BERICHT

DDS – Die Deutsche Schule
104. Jahrgang 2012, Heft 1, S. 88-99

Uwe Maier/Harm Kuper

Vergleichsarbeiten als Instrumente der Qualitätsentwicklung an Schulen

Überblick zum Forschungsstand

Zusammenfassung
Vergleichsarbeiten wurden vor ca. zehn Jahren in Deutschland als Qualitätsentwicklungsmaßnahme eingeführt und sollen vor allem auf Schul-, Fachkonferenz- und Lehrerebene zur Reflexion über und Verbesserung von Unterricht genutzt werden. Qualitätsentwicklung und -sicherung durch zentrales Testen ist jedoch keine neue Idee. In diesem Überblicksartikel werden Ergebnisse der internationalen und deutschsprachigen Literatur zu Effekten zentraler Tests zusammengefasst und Implikationen für die Weiterführung der deutschsprachigen Forschung in diesem Bereich diskutiert.
Schlüsselwörter: Vergleichsarbeiten, Leistungsmessung, Qualitätsentwicklung, Schulentwicklung, Unterrichtsentwicklung, Evaluation

Mandatory Testing as an Instrument for School Quality Development
A Research Survey from an International Perspective

Abstract
Ten years ago, German states implemented mandatory testing policies which aim at instructional improvement on school, department and classroom level. However, the concept of test-based quality assurance and quality development is not a new idea. This research survey summarizes main empirical findings on how mandatory testing affected teaching and student learning in different educational contexts. The paper eventually discusses implications for further research on mandatory testing effects in Germany.
Keywords: mandatory testing, student assessment, quality development, school improvement, school evaluation

1. Vergleichsarbeiten als Schulreforminstrument

Vergleichsarbeiten sind Teil einer Gesamtstrategie der Kultusministerkonferenz für das Bildungsmonitoring. Ihnen liegt die Idee zugrunde, nicht nur auf internationaler bzw. bildungspolitischer Ebene Leistungsdaten zur Qualitätssicherung und Qualitätsentwicklung zu nutzen; vielmehr möchte man mit Instrumenten wie den Vergleichsarbeiten die von vielen Bildungspolitikern und -politikerinnen als erfolgreich wahrgenommene Logik einer datengestützten Qualitätsentwicklung durch Leistungsmessung flächendeckend auf alle Schulen eines Bundeslandes übertragen.

Die Idee, mit zentralen Leistungstests Schule oder Unterricht zu verbessern oder gar die Schüler und Schülerinnen zu vertieftem und nachhaltigem Lernen zu motivieren, ist nicht neu. Zunächst einmal muss man die historische Tatsache zur Kenntnis nehmen, dass die zentrale Messung und Zertifizierung von Schülerleistungen bereits immer ein besonderes Qualitätsmerkmal eines Schulsystems war. Die Attraktivität des Gymnasiums für das aufstrebende Bürgertum im 19. Jahrhundert war wesentlich an die Einführung des Abiturs als Zulassungsvoraussetzung für ein Studium gekoppelt. Asiatische Länder haben eine Jahrtausende alte Tradition des zentralen Prüfens und Selektierens, um die besten Köpfe für den Staatsdienst zu gewinnen. In den USA spielen zentrale Abschluss- oder Eingangstests seit über 100 Jahren eine Rolle für den akademischen Karriereweg.

