „test use" oder „washback". Diese Suchbegriffe sollten entweder im Titel, im Abstrakt oder als Deskriptor erscheinen.

Die Rechercheergebnisse (ERIC: 163 Treffer; Science Direct: 78 Treffer; PsycInfo: 205 Treffer) wurden durch folgende Auswahlkriterien reduziert: Von Interesse waren lediglich Originalstudien oder Übersichtsartikel, die Effekte von testbasierter Rechenschaftslegung auf Lehrereinstellungen, Unterricht oder Schülerleistungen empirisch untersucht haben. Nicht berücksichtigt wurden Texte ohne empirischen Bezug, Literatur zu messtheoretischen Problemen und zur Frage der Validität von „high-stakes tests". Ebenfalls nicht aufgenommen wurden empirische Studien mit folgenden Merkmalen: Schülerbefragungen zu „high-stakes testing" (z.B. ethnographische Studien), Auswirkungen auf Schülermotivation, Studien aus Schwellen- oder Entwicklungsländern (z.B. Indien) und Texte zu Auswirkungen auf spezielle Akteure im Schulsystem (z.B. „students with special needs", Lehramtsstudierende, Schulverwaltung).

Für den Zeitraum 2000 bis 2009 wurden 70 Zeitschriftenartikel, die sämtlichen Suchkriterien entsprechen, ausgewählt. Es ist auffallend, dass trotz der immensen Literatur zu dieser bildungspolitisch und erziehungswissenschaftlich brisanten Thematik bei einer systematischen Literaturrecherche nur relativ wenige Studien übrig bleiben, die den oben genannten Kriterien entsprechen. Weiter fällt auf, dass die empirischen Zugänge sehr heterogen sind: Von Fallstudien (z.B. DeBray 2005; Wikeley/Stoll/Lodge 2002) über Studien mit mehrperspektivischen Zugängen (z.B. Luxia 2007) bis hin zu rein quantitativen Lehrerbefragungen (z.B. Parke u.a. 2006) oder quantitativen Analysen von Schulleistungsdaten (Borg u.a. 2007). Die von Watanabe (2004) beschriebenen „washback"-Dimensionen „Intentionalität" und „positive vs. negative Konsequenzen"

sowie die Unterscheidung von Testauswirkungen auf Produkte, Prozesse oder Personen (Cheng/Curtis 2004) dienen zur Systematisierung der Befunde.

3. Auswirkungen testbasierter Rechenschaftslegung auf Schule und Unterricht

3.1 Nicht intendierte, negative Konsequenzen zentraler Tests

In einer Vielzahl empirischer Untersuchungen wurden negative und nicht erwünschte Effekte von „high-stakes testing" untersucht (Stecher 2002). Ein empirisch sehr gut abgesichertes Ergebnis ist die Reduktion der verfügbaren Unterrichtszeit durch Testvorbereitungsaktivitäten. Die Testvorbereitungszeit ist höher an Schulen mit benachteiligter Schülerschaft. Nichtgetestete Fächer und Inhalte verlieren an Bedeutung oder werden gestrichen (Jones/Johnston 2004). Der Test wird das eigentliche Curriculum (Crocco/Costigan 2007). Eine Analyse von Schulcurricula in Kentucky (Stecher/Barron 2001) ergab eine stärkere Orientierung an staatlichen Tests als an den eigentlichen Bildungsstandards. Stecher und Barron (2001) beschreiben in diesem Zusammenhang sog. „curriculum swings". Einem Fach werden, je nachdem ob es in einer Klassenstufe getestet wird oder nicht, mehr oder weniger Wochenstunden zugewiesen.

Analog dazu führen Tests mit weitreichenden Sanktionen für Lehrer zu einer verstärkten Fokussierung auf testspezifische Inhaltsgebiete eines Faches. Wichtige und sinnvolle Inhalte, die der externe Test nicht enthält, werden nicht mehr unterrichtet: ganze Bücher lesen, authentische Schreibaufgaben, komplexe mathematische Problemstellungen, etc. (Stecher 2002). Für die Testvorbereitung werden vor allem simple, lehrergelenkte Unterrichtsmethoden und testähnliche Aufgabenformate eingesetzt (Vogler 2005; Grant 2007). Studien zum Sprachunterricht zeigten, dass Lehrer verstärkt Fehler in Texten suchen lassen, anstatt die Schüler selbst Texte produzieren zu lassen. Einfache Grammatikübungen und Auswendiglernen stehen im Vordergrund (Lipman 2002; Olson 2007; Valli/Chambliss 2007). Lehrkräfte geben zum Teil ihre persönlichen Prioritäten und Prinzipien für den Lese- und Schreibunterricht auf: z.B. so schreiben wie ein richtiger Autor (Assaf 2006; Watanabe 2007) und orientieren ihre eigenen Bewertungskriterien für Schreibprodukte an den staatlichen Tests (Ketter/Pool 2001; Slomp 2008).

Im naturwissenschaftlichen Unterricht in Kanada wurden schüleraktivierende und forschungsorientierte Unterrichtsstrategien aufgrund standardisierter Tests deutlich reduziert (Wideen u.a. 1997). Auch bezüglich der Einführung zentraler „science"-Tests in den USA ab 2008 werden ähnliche Konsequenzen befürchtet (Pringle u.a. 2005; Shaver u.a. 2007). An New Yorker „high schools" und „middle schools" wurden nach der Einführung eines den Anforderungen von NCLB genügenden Testsystems vor allem Kurse, die speziell auf Interessen der Schüler zugeschnitten waren, gestrichen (Hursh 2005). Ketter und Pool (2001) berichten von Fallstudien in U.S. „high schools", die zeigen, dass Testvorbereitungsaktivitäten Lehrer eher davon abhalten, sich mit den individuellen Lernvoraussetzungen der Schüler auseinanderzusetzen.

Trotz des Anspruchs, mit testbasierter Rechenschaftslegung die soziale Selektivität zu reduzieren, gibt es in den USA einen konstanten „test score gap“: Schüler ethnischer Minderheiten und sozioökonomisch benachteiligter Gruppen erreichen im Schnitt wesentlich niedrigere Werte bei zentralen Leistungstests als ihre weißen Mitschüler aus Mittelschichtfamilien (z.B. Burns u.a. 2004). Zahlreiche Analysen zeigen, dass „high-stakes tests“ dabei eher zu einer Verstetigung sozialer Disparitäten beitragen (Sloan 2007; Diamond 2007) und die in den USA neu aufkommende Tendenz zur Resegregation verstärken (Borman u.a. 2004). Wie in einem Teufelskreis sind diese Schulen dann immer weniger attraktiv für Eltern und qualifizierte Lehrkräfte (Rau/Shelley/Beck 2001; Lipman 2002; Tuerk 2005; Powers 2007).

