Journal for Educational Research Online
Journal für Bildungsforschung Online
Volume 2 (2010), No. 2, 125–152

Uwe Maier

Effekte von testbasiertem Rechenschaftsdruck auf Schülerleistungen: Ein Literaturüberblick zu quasi-experimentellen Ländervergleichsstudien

Zusammenfassung

Im vergangenen Jahrzehnt wurden in Deutschland Bildungsstandards und externe Vergleichsarbeiten nach internationalem Vorbild eingeführt, um die schulische Bildungsqualität zu sichern und zu verbessern. Diese testbasierten Schulreformen üben auf Einzelschulen jedoch einen weitaus geringeren Rechenschaftsdruck aus, als dies in angloamerikanischen Staaten seit Längerem der Fall ist. Vor allem durch die „No Child Left Behind"-Gesetzgebung wurde in den USA der testbasierte Rechenschaftsdruck auf Einzelschulen empfindlich erhöht. Dieser Literaturüberblick geht der Frage nach, wie sich eine Erhöhung des schulischen Rechenschaftsdrucks auf Schülerleistungen[1] *auswirkt. Hierzu werden verschiedene Typen von Vergleichsstudien, die Ländervarianzen im testbasierten Rechenschaftsdruck als quasi-experimentelle Versuchsbedingung modellieren, zusammengefasst und kritisch diskutiert. Es wird angenommen, dass diese Befunde auch für die Weiterentwicklung testbasierter Schulreform in Deutschland von Relevanz sind.*

Schlagworte

Bildungsstandards; Vergleichsarbeiten; Testbasierte Schulreformen; Ländervergleichsstudien; Schülerleistungen; Unterricht

Impact of test-based accountability pressure on student achievement: A literature review on cross-country comparative studies

Abstract

In the past decade, German federal states enacted educational standards and mandatory testing which aim at quality assurance and data-based school improvement. Accountability policies and high stakes testing in Anglo-Saxon countries were a pre-

Prof. Dr. Uwe Maier, Lehrstuhl für Schulpädagogik, Friedrich-Alexander-Universität Erlangen-Nürnberg, Regensburger Straße 160, 90478 Nürnberg, Deutschland
E-Mail: uwe.maier@ewf.uni-erlangen.de

1 Aus Gründen der Lesbarkeit wird im Text lediglich die männliche Form verwendet. Es sind jedoch stets beide Geschlechter gemeint.

cursor for the test-based reform movement in Germany. However, the stakes attached to test results in Germany are rather low and schools are not put under rigorous accountability pressure. This is one major difference to test-based school reforms in the United States where the "No Child Left Behind" legislation requires states to raise school accountability pressure. This literature review synthesizes evidence from international and US cross-country comparative studies on the effects of test-based school accountability pressure on student achievement. It is assumed that research evidence on the impact of accountability pressure might be of interest for further developments of mandatory testing systems in Germany.

Keywords
Test-based school reforms; Mandatory testing; Accountability pressure; Student achievement; Cross-country comparative studies

1. Rechenschaftsdruck als kritisches Merkmal testbasierter Schulreformen

Seit Anfang der 2000er Jahre wird an einer Neuausrichtung des deutschen Schulsystems durch testbasierte Schulreformen gearbeitet. Wesentliche Reformelemente sind *large scale assessments* als Ausgangspunkt für eine evidenzbasierte Bildungspolitik und länderspezifische Bildungsstandards sowie die darauf bezogenen Vergleichsarbeiten, die auf der Einzelschulebene eine datenbasierte Schulentwicklung unterstützen oder voranbringen sollen (Baumert, 2001; Klieme, 2004). Testbasierte Schulreformen folgen der Logik, durch Standardisierung und zentrale Messung von Schülerkompetenzen auf unterschiedlichen Systemebenen sowohl die Qualität von Bildungsangeboten zu sichern als auch diese schrittweise zu optimieren. Es wird angenommen, dass durch testbasierte Rechenschaftslegung nicht nur dem Bedürfnis nach Kontrolle der öffentlichen Mittelverwendung entsprochen wird, sondern dass der durch testbasierte Rechenschaftslegung erzeugte Druck auf die Einzelschule auch zu Anpassungen innerhalb der Institutionen und damit zu Qualitätsverbesserungen führt.