International vergleichende Studien zeigen immer wieder, dass diese Form von testbasierter *student accountability* (Schüler und Schülerinnen werden getestet und tragen die Konsequenzen des Testergebnisses) in der Regel zu mehr Anstrengung (bei Schülern und Schülerinnen, Eltern und Lehrkräften) und besseren Leistungsergebnissen (wiederum gemessen mit zentralen Vergleichstests) führt (vgl. z.B. Bishop 1995; Wößmann 2007). Vergleichsarbeiten folgen jedoch nicht der Logik von zentralen Abschlussprüfungen. Adressaten der Testrückmeldungen sind vielmehr Schulen, Fachkonferenzen und einzelne Lehrkräfte (*school accountability*). Diese sollen die Ergebnisse analysieren und daraus Konsequenzen für die Optimierung des Unterrichts ziehen. Diese Funktion zentraler Tests findet man erstmals in den USA und in England in den 1980er-Jahren. In dieser Blütezeit des Neoliberalismus hielten verstärkt Ideen der leistungsorientierten Gestaltung öffentlicher Dienste Einzug. Auch Schulen sollten sich an harten Leistungsdaten orientieren und diese als Ausgangspunkt für Veränderungen heranziehen. Diese Denkweise gipfelte in der *No-Child-Left-Behind*-Gesetzgebung (NCLB) (2001) aus der Regierungszeit von G.W. Bush. Alle US-Bundesstaaten wurden zur Einführung von testbasierter Rechenschaftslegung für Schulen verpflichtet, und mit ausgefeilten Sanktionssystemen wurde versucht, Lehrkräfte zur Arbeit zu motivieren.

Dieser Hintergrund ist wichtig, um verschiedene Forschungsbereiche zur Thematik verstehen und ihre Relevanz für die Bewertung von Vergleichsarbeiten in Deutschland (z.B. VERA) einschätzen zu können. In Deutschland ist man zwar weit entfernt von harten Sanktionen für Schulen, die unterdurchschnittliche VERA-Testergebnisse aufweisen. Dennoch sollte man zur Kenntnis nehmen, dass zumindest der Schulverwaltung VERA-Testergebnisse auf Schul- und auch Lehrerebene bekannt sein dürften. Ebenso gehört zur VERA-Logik, dass Schulleitungen oder Fachkonferenzleiter und -leiterinnen die Testergebnisse kennen und damit Unterrichtsentwicklung anstoßen sollen. Damit richten sich an deutsche Vergleichsarbeiten und „high-stakes tests" in den USA oder England prinzipiell die gleichen Funktionserwartungen: Sie sollen dazu dienen Qualität zu sichern und gleichzeitig Lehrkräfte zur Verbesserung des Unterrichts stimulieren. Verschieden sind jedoch die angestrebten Steuerungsmechanismen, die im Falle der „high-stakes tests" auf administrativ induzierte Leistungskonkurrenz und im Falle der Vergleichsarbeiten auf die informationelle Unterstützung professioneller Schulentwicklung setzen.

2. Forschungsstand

Vor dem Hintergrund der Eingangsüberlegungen kann man sagen, dass über die Effekte zentraler Tests in einem „high-stakes" Kontext bereits sehr viel geforscht wurde. Auch wenn diese bildungspolitischen Regelungsstrukturen nicht direkt auf deutsche Verhältnisse übertragbar sind, sollte man diese Befunde für eine Einschätzung von standardbasierten Vergleichsarbeiten im „low-stakes" Kontext Deutschland dennoch zur Kenntnis nehmen (2.1). Anschließend werden Studien skizziert, die sich mit der Frage beschäftigen, ob und wie über zentrale Leistungsmessungen didaktische Innovationen transportiert werden können (2.2). In einem dritten Schritt geht es um Forschungsarbeiten, die Effekten von Testrückmeldungen auf Schulebene nachgehen (2.3). Abschließend werden zentrale Ergebnisse der deutschen Forschung zu Vergleichsarbeiten skizziert.

2.1 Forschung zu zentralen Tests im „High-Stakes"-Kontext

In einer Vielzahl empirischer Untersuchungen wurden negative und nicht erwünschte Effekte von „high-stakes testing" untersucht (vgl. Stecher 2002; Cheng/Curtis 2004; Herman 2004). Die große Mehrheit der Studien entstand im Zusammenhang mit der ersten Welle staatlicher Tests, die als Folge allgemeiner Unzufriedenheit mit dem US-Schulsystem ab den 1970er-Jahren sukzessive eingeführt wurden. Weitere Studien beschreiben die nicht intendierten, negativen Konsequenzen der Testsysteme von US-Bundesstaaten im Rahmen der aktuellen NCLB-Gesetzgebung.