Wenn Indikatoren dermaßen bedeutsam werden, gibt es immer auch Möglichkeiten der Korrumpierung (Amrein/Berliner 2003; Koretz 2008; Heilig/Darling-Hammond 2008). Schwächere Schüler werden von den Tests ausgeschlossen oder sogar von der Schule gedrängt (Clotfelter/Ladd 1996; Rau/Shelley/Beck 2001; Anagnostopoulos 2006). Sogenannte „low performing schools“ konzentrieren sich auf die noch aussichtsreiche Förderung von Schülergruppen mit mittleren Testwerten (Diamond/Spillane 2004). Sims (2008) fand heraus, dass leistungsschwache Schuldistrikte durch extra Schultage und Schulstunden die Testvorbereitungszeit erhöhen und damit leichte Zugewinne im Mathematiktest für Viertklässler erzielen können.

Ebenfalls sind deutliche Zusammenhänge zwischen Rechenschaftslegung und Abschlussquoten bzw. „drop out rates“ von Minderheitenschülern nachweisbar (Darling-Hammond 2004; Allensworth 2005; Roderick/Nagaoka 2005; Borg/Plumlee/Stranahan 2007; Hong/Youngs 2008). Dies gilt vor allem für Schüler mit Englisch als Zweitsprache, die in den USA den gleichen Testbedingungen ausgesetzt werden wie Muttersprachler (Solorzano 2008). Die staatlichen Mathematiktests beispielsweise übersteigen das sprachliche Niveau dieser Schüler, die zudem den staatlichen Leistungstest ohne zusätzliche Hilfsmittel schreiben müssen (Wright/Li 2008). Lehrer sehen keine Verbesserung der Unterrichtsbedingungen für diese Schüler durch NCLB (Wright/Choi 2006). Ähnliche Schwierigkeiten werden aus Kanada berichtet (Fox/Cheng 2007).

Eine Reihe von Studien beschäftigte sich auch mit den Einstellungen von Lehrkräften gegenüber zentralen Tests (Brown u.a. 2004; Crocco/Costigan 2007). Zentrale Leistungsmessungen werden in einigen Studien von den Lehrkräften als fair eingeschätzt. In anderen Berichten wird von einer Beeinträchtigung der Arbeitszufriedenheit geredet. Vor allem die Einschränkung der professionellen Entscheidungsfreiheiten durch die von außen aufgezwängten Konsequenzen werden hierfür als Grund genannt. Die Nützlichkeit für die Verbesserung des Unterrichts wird dagegen als sehr gering betrachtet (Loeb/Knapp/Elfers 2008).

3.2 Studien zu intendierten Effekten zentraler Tests auf Unterrichtsebene

Die zum Teil massive Kritik an negativen Konsequenzen von „high-stakes testing“ auf der Basis eines zu eng gefassten Leistungsbegriffs führte in zahlreichen US-Bundesstaaten zu Testreformen und einer neuen Generation von Testaufgaben. Dabei wurden auch aus ad-

ministrativer Sicht die Tests immer mehr als Hebel für die positive Veränderung von Unterrichtspraxis betrachtet (Stecher 2002). Popham (1987) prägte hierzu den Begriff „measurement driven instruction". Ebenso entstand die Überzeugung, mit einer qualitativen Verbesserung der Aufgabenstellungen und Testrückmeldungen die schädlichen Effekte von „high-stakes tests" vermeiden oder zumindest reduzieren zu können (Cheng 1999).

Diese Form der staatlichen Standards und Tests unterscheidet sich von Vorgängerversionen in der Regel durch folgende Merkmale (McDonnell/Choisser 1997):

- Die Tests sind mit Bildungsstandards verknüpft, die neben Grundfertigkeiten auch höherwertige Denkprozesse und Informationsverarbeitungsstrategien betonen.
- Zunehmend finden alternative Aufgabenformate Einzug in die standardisierten Tests („performance based assessment"): offene Fragen, Textproduktion, Experimente oder Portfolios.
- Die Tests sollen den für das Lernen Verantwortlichen eine Rückmeldung geben und damit den Unterricht positiv beeinflussen.

Die intendierten Effekte dieser neuen Generation staatlich verordneter Leistungsmessungen wurden in einer Reihe von Studien aus den Bereichen „general education" und angewandter Linguistik („washback"-Studien) untersucht. Beispielsweise berichtet Au (2007) in einer qualitativen Metasynthese von einer Erweiterung und Integration von Lerninhalten sowie kooperativen und schülerorientierten Unterrichtsmethoden aufgrund innovativer Tests. Allerdings spielt die Konzeption des Testsystems eine entscheidende Rolle.

Ein Terrain für Untersuchungen dieser Art war beispielsweise der US-Bundesstaat Kentucky. Dort wurde im Zusammenhang einer umfassenden Reform des Schulsystems das zentrale Testsystem KIRIS („Kentucky Instructional Results Information System") eingeführt. McDonnell und Choisser (1997) verglichen die unterrichtlichen Auswirkungen der KIRIS-Reform mit dem weiterhin hauptsächlich auf MC-Aufgaben basierenden Testsystem in North Carolina. Eine Dokumentenanalyse bestätigte einen signifikant höheren Unterrichtsanteil von Multiple-Choice Aufgaben in North Carolina. Dagegen verlangen die in Kentucky untersuchten Lehrer öfter von ihren Schülern produzierte Texte. Der Vergleich der Unterrichtsmethoden ergab allerdings, dass lehrergelenkte Unterrichtsphasen in beiden Staaten weiterhin dominieren.

Weitere Studien bestätigten die allenfalls moderaten Innovationseffekte der KIRIS-Reform in Kentucky. Stecher und Barron (2001) berichten von leichten Veränderungen bei Unterrichtsstrategien und Aufgabenstellungen. Dagegen erhöhte sich der Einsatz von testspezifischen Aufgaben erheblich. Stecher und Barron (2001) sowie Faulkner und Cook (2006) beschreiben die Auswirkungen auf Unterricht als kurzsichtig, weil auf spezifische Testelemente reagiert wird und nicht die breiten lernpsychologischen und fachdidaktischen Ideen der Standards im Mittelpunkt stehen.

Ein weiterer US-Bundesstaat mit reformorientierten staatlichen Standards und Tests war Maryland. Das „Maryland School Performance Assessment Program" (MSPAP) verlangt von Schülern beispielsweise, argumentativ auf einen Schreibimpuls zu reagieren oder mathematisches Wissen in möglichst realen Situationen anwenden zu können.