Testbasierte Schulreformen sind eine internationale Bewegung, deren Wurzeln vor allem im angloamerikanischen Bereich zu suchen sind. Seit den 1980er Jahren wurde in den USA die Evaluation von Bildungsprogrammen verstärkt an Testleistungen gekoppelt. Testbasierter Rechenschaftsdruck (*accountability pressure*) kann durch unterschiedliche Instrumente auf unterschiedlichen Schulsystemebenen erzeugt werden (Linn, 2004; Porter, 1994; Stecher, 2002). In den USA wurden zunächst staatliche Interventionen im Bildungssystem evaluiert. D. h. es entstand Rechenschaftsdruck für die Bildungsadministration, die für den Einsatz staatlicher Fördermittel verantwortlich gemacht wurde. Seit den 1990er Jahren wurde von den US-Bundesstaaten nach und nach durch die Einführung zentraler Abschlussprüfungen oder versetzungsrelevanter Prüfungen der Rechenschaftsdruck an einzelne Schüler weitergegeben (*student accountabi-*

lity). Diese Art von Rechenschaftsdruck wirkte sich über die Steuerungswirkung der Versetzungs- oder Abschlusszahlen einzelner Schulen wiederum indirekt auf Lehrkräfte und Einzelschulen aus.

Parallel hierzu entstand in den USA die Idee, vor allem durch zwei Steuerungsinstrumente den Veränderungsdruck auf institutioneller Ebene zu erhöhen. Einmal wurde versucht, durch neuartige Formen der externen Leistungsmessung (*portfolio assessments, performance-based assessments*) und bestimmte Rückmeldepraktiken gezielt den Unterricht zu reformieren und damit die Schüler zu höherwertigen Lernleistungen zu führen (Popham, 1987). Zudem wurde in den jüngsten US-Reformen die direkte Erzeugung von Rechenschaftsdruck auf die einzelne Schule bundesstaatlich festgeschrieben (*school accountability pressure*). Die im Jahr 2002 eingeführte „No Child Left Behind"-Gesetzgebung (NCLB) fordert von allen US-Bundesstaaten eine rigorose, testbasierte Rechenschaftslegung auf Einzelschulebene (Hamilton & Koretz, 2002). Schulen mit unterdurchschnittlichen Testwerten über mehrere Jahre hinweg müssen mit empfindlichen Sanktionen rechnen. Beispielsweise können Schulen von halbstaatlichen Trägern übernommen werden oder Eltern können auf Kosten der bisherigen Schule ihre Kinder auf eine Schule ihrer Wahl schicken.

Die deutsche Variante testbasierter Schulreformen unterscheidet sich davon in zwei zentralen Punkten: Funktion der Tests und Konsequenzen. Es gibt in einigen Bundesländern zwar seit langem zentrale Abschlussprüfungen, wie z. B. das Zentralabitur. Diese werden allerdings nicht als Indikator für Schulqualität herangezogen. Dafür wurden Vergleichsarbeiten in der Grundschule und der Sekundarstufe I eingeführt, die Lehrkräften und Schulen informative, standardbasierte Rückmeldungen geben sollen. Die öffentliche Rechenschaftsfunktion wird allenfalls an die aggregierten Daten auf Landesebene geknüpft oder an das nationale Bildungsmonitoring delegiert. Auch die mit den zentralen Testergebnissen verknüpften Konsequenzen sind im deutschsprachigen Raum sehr schwach bzw. überhaupt nicht vorhanden. Ziel ist eine neue Form der „testdatenbasierten" Schulentwicklung, d. h. eine interne Datennutzung, weitgehend ohne externen, testbasierten Rechenschaftsdruck. Warum sollte man also internationale Forschungsliteratur zu testbasiertem Rechenschaftsdruck rezipieren, wenn dieses Instrument gar nicht zur Debatte steht? Ich gehe aus folgenden Gründen davon aus, dass die Entwicklungen im angloamerikanischen Raum mittelfristig auch für die deutschsprachige Diskussion über die zukünftige Entwicklung von Bildungsstandards, Bildungsmonitoring und Vergleichsarbeiten von Relevanz sind:

Zunächst einmal gibt es ein theoretisches Argument: Rechenschaftsdruck ist ein konstitutives Element testbasierter Rechenschaftslegung als neuer Schulsystemsteuerungsstrategie. Wenn man ganz auf externe Konsequenzen zentraler Tests für Lehrer, Schulen, Schulverwaltung oder auch Schüler verzichtet, wird testbasierte Rechenschaftslegung auf schulische Selbstevaluation reduziert und die ursprüngliche Idee verliert ihre zentrale Steuerungskomponente. Diese theoretische Problematik lässt sich auch an der schrittweisen Weiterentwicklung von Vergleichsarbeiten in einzelnen Bundesländern erkennen. Verfolgt man die

Befunde der deutschsprachigen Rezeptionsstudien, gewinnt man den Eindruck, dass ein nicht unerheblicher Teil der Lehrkräfte und Schulleitungen den Vergleichsarbeitsrückmeldungen recht indifferent gegenüber steht, weil keinerlei Konsequenzen folgen. In Baden-Württemberg wurde sogar die Benotung der Vergleichsarbeiten abgeschafft und damit die Bedeutung für Schüler und indirekt auch für Lehrkräfte weiter gesenkt. Es gibt jedoch auch Anzeichen einer leichten Erhöhung des testbasierten Rechenschaftsdrucks auf Schulen. In NRW werden beispielsweise die besten Schulen in den Lernstandserhebungen öffentlich ausgezeichnet.

Die Bildungspolitik könnte den externen Rechenschaftsdruck auf Schulen durch testbasierte Schulreformen erhöhen, wenn die Qualität schulischer Bildung weiter kritisch in der Öffentlichkeit diskutiert wird. Ebenso sollte die Bildungsforschung darauf vorbereitet sein, dass die Bildungspolitik über kurz oder lang die Frage stellt, zu welchen Qualitätsverbesserungen die neuen Steuerungsinstrumente überhaupt führen. Als bildungspolitische Reaktion sind zwei Lösungsvarianten denkbar: Vergleichsarbeiten reduzieren und die testbasierten Schulreformen einfach ins Leere laufen lassen. Die zweite Variante wäre, die Bedeutung externer Leistungsmessungen auf Schulebene deutlich zu erhöhen, z. B. durch Veröffentlichung schulbezogener Daten oder eine konsequentere Einbindung der Vergleichsarbeitsergebnisse in neue Formen der externen Schulevaluation bzw. Schulinspektion. Um für diese Debatte gerüstet zu sein, sollten empirisch nachgewiesene Effekte eines erhöhten Rechenschaftsdrucks geordnet, kritisch diskutiert und auf die Lage in Deutschland bezogen werden. Der Artikel soll hierzu einen Beitrag leisten, indem die Befunde methodologisch besonders überzeugender Studien zu den Effekten von testbasiertem Rechenschaftsdruck auf Schülerleistungen in Form einer systematischen, narrativen Literaturübersicht zusammengefasst werden. Zunächst müssen zentrale Begriffe und Theorien geklärt werden.