Ein empirisch sehr gut abgesichertes Ergebnis ist die Reduktion der verfügbaren Unterrichtszeit durch Testvorbereitungsaktivitäten. Lehrkräfte setzen veröffentlichte Aufgaben früherer Tests oder kommerzielle Materialen ein, um Schüler und Schülerinnen auf die Prüfungen vorzubereiten. Nicht getestete Fächer und Inhalte verlieren an Bedeutung bzw. werden gestrichen. Analog dazu führen Tests mit weitreichenden Sanktionen für Lehrkräfte zu einer verstärkten Fokussierung auf testspezifische Inhaltsgebiete eines Faches. Für die Testvorbereitung werden vor allem simple, lehrergelenkte Unterrichtsmethoden und testähnliche Aufgabenformate eingesetzt (vgl. Ketter/Pool 2001; Vogler 2005; Assaf 2006; Grant 2007; Olson 2007; Valli/Chambliss 2007; Watanabe 2007).

Trotz des Anspruchs, mit testbasierter Rechenschaftslegung die soziale Selektivität zu reduzieren, gibt es in den USA einen konstanten „test score gap": Schüler und Schülerinnen ethnischer Minderheiten und sozioökonomisch benachteiligter Gruppen erreichen im Schnitt wesentlich niedrigere Werte bei zentralen Leistungstests als ihre weißen Mitschüler und Mitschülerinnen aus Mittelschichtfamilien. Zahlreiche Analysen zeigen, dass „high-stakes tests" dabei eher zu einer Verstetigung sozialer Disparitäten beitragen (vgl. z.B. Sloan 2007; Diamond 2007). Wie in einem Teufelskreis sind diese Schulen dann immer weniger attraktiv für Eltern und qualifizierte Lehrkräfte. Studien wiesen auch die Korrumpierung der externen Leistungsindikatoren nach, beispielsweise durch eine clevere Kategorisierung von Schülern und Schülerinnen. Um Testwertsteigerungen zu erhalten, wurden Angehörige von Minderheitengruppen in sog. „disability programs" überführt.

Eine Serie von Forschungsarbeiten beschäftigte sich in den USA mit den Effekten zentraler Leistungstests auf Schülerleistungen (deutschsprachige Zusammenfassung: Maier 2010). Ausgangspunkt dieser Studien war die Arbeit von Amrein und Berliner (2003). Sie teilten die US-Bundesstaaten in „high-stakes testing" (HST)-Staaten und „low-stakes testing" (LST)-Staaten und verglichen die Schülerleistungszuwächse in diesen beiden Gruppen auf Basis nationaler Vergleichsdaten. Das Fazit dieser Analyse war ernüchternd: Die Werte der HST-Staaten waren in nationalen Längsschnittstudien (z.B. *National Assessment of Educational Progress*: NAEP) insgesamt rückläufig. Rosenshine (2003) reagierte mit einer kritischen Reanalyse dieser Studie durch Hinzunahme einer Kontrollgruppe von 15 LST-Staaten. Rosenshine fand einen moderaten Leistungszuwachs bei den NAEP 4-Mathematiktests und einen substanziellen Effekt bei den NAEP 8-Mathematik- und NAEP 4-Lesetests jeweils zugunsten der HST-Staaten. Weitere Forschergruppen folgten diesem Design und legten jeweils auf Basis unterschiedlicher Leistungsvergleichsdaten und mit unterschiedlicher forschungsmethodischer Präzision Forschungsergebnisse vor (vgl. Carnoy 2005; Lee/Wong 2004). Diese Studien weisen auf differentielle Effekte hin. Eine Kombination aus hohem Rechenschaftsdruck auf Schulebene und zentralen Leistungsmessungen in Abhängigkeit von Randbedingungen in einzelnen Teilbereichen kann zu Schülerleis-

tungssteigerungen führen. Letztlich wird damit ein empirischer Beleg für die Kontingenz der Wirkungen testbasierter Schulreform gegeben.