Parke u.a. (2006) fanden mit Dokumentenanalysen heraus, dass 70% der kodierten Schreib- und Leseaktivitäten mindestens ein Element oder Prinzip der MSPAP-Testaufgaben reflektieren. Zu einer kritischeren Einschätzung der Reformen in Maryland kommen Firestone, Winter und Fitz (2000). Die Mehrheit der Lehrer in dieser Studie berichteten zwar von bestimmten Testvorbereitungsaktivitäten, die sich speziell auf die neuen Aufgaben des MSPAP bezogen: z.B. längere Problemlöseaufgaben in Mathematik. Allerdings wurde das mathematikdidaktische Potenzial der komplexeren Aufgabenstellungen nach Ansicht der Autoren nicht voll ausgenutzt. Beispielsweise schränkten die Lehrer schon im Voraus die möglichen Lösungen ein, um somit die Bearbeitungszeit für die Schüler zu reduzieren.

Auch aus anderen US-Bundesstaaten und Ländern liegen empirische Berichte vor, dass Testreformen den Unterricht von Lehrkräften nicht unberührt lassen (Bol 2004; Cankoy/Tut 2005). Cheng (1999; 2003) untersuchte, ob sich die Revision des „Hong Kong Certificate Examination in English" auf den Englischunterricht der Sekundarstufe auswirkt. Bei der Revision wurden ein integrierter Hörverstehenstest und eine ausführliche Schreibaufgabe mit realistischem Schreibimpuls hinzugefügt. Cheng (2003) konnte gewisse Veränderungen im Unterricht von drei intensiv beobachteten Lehrern nach der Testrevision feststellen. Für den Unterricht in mündlicher Kommunikation nutzten alle drei Englischlehrer mehr Rollenspiele und Gruppendiskussionen. Auch eine neu hinzugefügte, authentische Schreibaufgabe im chinesischen „National Matriculation English Test" führte zu einer deutlichen Erhöhung der Unterrichtszeit für Schreibübungen (Luxia 2007). Allerdings fand in dieser Zeit nicht der von den Testkonstrukteuren intendierte, produktive Schreibunterricht statt. Es wurde weiterhin auf grammatikalische Korrektheit geachtet, anstatt auf kreatives Schreiben. Lehrer und Schüler konzentrierten sich vor allem auf stilistische Mittel, um die unabhängigen „Rater" zu beeindrucken: z.B. bestimmte Formulierungen, sauberes Papier und Handschrift, das Wortlimit nicht überschreiten.

3.3 Studien zu Effekten zentraler Tests auf institutioneller Ebene

Ein weiterer Typ von Studien analysiert die datenbasierten Kommunikationsstrukturen und die Nutzung der Rückmeldungen zur Ableitung von Verbesserungsmaßnahmen innerhalb von Einzelschulen. Von erfolgversprechenden, lokalen Modellen für die Nutzung zentraler Testdaten für die Schulentwicklung in England bzw. China berichten eine Reihe von Autoren (Wikeley/Stoll/Lodge 2002; Demie 2003; Peng u.a. 2006). Zentrale Herausforderungen für datenbasierte Schulentwicklung vor dem Hintergrund einer Theorie des Organisationslernens fassen Ingram, Louis und Schroeder (2004) zusammen.

Ein wichtiges Ergebnis dieser Literatur ist, dass die schulinterne Nutzung externer Leistungsdaten von außen gezielt unterstützt werden muss (Saunders 2000). Wikeley, Stoll und Lodge (2002) berichten beispielsweise, dass Lehrer durch die Teilnahme an spezifischen Fortbildungen zusehends in die Lage versetzt wurden, externe Leistungsdaten schüler- und schülergruppenspezifisch zu interpretieren. Daraufhin konnten spezifische Programme für diese Schülergruppen entwickelt und durchgeführt werden.

Wenn allerdings aufgrund eines schmalen Projektbudgets keine Fortbildungen und Besprechungen stattfanden, konnten Schulen mit ohnehin geringen Veränderungskapazitäten die externen Leistungsinformationen nicht nutzen.

Die vertrauensvolle Zusammenarbeit mit den lokalen Schulbehörden ist eine ebenso wichtige Voraussetzung für eine sinnvolle Dateninterpretation auf Schulebene (Rudd/Davies 2002). Als besonders hilfreich stellte sich heraus, wenn lokale Schulbehörden sinnvolle Leistungsindikatoren auswählen, mit weiteren organisationalen Variablen verknüpfen und in lesbarer Form den Schulen zur Verfügung stellen (Demie 2003; Louis/Febey/Schroeder 2005). Die Schnittstelle dieser Zusammenarbeit zwischen Schulverwaltung und Einzelschule sind Experten, die sich auf die Dateninterpretation und -nutzung spezialisiert haben.

Auch innerhalb der Einzelschule kommt es darauf an, dass Fachabteilungsleiter oder Lehrkräfte mit einer gewissen statistischen Expertise bei der Übersetzung und Interpretation von Daten helfen können (Wikeley/Stoll/Lodge 2002). Unterstützend wirkte zudem die schulinterne Belohnung von Lehrern und Abteilungen, wenn aufgrund der Datenrückmeldung unterrichtliche Innovationen erprobt wurden. Von Vorteil war ebenso, wenn Lehrer die korrigierten Leistungsdaten ihrer Schüler mit weiteren Performanzindikatoren oder qualitativen Beobachtungen in Verbindung bringen konnten (Saunders 2000).

Das generelle Innovationsklima an Schulen ist eine weitere, wichtige Kontextvariable. In Schulen mit geringen Veränderungskapazitäten beispielsweise standen die entwickelten Testrückmeldesysteme recht unverbunden anderen Initiativen gegenüber (Wikeley/Stoll/Lodge 2002). Dagegen steigt die Intensität der Datennutzung mit der Gesamtdauer einer Schule in einem Rückmeldeprojekt (Saunders 2000; Louis/Febey/Schroeder 2005). Die Nutzung externer Leistungsdaten muss somit als Prozess an den Schulen betrachtet werden.

Es gibt allerdings auch empirische Studien, die grundsätzlich in Frage stellen, dass Rückmeldungen aus „high-stakes tests“ valide Indikatoren für Schulqualität darstellen und von Lehrkräften zur Weiterentwicklung von Unterricht genutzt werden können. Mintrop und Trujillo (2007) kommen in einer mehrperspektivischen Studie zu dem Ergebnis, dass die absoluten Testwerte des kalifornischen Leistungstests in keiner systematischen Verbindung mit weiteren Kriterien zur Schul- und Organisationsqualität stehen. Ähnlich kritisch argumentieren Ingram, Louis und Schroeder (2004). Die Autoren untersuchten ebenfalls in einer mehrperspektivischen Studie datenbasierte Entscheidungsprozesse in US High Schools, die Teilnehmer eines staatlichen Qualitätsmanagements (Continuous Improvement) sind und als „best practice“-Schulen ausgewiesen wurden. Selbst in diesen Schulen sind Diskrepanzen zwischen professionellem Wissen und Evaluationswissen deutlich sichtbar.