1.1 Begriffsklärungen

In der Literatur wird vor allem zwischen „student accountability“ und „school accountability“ differenziert (z. B. Hamilton & Koretz, 2002; Linn, 2004). Mit „student accountability“ bezeichnet man den Einsatz zentraler Leistungstests für selektionsdiagnostische Zwecke. Diese Form testbasierter Rechenschaftslegung hat eine lange Tradition in den USA (z. B. „aptitude tests“ für Aufnahmeverfahren zum Militär, College-Zulassung, etc.). In den 1980er Jahren wurde in den USA als Reaktion auf den Bericht „A Nation at Risk“ (1983) gefordert, dass sich zentrale Abschlusstests und „end-of-course examinations“ stärker auf das unterrichtete Curriculum beziehen sollen („curriculum-embedded graduation exams“). Zentrale Abschlussprüfungen auf Sek-II-Niveau wurden von den einzelnen US-Staaten jedoch unterschiedlich schnell, in unterschiedlichem Umfang und mit unterschied-

lichen Konsequenzen eingeführt. Ebenso ist von Interesse, welche Ebene den Test administriert (Schule, Distrikt oder Bundesstaat).

Werden Schulleistungsdaten auf Einzelschulebene publiziert und ist die Einzelschule entweder den Eltern oder den lokalen Schulbehörden gegenüber rechenschaftspflichtig, spricht man von „school accountability". O'Day (2004) unterscheidet, je nachdem wie stark die elterliche Schulwahl von Leistungsdaten abhängt, noch einmal zwischen „market-based school accountability" und „bureaucratic school accountability". Im letzteren Fall gehen testbasierte Sanktionen bzw. Belohnungen von der Schulverwaltung aus. Einzelne Komponenten von „school accountability" wurden bereits in den 1990er Jahren von US-Staaten eingeführt (Vorreiter: z. B. Texas, North Carolina). Die NCLB-Reform der Bush-Administration im Jahr 2002 verlangt von allen US-Staaten die Einführung von „school accountability", d. h. von zentralen Testsystemen mit einem sehr rigiden Sanktionsmechanismus für Einzelschulen, die jährlich festgelegte Testwertzuwächse nicht erreichen.

Es ist von zentraler Bedeutung, bei der Aufarbeitung von empirischen Studien zu testbasiertem Rechenschaftsdruck so gut wie möglich zwischen „student accountability" und „school accountability" zu unterscheiden. Problematisch ist allerdings, dass US-Staaten zu unterschiedlichen Zeitpunkten und verschieden schnell die jeweiligen Reformen umsetzen und es zu einer Überlagerung von Instrumenten kommt, d. h. oft nicht genau zwischen „student accountability" und „school accountability" differenziert werden kann. Ebenso muss bedacht werden, dass die NCLB-Reform grundsätzlich von einem kombinierten Effekt beider Formen testbasierter Rechenschaftslegung ausgeht.

1.2 Theoretische Modelle zur Erklärung von Effekten testbasierten Rechenschaftsdrucks

Die Zusammenhänge zwischen verschiedenen Formen von Rechenschaftsdruck und einer Verbesserung von Schülerleistungen sind sehr komplex. Es gibt unterschiedliche Modelle bzw. Theorien, die ineinander verschachtelten Wirkmechanismen adäquat abzubilden. Als theoretische Rahmung für diese Literaturübersicht dienen zwei Ansätze: (1) Eine bildungsökonomische Perspektive, in der vor allem die Effekte von Rechenschaftsdruck auf die Motivation von Schülern als wichtiger Prädiktor für Lernleistungen im Mittelpunkt steht. (2) Eine organisations- und professionstheoretische Perspektive, in der die Nutzung bzw. Nicht-Nutzung externer Leistungsdaten im komplexen System Schule modelliert wird.