2.2 Unterrichtsreformen über zentrale Tests

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 administrativer Sicht die Tests immer mehr als Hebel für die positive Veränderung der Unterrichtspraxis betrachtet (vgl. Popham 1987; Cheng 1999; Stecher 2002). Diese Form der staatlichen Standards und Tests unterscheidet sich von Vorgängerversionen in der Regel durch folgende Merkmale: (1) Die Tests sind mit Bildungsstandards verknüpft, die neben Grundfertigkeiten auch höherwertige Denkprozesse und Informationsverarbeitungsstrategien betonen. (2) Zunehmend finden alternative Aufgabenformate Einzug in die standardisierten Tests („performance based assessment"): offene Fragen, Textproduktion, Experimente oder Portfolios. (3) Die Tests sollen den für das Lehren Verantwortlichen eine Rückmeldung geben und damit den Unterricht positiv beeinflussen.

Die intendierten Effekte dieser neuen Generation staatlich verordneter Leistungsmessungen (z.B. in Kentucky oder Maryland) wurden in einer Reihe von Studien aus den Bereichen „general education" und angewandter Linguistik („washback"-Studien) untersucht (vgl. McDonnell/Choisser 1997; Stecher/Barron 2001; Faulkner/Cook 2006; Parke/Lane/Stone 2006; Firestone/Winter/Fitz 2000). Diese Studien lassen sich folgendermaßen zusammenfassen: Es gibt nachweisbare, intendierte Effekte staatlicher Testreformen auf Unterrichtsprozesse. Lehrkräfte nutzen innovative Schreib- oder Problemlöseaufgaben auch in ihrem regulären Unterricht. Allerdings zeigten die Studien ebenso, dass dieser Reformimpuls deutlich begrenzt ist. Es gab keine empirischen Belege dafür, dass ein nennenswerter Anteil der Lehrerschaft aufgrund fachdidaktisch innovativer Testsysteme ihre didaktischen Ansätze grundsätzlich überdenkt. Auch die „washback"-Forschung zu Testreformen in China und Hong Kong kommt zu ähnlichen Schlüssen (vgl. Cheng 1999, 2003; Luxia 2007).

2.3 Studien zum schulinternen Umgang mit Testrückmeldungen

In Studien zu Effekten zentraler Tests auf institutioneller Ebene werden Kommunikationsstrukturen und die Nutzung der Rückmeldungen zur Ableitung von Verbesserungsmaßnahmen innerhalb von Einzelschulen analysiert (vgl. Yang u.a. 1999; Hayes/Rutt 1999; Wikeley/Stoll/Lodge 2002; Demie 2003). Ein wichtiges Ergebnis dieser Literatur ist, dass die schulinterne Nutzung externer Leistungsdaten von außen gezielt unterstützt werden muss. Wikeley, Stoll und Lodge (2002) berichten beispiels-

weise über Lehrkräfte, die aufgrund der Teilnahme an spezifischen Fortbildungen zunehmend in die Lage versetzt wurden, externe Leistungsdaten schüler- und schülergruppenspezifisch zu interpretieren. Daraufhin konnten Programme für spezifische 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 (vgl. 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 (vgl. Demie 2003; Louis/Febey/Schroeder 2005). Die Schnittstelle dieser Zusammenarbeit zwischen Schulverwaltung und Einzelschule sind Experten und Expertinnen, die sich auf die Dateninterpretation und -nutzung spezialisiert haben (vgl. Hayes/Rutt 1999).