3.4 Testbasierte Rechenschaftslegung und schulischer Kompetenzerwerb

Von besonderem Interesse sind Studien zu Effekten testbasierter Rechenschaftslegung auf den Kompetenzerwerb der Schüler. In den USA gelten diese Forschungen oft als

„harter Beweis“ für oder gegen „accountability policies“. Trotz methodisch und statistisch anspruchsvoller Forschungsdesigns können diese Studien allerdings nur eine grobe Abschätzung leisten. Die Prozesse innerhalb der Schulen werden vollständig ausgeblendet. Als quasi-experimentelle Bedingung gilt entweder die Einführung von Rechenschaftslegung (z.B. Ding/Davison 2005) oder die Varianz auf verschiedenen Dimensionen zur Charakterisierung staatlicher Systeme der Rechenschaftslegung (z.B. Muller/Schiller 2000; Amrein/Berliner 2003).

Von entscheidender Bedeutung für die Bewertung dieser Studien ist, welche Leistungsindikatoren auf welchem Aggregationsniveau genutzt werden. Methodisch überzeugend sind Studien, die Wirkungen unterschiedlicher Testsysteme innerhalb der USA miteinander vergleichen und dabei auf die US-weiten NAEP-Leistungsdaten zurückgreifen. Muller und Schiller (2000) fanden beispielsweise heraus, dass in Staaten, die zentrale Tests zur individuellen Schülerzertifizierung einsetzen, ungerechtfertigt niedrige Erwartungen von Lehrern reduziert werden konnten. Es kam zu einer teilweisen Entkopplung von Leistungsbeurteilungstendenzen und sozioökonomischem Status. Andererseits kommt es zu negativen Entwicklungen, wenn die Schulen für die Ergebnisse ihrer Schüler verantwortlich gemacht werden und bestimmte Konsequenzen tragen müssen. Diese Befunde konnten durch weitere Analyse von NAEP-Daten bestätigt werden (Schiller/Muller 2003). In Staaten mit höheren Anforderungen an den „high school“-Abschluss und auch in Staaten, die Testleistungen mit Konsequenzen für die Schule verbinden, melden sich mehr Schüler zu höheren Mathematikkursen an und verbleiben auch eher in diesen Kursen. Häufiges Testen hat dagegen keine entsprechenden Auswirkungen.

Sehr populär rezipierte Studien in diesem Forschungsparadigma sind die ebenfalls mit NAEP-Daten durchgeführten Bundesstaatenvergleiche der Arbeitsgruppen um Berliner (Amrein/Berliner 2003; Nichols/Glass/Berliner 2006). Diesen Studien zufolge scheint es keinerlei Hinweise zu geben, dass „high-stakes tests“ zu einem vertiefteren und anwendungsfähigeren Lernen führen. Gerade für ethnische Minderheiten und sozioökonomisch benachteiligte Schüler ergeben sich noch schlechtere Lernmöglichkeiten und Lernergebnisse. Rosenshine (2003) reanalysierte die von Amrein und Berliner verwendeten Daten und fügte eine Kontrollgruppe von Bundesstaaten hinzu, die keine Konsequenzen mit Testresultaten verknüpfen. In diesem Vergleich schneiden Staaten mit Testkonsequenzen, d.h. höherem Rechenschaftsdruck, besser ab. Die Kontroverse zwischen Rosenshine und Berliner weist auf die Problematik dieser Ländervergleiche hin. Je nach Auswahl von Ländern und Dimensionen zur Einteilung von Testsystemen können sich sehr unterschiedliche Befunde ergeben.

Auch Lee und Wong (2004; Lee 2006) führten einen Ländervergleich durch, in dem die Testsysteme aller 50 US-Staaten anhand der Dimensionen „state support for school resources“ und „state pressure for school accountability“ in vier unterschiedliche Gruppen eingeteilt wurden. Es gab keine Korrelation zwischen beiden Dimensionen, d.h. höherer „accountability“-Druck wurde von den Staaten nicht durch mehr Unterstützung kompensiert. Ebenso gelang es den Staaten mit höherem „accountability“-Druck nicht, die sozioökonomischen und ethnischen Disparitäten im Schulsystem zu minimieren. Es gab allerdings in dieser Studie auch keine Hinweise auf eine Verstärkung der Ungleich-

heit. Dafür zeigten Wachstumsanalysen auf der Grundlage nationaler Testdaten, dass sich die Kombination von „accountability"-Druck und Unterstützung positiv auf Schülerleistungen auswirkt. Keine leistungssteigernden Auswirkungen hat dagegen der reine „accountability"-Druck.

Zu einem ähnlichen Befund kommt auch eine Analyse der Wirkungen von NCLB „accountability policies" auf Leistungszuwächse bei Schulen mit hohen Nichtbestehensquoten (Springer 2008). Wenn Schulen durch Sanktionen dazu gezwungen werden, ihren Unterricht auf Schüler zu konzentrieren, die den staatlichen „high-stakes test" nicht bestanden haben (below proficient), sind Leistungssteigerungen bei dieser Gruppe zu verzeichnen.

In diesen Abschnitt müssen auch noch Studien eingereiht werden, die mit Daten aus internationalen „large scale assessments" bildungsökonomische Produktivitätsmodelle berechnen. Wößmann und Fuchs (2007; Wößmann 2007) fanden einen signifikanten Interaktionseffekt zwischen standardisiertem Testen und zentralen Abschlussexamen. In Schulsystemen ohne zentrale Abschlussprüfungen ist der Einfluss standardisierter Tests auf die Schulleistung signifikant negativ. Mit zentralen Examen kommt es zu positiven Effekten von standardisierten Tests auf alle drei Bereiche der in PISA erfassten Schulleistung. Wößmann und Fuchs (2007) führen diesen Befund auf mangelnde Zielkriterien in Schulsystemen ohne zentrale Examina zurück.

4. Transfer der Befunde und Anknüpfungspunkte für die deutschsprachige Rezeptions- und Wirkungsforschung

Wie lässt sich die internationale Befundlage zu Effekten testbasierter Rechenschaftslegung zusammenfassen und welche Teilergebnisse lassen sich auf die Situation im deutschsprachigen Raum übertragen?