(1) Mit bildungsökonomischen Produktivitätsmodellen sollen Effekte von Schulsystemmerkmalen, u. a. zentrale Abschlussprüfungen aber auch erweiterte Schulautonomie, elterliche Schulwahlfreiheit, etc., auf Schülerleistungen erklärt werden. Bishop (1995) geht im Gegensatz zu einfachen Produktivitätsmodellen (Input erklärt Output) von einer erweiterten bildungsökonomischen Theorie aus, in der Schüler und Lehrer gemeinsam an dem Produkt „Bildung" arbeiten

(Koproduktion). Die Qualität des Produkts (Lernleistungen) hängt also davon ab, wie sehr sich sowohl Schüler als auch Lehrkräfte, Eltern und Schuladministration anstrengen. Das Verhalten der einzelnen Akteure hängt wiederum von den externen Anreizsystemen ab. Schüler beispielsweise kalkulieren das Kosten-Nutzen-Verhältnis ihrer Lernanstrengungen. Ein zentraler Nutzen für Schüler (im amerikanischen Schulsystem) ist die Zulassung zum College. Bishop geht davon aus, dass sich Schüler in besonderem Maße engagieren, wenn die College-Zulassung eine Hohe Hürde darstellt und beispielsweise an externe, curriculumbasierte Examina gekoppelt wird (vgl. „student accountability").

Indirekt entsteht dadurch Druck auf Lehrer und die Schuladministration, die zur Verfügung stehenden Ressourcen möglichst optimal für das Lernen der Schüler einzusetzen. Bishop prüft in seinen Studien deshalb auch, wie sich Lehrerausbildung, Lehrergehälter und Bildungsausgaben in Abhängigkeit zentraler Testsysteme verhalten. Auch Wößmann (2007) sowie Wößmann und Fuchs (2007) betonen diesen indirekten Effekt externer Tests, vor allem externer Abschlussprüfungen. Sie geben dem Schulsystem indirekt, jedoch sehr prägnant die zentralen Zielkriterien vor und sind damit so etwas wie die „Währung" eines Bildungssystems.

(2) In der erziehungswissenschaftlichen und organisationssoziologischen Literatur werden alternative Modelle zur theoretischen Erklärung von testbasierter Rechenschaftslegung diskutiert. Grundlage dieser Modelle sind Vorstellungen über die komplexe Struktur sozialer Organisationen, speziell des Schulsystems. Dabei stellt sich die Frage, wie die interne Nutzung extern generierter Schulleistungsdaten zur Verbesserung von Unterricht bzw. innerschulischer Kooperation und infolgedessen zur Verbesserung von Schülerleistungen beitragen kann („internal accountability", „data-based school improvement"). Vor dem Hintergrund system- und professionstheoretischer Prämissen entwickelte O'Day (2002, 2004) ein Rahmenmodell für die Effektivität testbasierter Rechenschaftslegung. Testbasierte Schulreformen sind dann nützlich für die Verbesserung von Lernen und Unterricht, wenn:

- Informationen bereitgestellt werden, die Lehrer und Schüler auf Unterricht und Lernen aufmerksam machen (curriculare Validität des Tests Periodizität, Spezifität);
- Lehrer zum aktiven Umgang mit diesen Informationen, d. h. zu einer ergebnisorientierten Weiterentwicklung von Strategien motiviert werden können;
- im Schulsystem Wissen und Fähigkeiten entwickelt werden können (auf Individual- und Systemebene), die eine valide Interpretation der Informationen unterstützen;
- es Anreizstrukturen gibt, die eine informationsbasierte Ressourcenverteilung ermöglichen (z. B. Stehen überhaupt zu verteilende Ressourcen zur Verfügung? Balance zwischen individuellen (Lehrer) und kollektiven (Schule, Kollegium) Anreizstrukturen?).

Das Rahmenmodell wurde bisher vor allem genutzt, um empirische Befunde zur Umsetzung testbasierter Rechenschaftslegung in Schulen einzuordnen. In diesem Artikel ergänzt es die bildungsökonomische Perspektive, die den Nachteil hat, dass schulinterne Abläufe und Prozesse lediglich als „black box" behandelt werden.