Auch innerhalb der Einzelschule kommt es darauf an, dass Lehrkräfte mit einer gewissen statistischen Expertise bei der Übersetzung und Interpretation von Daten helfen können (vgl. Yang u.a. 1999; Wikeley/Stoll/Lodge 2002). Fachabteilungsleitern und -leiterinnen muss es gelingen, fachdidaktisch orientierte Visionen im Kollegium zu bündeln und diese mit den Implikationen externer Leistungsindikatoren zu verknüpfen. Zentrale Testergebnisse werden auf Abteilungsebene vor allem dann systematisch analysiert, wenn sich das Lehrerkollegium bereits über die Notwendigkeit von Reformmaßnahmen verständigt hat. Das generelle Innovationsklima an Schulen ist eine weitere, wichtige Kontextvariable. In Schulen mit geringen Veränderungskapazitäten beispielsweise standen die entwickelten Testrückmeldesysteme recht unverbunden anderen Initiativen der Qualitätsentwicklung gegenüber.

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-amerikanischen *high schools*, die Teilnehmer eines staatlichen Qualitätsmanagements sind und als „best practice“-Schulen ausgewiesen wurden. Selbst in diesen Schulen sind Diskrepanzen zwischen professionellem Wissen und Evaluationswissen deutlich sichtbar. Lehrer und Lehrerinnen haben in der Regel einen eigenen Qualitätsmaßstab für ihren Unterricht, der sich deutlich von

externen Indikatoren unterscheidet und auf anekdotischer Evidenz und informellem Wissen basiert.

2.4 Deutschsprachige Forschung zur Rezeption und Nutzung von Vergleichsarbeiten

Mit der Einführung von Vergleichsarbeitsprojekten auf Länderebene entstand eine eigenständige Forschungslinie, die sich mit der Akzeptanz, Rezeption und Nutzung von Rückmeldedaten durch Lehrkräfte und Schulen beschäftigt.

Im Rahmen der Thüringer Kompetenztests wurden beispielsweise regelmäßige Fragebogenstudien zur Evaluation der Nutzung von Rückmeldedaten durchgeführt (vgl. Nachtigall/Jantowski 2007). Es zeigte sich, dass die organisatorisch-technische Realisierung der Dateneingabe und -auswertung für Lehrkräfte in der Regel kein größeres Problem darstellt. Durch die Einführung von Sofortrückmeldungen konnte die Nutzung der Leistungsdaten erhöht werden. Lehrkräfte berichteten ebenfalls über eine Verlagerung von unterrichtlichen Inhalten, eine Anpassung von Methoden und eine Intensivierung der Zusammenarbeit im Kollegium als Folge der Kompetenztests. Allerdings sehen Lehrer und Lehrerinnen den Nutzen der Kompetenztests vorrangig im Bereich der Leistungsdiagnostik und des Leistungsvergleichs. Als Evaluationsinstrument für den Unterricht werden sie nur teilweise wahrgenommen und genutzt.

Auch in Online-Befragungen zur Nutzung und Rezeption der nordrhein-westfälischen Lernstandserhebungen in der Sekundarstufe I äußern die Lehrkräfte insgesamt eine hohe Bereitschaft, sich mit den Daten auseinanderzusetzen (vgl. z.B. Kühle/Peek 2007; Kühle 2010). Die Bereitschaft zur nutzungsorientierten Rezeption der Ergebnisrückmeldungen hängt allerdings von der eingeschätzten Nützlichkeit und von der Akzeptanz ab und nicht von der Verständlichkeit der Rückmeldungen bzw. der Intensität der Auseinandersetzung. Auch im Rahmen des Verbundprojektes VERA 3/4 wurden regelmäßige Online-Lehrerbefragungen zur Rezeption und Nutzung der Testergebnisse durchgeführt (vgl. z.B. Groß Ophoff/Hosenfeld/Koch 2007). Die Vergleichsarbeiten wurden von den Grundschullehrkräften als informativ wahrgenommen, und die Durchführung sowie Auswertung bereiteten kaum Probleme. Im Vordergrund stand aus Lehrersicht die Ableitung von geeigneten Fördermaßnahmen für einzelne Schüler und Schülerinnen. Nur in einem geringeren Maße wurde die Bedeutung der Tests für die Reflexion des eigenen Unterrichts und eine ergebnisorientierte Unterrichtsentwicklung gesehen.