1. Die Belege für negative, nicht erwünschte Konsequenzen zentraler Tests sind zahlreich und unmissverständlich. Dies ist vor allem dann der Fall, wenn die Tests eingeschränkte Lernziele überprüfen und die Konsequenzen für Lehrer, Schüler und Schulen hoch sind. Lehrer fühlen sich kontrolliert und bereiten ihre Schüler gezielt auf die Leistungsmessung vor. Einzelschulen versuchen die Leistungsindikatoren geschickt zu beeinflussen. Leidtragende sind in der Regel Schulen mit sozial benachteiligter Schülerschaft. Vor allem in den USA könnte die Diskrepanz zwischen bildungspolitischer Rhetorik und schulpraktischen Folgen kaum größer sein.

Die Stärke dieser negativen Effekte hängt vor allem mit dem Belohnungs- und Bestrafungssystem testbasierter Rechenschaftslegung in den USA zusammen. Auf die Situation im deutschsprachigen Raum übertragen, würde dies bedeuten, dass mit pädagogisch unerwünschtem „teaching to the test" zwar zu rechnen ist, allerdings nicht in der gleichen Intensität. Ergebnisse von Rezeptionsstudien in Deutschland bestätigen dies prinzipiell. Lehrer akzeptieren Vergleichsarbeiten und bereiten selbstverständlich ihre Schüler darauf vor (Tresch 2007; Nachtigall/Jantowski 2007; Hosenfeld/Groß

Ophoff 2007; Maier 2008a). Eine extreme Deformation des Unterrichts durch zentrale Tests wurde dagegen noch nicht empirisch nachgewiesen.

2. Positive, durch Testreformen intendierte Unterrichtsverbesserungen sind nachweisbar aber begrenzt. Lehrkräfte nehmen fachdidaktisch innovative Testaufgaben als externen, bürokratischen Input wahr und richten ihren Unterricht nur oberflächlich daran aus. An der grundlegenden Logik testbasierter Rechenschaftslegung ändert sich aus Perspektive der Lehrkräfte nichts. Eine professionelle Auseinandersetzung mit neuen fachdidaktischen Prinzipien kann auf diese Weise nicht stattfinden, weil Unterrichtsentwicklung an Einzelschulen immer auf einem professionellen, normativen Diskurs basiert (O'Day 2004). Testbasierte Rechenschaftslegung dagegen – und ist sie auch noch so innovativ – wird immer als Teil der bürokratischen Regelung von Schule wahrgenommen und kann deshalb nur äußerst begrenzt die normativen Vorstellungen über guten Unterricht beeinflussen.

Mit dieser Problematik muss man sich auch im deutschsprachigen Raum über kurz oder lang auseinandersetzen. Es mangelt zwar nicht an Modellen zur datenbasierten Unterrichtsentwicklung (z.B. Tresch 2007; Hosenfeld/Groß Ophoff 2007) und man denkt in den einzelnen Vergleichsarbeitsprojekten auch sehr differenziert und fachdidaktisch fundiert über innovative Testaufgaben nach (z.B. Büchter/Leuders 2005; Lorenz 2005). Aber die Veränderung von Unterricht ist damit nicht garantiert. Beispielsweise zeigt die Begleitbefragung zu den Thüringer Kompetenztests, dass Lehrer den Nutzen vor allem in der Leistungsdiagnostik und weniger in der Evaluation und Weiterentwicklung von Unterricht sehen (Nachtigall/Jantowski 2007). Aus Lehrersicht eigentlich naheliegend.

3. Betrachtet man Studien zur schulinternen Rezeption und Kommunikation externer Leistungsindikatoren, erkennt man gewisse Lösungsansätze, die sich allerdings kaum von Befunden der Schulentwicklungsforschung unterscheiden (Hulpia/Valcke 2004). Externe Leistungsindikatoren müssen gelesen und übersetzt werden. Schulen und Fachkollegien können sich diese Expertise aneignen oder an einzelne Personen deligieren. Die notwendigen Voraussetzungen hierfür sind allerdings bestimmte Kommunikationsstrukturen auf Schul- oder Fachabteilungsebene und ein gezieltes Unterstützungsangebot auf Schulsystemebene. Kurz zusammengefasst: Top-down Reformen bedürfen eines gezielten „change managements". Eine schulreformerische Binsenweisheit, die man getrost auch auf die datenbasierte Schul- und Unterrichtsentwicklung mittels Vergleichsarbeiten im deutschsprachigen Raum übertragen kann. Eine Aufarbeitung internationaler Praxisberichte könnte viele Impulse geben, wie lokale Schulverwaltung, Testinstitute und Fachlehrkräfte gezielt zusammenarbeiten können.

4. Testbasierte Rechenschaftslegung soll zu einer Verbesserung des schulischen Kompetenzerwerbs beitragen. In den USA wurde deshalb versucht, mit Ländervergleichen den Zusammenhängen zwischen „accountability" und Leistungssteigerungen in nationalen Tests nachzugehen. Die Ergebnisse sind auch hier widersprüchlich. Im besten Fall zeigt sich ein positiver Effekt von „accountability"-Druck und staatlicher Unterstützung. Es

gibt dagegen keine Anzeichen, dass testbasierte Rechenschaftslegung für sich genommen zu besseren Schülerleistungen führt.

Die inhaltliche Übertragung dieser Befunde sollte allerdings stark relativiert geschehen. In den US-internen Ländervergleichen wurden vor allem funktionelle Differenzen als unabhängige Variablen modelliert. Gerade in diesem Bereich gibt es jedoch die größten Unterschiede zur deutschsprachigen Form der testbasierten Rechenschaftslegung. Umso wichtiger ist vielmehr die Tatsache, dass es solche Studien überhaupt gibt. Sie machen darauf aufmerksam, um was es bei der Einführung zentraler Tests eigentlich gehen sollte. Allerdings ist fraglich, ob es jemals gelingen wird, einen positiven oder auch negativen Effekt einer isolierten staatlichen Reformmaßnahme auf Leistungssteigerungen bei „large scale assessments" stichhaltig empirisch modellieren und nachweisen zu können.

Zum Abschluss wäre noch zu fragen, welche methodologischen Impulse für die Erforschung von Vergleichsarbeitswirkungen aufgenommen werden sollten:

- Mit quantitativen Selbstauskünften von Lehrern werden hierzulande zum Teil sehr weitreichende Schlussfolgerungen über die Rezeption und Nutzung von Rückmeldedaten für die Schul- und Unterrichtsentwicklung gezogen. Eine Validierung solcher Befunde mit z.B. querschnittlich angelegten Dokumentenanalysen wie bei etlichen „washback"-Studien wäre notwendig (z.B. Cheng 1999).
- Ein weiterer Impuls könnte von den Ländervergleichen in den USA ausgehen (z.B. Firestone/Winter/Fitz 2000). Auch in Deutschland findet man je nach Bundesland unterschiedliche Realisierungen testbasierter Rechenschaftslegung. Dies wäre ein guter Ausgangspunkt für kontrastierende Studien (z.B. Maier 2008b).
- Nicht zuletzt sollte man auf eine methodologische Forderung reagieren, die auch international noch kaum eingelöst wurde: sog. „baseline" Studien (Wall/Horak 2007). Vor allem Revisionen oder Neueinführungen von Vergleichsarbeiten könnten genutzt werden, um „Vorher-Nachher" Vergleiche durchzuführen.

All diese methodologischen Impulse und Forderungen sind anspruchsvoll und nur mit großem Aufwand realisierbar. Wenn Bildungspolitik allerdings evidenzbasiert sein möchte, muss man sich einer gründlichen, unabhängigen Erforschung von Effekten testbasierter Rechenschaftslegung auch stellen.

Literatur

Alderson, J.C./Wall, D. (1993): Does washback exist? Applied Linguistics 14, S. 115–129.

Allensworth, E.M. (2005): Dropout Rates After High-Stakes Testing in Elementary School: A Study of the Contradictory Effects of Chicago's Efforts to End Social Promotion. Educational Evaluation and Policy Analysis 27, H. 4, S. 341–364.

Amrein, A.L./Berliner, D.C. (2003): The effects of high stakes testing on student motivation and learning. In: Educational Leadership 60, H. 5, S. 32–38.

Anagnostopoulos, D. (2006): "Real Students" and "True Demotes": Ending Social Promotion and the Moral Ordering of Urban High Schools. In: American Educational Research Journal 43, H. 1, S. 5–42.

Assaf, L. (2006): One Reading Specialist's Response to High-Stakes Testing Pressures. Reading Teacher 60, H. 2, S. 158–167.

Au, W. (2007): High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. Educational Researcher 36, H. 5, S. 258–267.

Bol, L. (2004): Teachers' Assessment Practices in a High-Stakes Testing Environment. Teacher Education and Practice 17, H. 2, S. 162–181.

Borg, M./Plumlee, J.P./Stranahan, H.A. (2007): Plenty of Children Left Behind: High-Stakes Testing and Graduation Rates in Duval County, Florida. Educational Policy 21, H. 5, S. 695–716.

Borman, K.M./McNulty Eitle, T./Michael, D./Eitle, D.J./Lee, R./Johnson, L./Cobb-Roberts, D./Dorn, S./Shircliffe, B. (2004): Accountability in a Postdesegregation Era: The Continuing Significance of Racial Segregation in Florida's Schools. In: American Educational Research Journal 41, H. 3, S. 605–631.

Brown, D./Galassi, J.P./Akos, P. (2004): School Counselors' Perceptions of the Impact of High-Stakes Testing. In: Professional School Counseling 8, H. 1, S. 31–39.

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 (ZDM) 37, H. 4, S. 324–334.

Burns, M.K./Courtad, C.A./Hoffman, H./Folger, W. (2004): A comparison of district-level variables and state accountability test results for public elementary and middle schools. Psychology and Education 41, H. 2, S. 17–26.

Cankoy, O./Tut, M.A. (2005): High-Stakes Testing and Mathematics Performance of Fourth Graders in North Cyprus. Journal of Educational Research 98, H. 4, S. 234–244.

Cheng, L. (1999): Changing assessment: Washback on teacher perspectives and actions. In: Teaching and Teacher Education 15, S. 253–271.

Cheng, L. (2003): Looking at the impact of a public examination change on secondary classroom teaching: A Hong Kong case study. In: Journal of Classroom Interaction 38, H. 1, S. 1–10.

Cheng, L./Curtis, A. (2004): Washback or Backwash: A Review of the Impact of Testing on Teaching and Learning. In: Cheng, L./Watanabe, Y./Curtis, A. (Hrsg.): Washback in Language Testing. Research Contexts and Methods. Mahwah/London: Lawrence Erlbaum, S. 3–17.

Clotfelder, C.T./Ladd, H.F. (1996): Recognizing and Rewarding Success in Public Schools. In: Ladd, H.F. (Ed.): Holding Schools Accountable: Performance-based reform in education. Washington: Brookings Institution Press, S. 23–64.

Crocco, M./Costigan, A. (2007): The narrowing of curriculum and pedagogy in the age of accountability: Urban educators speak out. In: Urban Education 42, H. 6, S. 512–535.

Darling-Hammond, L. (2004): Standards, Accountability, and School Reform. In: Teachers College Record 106, H. 6, S. 1047–1085.

DeBray, E. (2005): A Comprehensive High School and a Shift in New York State Policy: A Study of Early Implementation. The High School Journal 89, H. 1, S. 18–45.

Demie, F. (2003): Using Value-added Data for School Self-evaluation: A case study of practice in inner-city schools. In: School Leadership & Management 23, H. 4, S. 445–467.

Diamond, J./Spillane, J. (2004): High Stakes Accountability in Urban Elementary Schools: Challenging or Reproducing Inequality? Teachers College Record 106, H. 6, S. 1145–1176.

Diamond, J.B. (2007): Where the Rubber Meets the Road: Rethinking the connection between high-stakes testing policy and classroom instruction. In: Sociology of Education 80, S. 285–313.

Ding, C.S./Davison, M.L. (2005): A longitudinal study of math achievement gains for initially low achieving students. In: Contemporary Educational Psychology 30, H. 1, S. 81–95.

Faulkner, S.A./Cook, C.M. (2006): Testing vs. Teaching: The Perceived Impact of Assessment Demands on Middle Grades Instructional Practices. Research in Middle Level Education 29, H. 7, S. 1–13.

Firestone, W.A./Winter, J./Fitz, J. (2000): Different assessments, common practice? Mathematics testing and teaching in the USA and England and Wales. In: Assessment in Education 7, H. 1, S. 13–37.

Fox, J./Cheng, L. (2007): Did we take the same test? Differing accounts of the Ontaria Secondary School Literacy Test by first and second language test-takers. Assessment in Education 14, H. 1, S. 9–26.

Grant, S.G. (2007): High-Stakes Testing: How Are Social Studies Teachers Responding? Social Education 71, H. 5, S. 250–254.

Green, A. (2007): Washback to learning outcomes: a comparative study of IELTS preparation and university pre-sessional language courses. In: Assessment in Education 14, H. 1, S. 75–97.

Heilig, J.V./Darling-Hammond, L. (2008): Accountability Texas-Style: The Progress and Learning of Urban Minority Students in a High-Stakes Testing Context. Educational Evaluation and Policy Analysis 30, H. 2, S. 75–110.

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.

Hong, W.-P./Youngs, P. (2008): Does high-stakes testing increase cultural capital among low-income and racial minority students? Education Policy Analysis Archives 16, H. 6, eingesehen am 23.12.2008 unter http://epaa.asu.edu/epaa/v16n6.