Die bis hierher referierten Befunde entstanden im Umkreis der Vergleichsarbeitsprojekte und können aus diesem Grund als „usability"-Studien bzw. Evaluationsvorhaben bezeichnet werden. Hierzu gehört beispielsweise auch die von Koch (2011) entwickelte und evaluierte Fortbildungsveranstaltung zur Verbesserung der

Datenkompetenz von Grundschullehrkräften. Darüber hinaus gibt es weitere empirische Arbeiten, die Rezeption und Nutzung von Leistungsrückmeldungen aus einer externen Perspektive analysieren.

Maier (2007, 2008a) führte eine qualitative Studie zur Nutzung von Rückmeldedaten in Baden-Württemberg durch. Lehrkräfte an Sekundarschulen in Baden-Württemberg nutzten die Vergleichsarbeitsrückmeldung vor allem als weitere Klassenarbeitsnote. Eine formativ-diagnostische Nutzung wurde aufgrund der eingeschränkten curricularen Validität des Tests und des ungünstigen Testzeitpunkts abgelehnt. Die von Lehrkräften in den Interviews berichteten Unterrichtsentwicklungsmaßnahmen waren eher rudimentär und zufällig. In einer weiteren Studie fand Maier (2008b) deutliche Unterschiede bei der Rezeption und Nutzung von Leistungsrückmeldungen in Abhängigkeit verschiedener Vergleichsarbeitssysteme. Bei den Kompetenztests in Thüringen (Klasse 6 und 8) erhielten die Lehrkräfte im Vergleich zu den baden-württembergischen Vergleichsarbeiten (Klasse 6) kriteriale und faire Vergleiche als Rückmeldung. Es zeigte sich, dass Thüringer Lehrkräfte sowohl die Akzeptanz, die curriculare Validität als auch die Nutzung der Leistungsdaten für Lerndiagnosen und die zukünftige Unterrichtsplanung höher einschätzten. In Thüringen waren die Leistungsrückmeldungen auch wesentlich häufiger Gegenstand systematischer Diskussionen in Fach- und Gesamtlehrerkonferenzen.

Kuper und Hartung (2007) fragten vor einem professions- und organisationstheoretischen Hintergrund, welche Rezeptions- und Deutungsmuster bei Lehrkräften im Umgang mit Vergleichsarbeitsrückmeldungen zu beobachten sind. Hierzu wurden „Orientierung an technologischen Gründen (empirisches Wissen)“ und „Orientierung an normativen Gründen“ theoretisch unterschieden. Mit qualitativen Daten (Fallanalysen in NRW zu Lernstand 9) zeigten Kuper und Hartung, dass beide Überzeugungen in Bezug auf Lernstandserhebungen vorkommen. Es kommt sowohl zur retrospektiven Ursachenzuschreibung durch Leistungsrückmeldungen (technologisches Handeln) als auch zur Betonung der Bedingungen professionellen Handelns (Ablehnung des Schlusses von Leistungsergebnissen auf Unterrichtsqualität). Kuper und Hartung gehen in ihrer Studie allerdings nicht von einer direkten Beeinflussung des pädagogischen Interaktionsverhältnisses aus. Vielmehr wäre denkbar, dass Lernstandserhebungen die Reflexion professioneller pädagogischer Arbeit organisatorisch unterstützen.