Hosenfeld, I./Groß Ophoff J. (Hrsg.) (2007): Nutzung und Nutzen von Evaluationsstudien in Schule und Unterricht. In: Empirische Pädagogik 21, H. 4, S. 352–457.

Hulpia, H./ Valcke, M. (2004): The Use of Performance Indicators in a School Improvement Policy: The Theoretical and Empirical Context. In: Evaluation and Research in Education 18, S. 102–120.

Hursh, D. (2005): The growth of high-stakes testing in the USA: accountability, markets and the decline in educational equality. In: British Educational Research Journal 31, H. 5, S. 605–622.

Ingram, D./Louis, K.S./Schroeder, R.G. (2004): Accountability Policies and Teacher Decision Making: Barriers to the Use of Data to Improve Practice. Teachers College Record 106, H. 6, S. 1258–1287.

Jones, B.D./Johnston, A.F. (2004): High-stakes testing in elementary school: Teachers' perceptions of the effects on teaching and student outcomes. In: Research in the Schools 11, H. 2, S. 1–16.

Ketter, J./Pool, J. (2001): Exploring the Impact of a High-Stakes Direct Writing Assessment in Two High School Classrooms. Research in the Teaching of English 35, H. 3, S. 344–393.

Klieme, E. (2004): Begründung, Implementation und Wirkung von Bildungsstandards: Aktuelle Diskussionslinien und empirische Befunde. In: Zeitschrift für Pädagogik 50, S. 625–634.

Kohn, A. (2000): The case against standardized tests: Raising the scores, ruining the schools. Portsmouth, NH: Heinemann.

Koretz, D. (2008): Test-Based Educational Accountability. Zeitschrift für Pädagogik 54, S. 777–790.

Lee, J. (2006): Input-Guarantee Versus Performance-Guarantee Approaches to School Accountability: Cross-State Comparisons of Policies, Resources, and Outcomes. In: Peabody Journal of Education 81, H. 4, S. 43–64.

Lee, J./Wong K.K. (2004): The Impact of Accountability on Racial and Socioeconomic Equity: Considering Both School Resources and Achievement Outcomes. In: American Educational Research Journal 41, H. 4, S. 797–832.

Linn, R.L. (2004): Accountability Models. In: Fuhrman, S.H./Elmore, R.F. (Hrsg.): Redesigning Accountability Systems for Education. New York/London: Teachers College Press, S. 73–95.

Lipman, P. (2002): Making the global city, making inequality: The political economy and cultural politics of Chicago school policy. In: American Educational Research Journal 39, H. 2, S. 379–419.

Loeb, H./Knapp, M.S./Elfers, A.M. (2008): Teachers' response to standards-based reform: Probing reform assumptions in Washington State. Education Policy Analysis Archives 16, H. 8, eingesehen am 23.12.2008 unter http://epaa.asu.edu/epaa/v16n8.

Lorenz, J.H. (2005): Zentrale Lernstandsmessung in der Primarstufe: Vergleichsarbeiten Klasse 4 (VERA) in sieben Bundesländern. Zentralblatt für Didaktik der Mathematik 37, H. 4, S. 317–324.

Louis, K.S./Febey, K./Schroeder, R. (2005): State-Mandated Accountability in High Schools: Teachers' Interpretations of a New Era. In: Educational Evaluation and Policy Analysis 27, H. 2, S. 177–204.

Luxia, Q. (2007): Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes writing test in China. In: Assessment in Education 14, H. 1, S. 51–74.

Maier, U. (2008a): Rezeption und Nutzung von Vergleichsarbeiten aus der Perspektive von Lehrkräften. In: Zeitschrift für Pädagogik 54, S. 95–117.

Maier, U. (2008b): Vergleichsarbeiten im Vergleich – Akzeptanz und wahrgenommener Nutzen standardbasierter Leistungsmessungen in Baden-Württemberg und Thüringen. Zeitschrift für Erziehungswissenschaft 11, S. 453–474.

Maier, U. (2009): Wie gehen Lehrkräfte mit Vergleichsarbeiten um? Eine Studie zu testbasierten Schulreformen in Baden-Württemberg und Thüringen. Hohengehren: Schneider Verlag.

McDonnell, L.M./Choisser, C. (1997): Testing and teaching: Local implementation of new state assessments (CSE Tech. Rep. H. 442). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Mintrop, H./Trujillo, T. (2007): The Practical Relevance of Accountability Systems for School Improvement: A Descriptive Analysis of California Schools. Educational Evaluation and Policy Analysis 29, H. 4, S. 319–352.

Muller C./Schiller, K.S. (2000): Leveling the Playing Field? Students' Educational Attainment and States' Performance Testing. In: Sociology of Education 73, S. 196–218.

Nachtigall, C./Jantowski, A. (2007): Die Thüringer Kompetenztests unter besonderer Berücksichtigung der Evaluationsergebnisse zum Rezeptionsverhalten. In: Empirische Pädagogik 21, H. 4, S. 401–410.

Nichols, S.L./Glass, G.V./Berliner, D.C. (2006): High-stakes testing and student achievement: Does accountability pressure increase student learning? Education Policy Analysis Archives 14, H. 1, eingesehen am 20.12.2008 unter http://epaa.asu.edu/epaa/v14n1.

O'Day, J.A. (2004): Complexity, Accountability, and School Improvement. In: Fuhrman, S.H./Elmore, R.F. (Hrsg.): Redesigning Accountability Systems for Education. New York/London: Teachers College Press, S. 15–43.

Olson, K. (2007): Lost opportunities to learn: The effects of education policy on primary language instruction for English learners. In: Linguistics and Education 18, H. 2, S. 121–141.

Parke, C.S/Lane, S./Stone, C.A. (2006): Impact of a State Performance Assessment Program in Reading and Writing. In: Educational Research and Evaluation 12, H. 3, S. 239–269.

Peng, W.J./Thomas, S.M./Yang, X./Li, J. (2006): Developing school evaluation methods to improve the quality of schooling in China: a pilot ‚value added' study. In: Assessment in Education 13, H. 2, S. 135–154.

Popham, W.J. (1987): The merits of measurement-driven instruction. In: Phi Delta Kappa, 68, S. 679–682.

Powers, J.M. (2007): High-Stakes Accountability and Equity: Using Evidence from California's Public Schools Accountability Act to Address the Issues in „Williams v. State of California" – In: American Educational Research Journal 41, H. 4, S. 763–795.

Pringle, R.M./Carrier Martin, S. (2005): The Potential Impacts of Upcoming High-Stakes Testing on the Teaching of Science in Elementary Classrooms. In: Research in Science Education 35, S. 347–361.

Rau, W.C./Shelley, N.M./Beck, F.D. (2001): The Dark Engine of Illinois Education: A Sociological Critique of a „Well Crafted (Testing) Machine". In: Educational Policy 15, H. 3, S. 403–430.

Rea-Dickins, P./Scott, C. (2007): Washback from language tests on teaching, learning and policy: evidence from diverse settings. In: Assessment in Education 14, H. 1, S. 1–7.

Roderick, M./Nagaoka, J. (2005): Retention Under Chicago's High-Stakes Testing Program: Helpful, Harmful, or Harmless? Educational Evaluation and Policy Analysis 27, H. 4, S. 309–340.

Rosenshine, B. (2003): High-Stakes Testing: Another Analysis. Education Policy Analysis Archives 11, H. 24, eingesehen am 11.1.2009 unter http://epaa.asu.edu/epaa/v11n24.

Rudd, P./Davies, D. (2002): A revolution in the Use of Data? The LEA Role in Data Collection, Analysis and Use and its Impact on Pupil Performance. Slough: NFER.

Saunders, L. (2000): Understanding schools' use of ‚value added' data: the psychology and sociology of numbers. In: Research Papers in Education 15, H. 3, S. 241–258.

Schiller, K.S./Muller, C. (2003): Raising the Bar and Equity? Effects of State High School Graduation Requirements and Accountability Policies on Students' Mathematics Course Taking. Educational Evaluation and Policy Analysis 25, H. 3, S. 299–318.

Shaver, A./Cuevas, P./Lee, O./Avalos, M. (2007): Teachers' perceptions of policy influences on science instruction with culturally and linguistically diverse elementary students. In: Journal of Research in Science Teaching 44, H. 5, S. 725–746.

Sill, H.-D./Sikora, C. (2007): Leistungserhebungen im Mathematikunterricht – Theoretische und empirische Studien. Hildesheim, Franzbecker.

Sims, D.P. (2008): Strategic Responses to School Accountability Measures: It's All in the Timing. Economics of Education Review 27, H. 1, S. 58–68.

Sloan, K. (2007): High-stakes accountability, minority youth, and ethnography: Assessing the multiple effects. Anthropology & Education Quarterly 38, H. 1, S. 24–41.

Slomp, D.H. (2008): Harming not helping: The impact of a Canadian standardized writing assessment on curriculum and pedagogy. Assessing Writing 13, H. 3, S. 180–200.

Solorzano, R.W. (2008): High Stakes Testing: Issues, Implications, and Remedies for English Language Learners. Review of Educational Research 78, H. 2, S. 260–329.

Springer, M.G. (2008): The influence of an NCLB accountability plan on the distribution of student test score gains. Economics of Education Review 27, H. 5, S. 556–563.

Stecher, B.M. (2002): Consequences of large-scale, high-stakes testing on school and classroom practice. In: Hamilton, L.S./Brian M./Stecher, S./Klein, P. (Hrsg.): Making sense of test-based accountability in education. RAND Education, S. 79–100.

Stecher, B.M./Barron, S. (2001): Unintended consequences of test-based accountability when testing in „milepost" grades. In: Educational Assessment 7, H. 4, S. 259–281.

Tresch, S. (2007): Potenzial Leistungstest. Wie Lehrerinnen und Lehrer Ergebnisrückmeldungen zur Sicherung und Steigerung ihrer Unterrichtsqualität nutzen. Bern: hep-Verlag.

Tuerk, P.W. (2005): Research in the High-Stakes Era: Achievement, Resources, and No Child Left Behind. In: Psychological Science 16, H. 6, S. 419–425.

Valli, L./Chambliss, M. (2007): Creating classroom cultures: One teacher, two lessons, and a high-stakes test. In: Anthropology & Education Quarterly 38, H. 1, S. 57–75.

Visscher, A.J./Coe, R. (2003): School performance feedback systems: Conceptualisation, Analysis, and Reflection. In: School effectiveness and school improvement 14, H. 3, S. 321–349.

Vogler, K.E. (2005): Impact of an accountability examination on Tennessee social studies teachers' instructional practices. In: Research in the Schools 12, H. 2, S. 41–55.

Wall, D./Horak, T. (2007): Using baseline studies in the investigation of test impact. In: Assessment in Education 14, H. 1, S. 99–116.
Watanabe, M. (2007): Displaced Teacher and State Priorities in a High-Stakes Accountability Context. Educational Policy 21, H. 2, S. 311–368
Watanabe, Y. (2004): Methodology in Washback Studies. In: Cheng, L./Watanabe, Y./Curtis, A. (Hrsg.): Washback in Language Testing. Research Contexts and Methods. Mahwah/London: Lawrence Erlbaum, S. 19–36.
Wideen, M.F./O'Shea, T./Pye, I./Ivany, G. (1997): High-stakes testing and the teaching of science. In: Canadian Journal of Education 22, H. 4, S. 428–444.
Wikeley, F./Stoll, L./Lodge, C. (2002): Effective School Improvement: English Case Studies. Educational Research and Evaluation. In: School effectiveness and school improvement in a European context 8, H. 4, S. 363–385.
Wößmann, L./Fuchs, T. (2007): What Accounts for International Differences in Student Performance? A Re-Examination Using PISA Data. In: Empirical Economics 32, S. 433–464.
Wößmann, L. (2007): International Evidence on School Competition, Autonomy and Accountability: A Review. In: Peabody Journal of Education 82, S. 473–497.
Wright, W.E./Choi, D. (2006): The Impact of Language and High-Stakes Testing Policies on Elementary School English Language Learners in Arizona. Education Policy Analysis Archives 14, H. 13, eingesehen am 11.1.2009 unter http://epaa.asu.edu/epaa/v14n13.
Wright, W.E./Li, X. (2008): High-Stakes Math Tests: How „No Child Left Behind" Leaves Newcomer English Language Learners behind. Language Policy 7, H. 3, S. 237–266.

Abstract: Assuming that accountability an international mandatory testing pattern of test-based exists, the author summarizes and categorizes international findings on the effects of accountability. The survey on existing literature shows that extremely negative consequences are above all linked with the functional integration of test-based accountability in countries such as the US or England. On the other hand, there are also studies that point to the relative insignificance of innovative mandatory tests for the development of teaching; – a research branch which has as yet not gained enough attention in the German-speaking countries. International research reports on the reception and use of feedback concerning achievement on the school level, too, could be instructive for the implementation of mandatory testing. In a final part, the author discusses research-methodological implications for studies on test effects and on feedback usage.

Anschrift des Autors

PD Dr. Uwe Maier, Akad. Oberrat, Institut für Erziehungswissenschaft, Pädagogische Hochschule Schwäbisch Gmünd, Oberbettringerstraße 200, 73525 Schwäbisch Gmünd
E-Mail: uwe.maier@ph-gmuend.de