3. Fazit und Forschungsperspektiven

Der Überblick zum Forschungsstand zeigt eine Vielzahl von Ergebnissen auf, die institutionelle und organisatorische Kontingenzen der Rezeption und Verwendung von Ergebnissen aus zentralen Leistungsüberprüfungen in den Schulen belegen. Eine professions-, organisations- und/oder institutionentheoretische Rahmung empirischer Projekte, die sich mit den Folgen testbasierter Schulreform befassen, erscheint daher geboten und aussichtsreich. Eine Fokussierung auf die Auswirkungen schulübergreifender institutioneller Regelstrukturen ist insbesondere in der US-amerikanischen Forschung erfolgt. Diese Forschung ist deutlich mit den Folgen von „high-stakes testing" befasst; die Ergebnisse sind daher auf die Konzepte testbasierter Schulreform in den deutschen Ländern nicht unmittelbar übertragbar. Wohl aber sind ihnen theoretische und methodische Anregungen zu entnehmen. Darüber hinaus regt der internationale Vergleich der Befunde institutionentheoretische Interpretationen der Auswirkungen regulativer Rahmenbedingungen auf die Verwendung der Ergebnisse in den Schulen an. Der deutsche Forschungsstand nimmt diese Anregungen zum Teil auf, enthält aber allenfalls ansatzweise und kaum systematische Ergebnisse, die über die Betrachtung der Schulebene hinausgehen. Forschung zu den Folgen testbasierter Schulreform im Kontext der institutionellen Strukturen eines Mehrebenensystems ist ein Desiderat.

Das von den beiden Autoren geleitete Forschungsprojekt „Testbasierte Schulreformen im Mehrebenensystem des Schulsystems" reagiert auf dieses Forschungsdefizit. Das Projekt wird vom BMBF im Rahmen der Ausschreibung „Steuerung im Bildungssystem" gefördert. Ein Ziel ist die Untersuchung des Zusammenspiels verschiedener Akteure und Akteursgruppen bei der Rezeption und Nutzung von Vergleichsarbeitsrückmeldungen in der Organisation Schule. Die Daten werden in zusammenhängenden Organisations- und Verwaltungseinheiten erhoben (Schulen in einem Schulverwaltungsbezirk; Interviews mit Lehrkräften, Fachgruppenleitern und -leiterinnen und Schulleitungen innerhalb einer Schule); die Leitfadeninterviews arbeiten Relationen zwischen den Rezeptions- und Nutzungsformen auf unterschiedlichen Ebenen heraus. Ein zweites Ziel ist die Klärung der Frage, inwiefern sich institutionelle Regelungskontexte auf die Rezeption und Nutzung von Vergleichsarbeitsdaten auswirken. Um dieser Fragestellung nachgehen zu können, führen wir die qualitativen Befragungen an Gymnasien in vier Bundesländern (Berlin, Brandenburg, Baden-Württemberg und Thüringen) durch. Wir gehen davon aus, dass die institutionelle Umwelt einer Einzelschule sowohl durch die bundeslandspezifischen Regelungen zu Vergleichsarbeiten als auch durch eher implizite Vorgaben und Botschaften der lokalen Schulverwaltung geprägt wird. In vergleichenden Fallstudien soll herausgearbeitet werden, inwiefern diese Regelungskontexte die Rezeption und Nutzung von Rückmeldedaten an einzelnen Schulen bedingen.

Literatur

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.

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

Bishop, J.H. (1995): The Impact of Curriculum-Based External Examinations on School Priorities and Student Learning. In: International Journal of Educational Research 23, H. 8, S. 653-752.

Carnoy, M. (2005): Have State Accountability and High-Stakes Tests Influenced Student Progression Rates in High School? In: Educational Measurement: Issues and Practice 24, H. 4, S. 19-31.

Cheng, L. (1999): Changing Assessment: Washback on Teacher Perceptions and Actions. In: Teaching and Teacher Education 15, H. 3, 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.

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.B. (2007): Where the Rubber Meets the Road: Rethinking the Connection Between High-Stakes Testing Policy and Classroom Instruction. In: Sociology of Education 80, H. 4, S. 285-313.

Faulkner, S.A./Cook, C.M. (2006): Testing vs. Teaching: The Perceived Impact of Assessment Demands on Middle Grades Instructional Practices. In: 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.

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

Groß Ophoff, J./Hosenfeld, I./Koch, U. (2007): Formen der Ergebnisrezeption und damit verbundene Schul- und Unterrichtsentwicklung. In: Empirische Pädagogik 21, H. 4, S. 411-427.

Hayes, S.G./Rutt, S. (1999): Primary Analysis for Secondary Schools: An LEA Research Officer's Perspective on Helping Secondary Schools Interpret Assessment Data for School Improvement Purposes. In: Improving Schools 2, H. 2, S. 44–52.

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.

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

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

Koch, U. (2011): Verstehen Lehrkräfte Rückmeldungen aus Vergleichsarbeiten? Datenkompetenz von Lehrkräften und die Nutzung von Ergebnisrückmeldungen aus Vergleichsarbeiten. Münster u.a.: Waxmann.

Kühle, B. (2010): Zentrale Lernstandserhebungen – Ergebnisorientierte Unterrrichtsentwicklung? Schulische Strategien beim Umgang mit Ergebnissen aus den Schulrückmeldungen im Kontext der ersten Lernstandserhebungen 2004/2005 in Nordrhein-Westfalen. Berlin: Köster.

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./Hartung, V. (2007): Überzeugungen zur Verwendung des Wissens aus Lernstandserhebungen. In: Zeitschrift für Erziehungswissenschaft 10, H. 2, S. 214-229.

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.

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. (2007): Welche Konsequenzen ziehen Mathematiklehrkräfte aus verpflichtenden Diagnose- und Vergleichsarbeiten? In: mathematica didactica 30, H. 2, S. 5-31.

Maier, U. (2008a): Vergleichsarbeiten im Vergleich – Akzeptanz und wahrgenommener Nutzen standardbasierter Leistungsmessungen in Baden-Württemberg und Thüringen. In: Zeitschrift für Erziehungswissenschaft 11, H. 3, S. 453-474.

Maier, U. (2008b): Rezeption und Nutzung von Vergleichsarbeiten aus der Perspektive von Lehrkräften. In: Zeitschrift für Pädagogik 54, H. 1, S. 95-117.

Maier, U. (2010): Effekte testbasierter Rechenschaftslegung auf Schule und Unterricht: Ist die internationale Befundlage auf Vergleichsarbeiten im deutschsprachigen Raum übertragbar? In: Zeitschrift für Pädagogik 56, H. 1, S. 112-128.

McDonnell, L.M./Choisser, C. (1997): Testing and Teaching: Local Implementation of New State Assessments (CSE Tech. Rep. No. 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. In: Educational Evaluation and Policy Analysis 29, H. 4, S. 319-352.

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

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.

Popham, W.J. (1987): The Merits of Measurement-Driven Instruction. In: Phi Delta Kappan 68, H. 9, S. 679-682.

Rosenshine, B. (2003): High-Stakes Testing: Another Analysis. In: Education Policy Analysis Archives 11, H. 24. URL: epaa.asu.edu/epaa/v11n24/; Zugriffsdatum: 01.02.2012.

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.

Sloan, K. (2007): High-Stakes Accountability, Minority Youth, and Ethnography: Assessing the Multiple Effects. In: Anthropology & Education Quarterly 38, H. 1, S. 24-41.
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. Santa Monica, CA: 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.
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.
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.
Watanabe, M. (2007): Displaced Teacher and State Priorities in a High-Stakes Accountability Context. In: Educational Policy 21, H. 2, S. 311-368.
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. (2007): International Evidence on School Competition, Autonomy and Accountability: A Review. In: Peabody Journal of Education 82, H. 2-3, S. 473-497.
Yang, M./Goldstein, H./Rath, T./Hill, N. (1999): The Use of Assessment Data for School Improvement Purposes. In: Oxford Review of Education 25, H. 4, S. 469-483.

Uwe Maier, Prof. Dr., geb. 1971, Lehrstuhl für Schulpädagogik an der Universität Erlangen-Nürnberg.

Anschrift: Regensburgerstraße 160, 90478 Nürnberg
E-Mail: uwe.maier@ewf.uni-erlangen.de

Harm Kuper, Prof. Dr., geb. 1966, Arbeitsbereich Weiterbildung und Bildungsmanagement an der Freien Universität Berlin.

Anschrift: Arnimallee 12, 14195 Berlin
E-Mail: harm.kuper@fu-berlin.de