2. Forschungsmethodische Zugänge und Schwierigkeiten

Die Effekte testbasierter Schulreformen lassen sich im angloamerikanischen Raum gut studieren. Problematisch ist allerdings, dass sich die Formen testbasierter Rechenschaftslegung in den letzten 30 Jahren stetig veränderten, es höchst unterschiedliche Forschungsansätze gibt und eine nicht unerhebliche Anzahl von Studien bereits mit normativen Vorbehalten an den Forschungsgegenstand herantritt. Entsprechend inkonsistent ist die Befundlage. Gut belegt sind beispielsweise negative Konsequenzen des sog. *minimum competency testing* in den 1980er Jahren auf Unterricht und Schülermotivation (z. B. Stecher, 2002). Im Widerspruch hierzu stehen Befunde zu positiven Effekten zentraler Prüfungen auf Schülerleistungen (z. B. Bishop, 1995) oder Befunde zu positiven „washback"-Effekten auf eine zielkonforme Veränderung von Unterricht (z. B. Cheng & Curtis, 2004). Insgesamt wird die Literatur jedoch von kritischen Stimmen, die vor negativen pädagogischen Konsequenzen eines zu hohen Rechenschaftsdrucks warnen, dominiert (z. B. Amrein & Berliner, 2003; Darling-Hammond, 2004; Hursh, 2007).

Ein forschungsmethodisches Problem ist, dass viele empirische Studien, die sich den Effekten von „school accountability" bzw. „high stakes testing" widmen, sog. „one shot case"-Studien (Cook & Campbell, 1976) sind, d. h. weder eine längsschnittliche noch eine experimentelle Vergleichsperspektive haben. Damit werden sowohl die interne als auch die externe Validität der Forschungsbefunde eingeschränkt (Campbell & Stanley, 1963; Shadish, Cook & Campbell, 2002). Es muss mit Interaktionen zwischen den Instrumenten testbasierter Rechenschaftslegung als Treatment und weiteren bildungspolitischen Reformen gerechnet werden („multiple treatment interference"). Ein weiteres Problem für die Validität kleinerer Studien sind Stichprobenselektionseffekte und die Reaktivität der Probanden auf bildungspolitisch brisante Forschungsfragen.

Um die Validität von Studien zu Effekten bildungspolitischer Maßnahmen zu sichern, müssten Experimentalstudien mit Vor- und Nachtest, sowie randomisierten Versuchsgruppen durchgeführt werden. Dieses Forschungsdesign ist zur Überprüfung bildungspolitischer Maßnahmen so gut wie ausgeschlossen. Es gibt lediglich eine in Irland durchgeführte Experimentalstudie, deren externe Validität allerdings zweifelhaft ist (Kellaghan, Madaus & Airasian, 1982). In der „washback"-Forschung, die der Frage nach Effekten spezieller Testreformen auf Unterricht nachgeht, findet man zwar vereinzelt längsschnittliche Vergleiche.

Beispielsweise untersuchte Cheng (2003) in einem „one group pre-posttest design“ Unterrichtsveränderungen nach einer Reform eines zentralen Tests in Hong Kong. Aufgrund der sehr kleinen Stichprobe ist auch hier die externe Validität gering. Zudem beantworten Studien, die Effekte testbasierter Schulreformen auf Unterrichtsprozesse untersuchen, noch lange nicht die Frage nach den Effekten auf Schülerleistungen.

Eine forschungsmethodisch interessante Alternative sind deshalb quasi-experimentelle Ländervergleiche. Die USA sind hierfür ein prädestiniertes Forschungsfeld, weil nationale „large scale assessments“ mit langen Datenreihen vorliegen (z. B. *National Assessment of Educational Progress*: NAEP, *National Educational Longitudinal Study*: NELS) und die Einführung testbasierter Rechenschaftslegung von Bundesstaat zu Bundesstaat sehr stark variiert. Lee (2008) legte zu diesem Typ ländervergleichender Studien eine Metaanalyse vor. Für die Berechnung eines gepoolten Treatmenteffekts von Rechenschaftsdruck werden allerdings sämtliche Studien gleich behandelt, obwohl es deutliche Qualitätsdifferenzen gibt. Ebenfalls wird darauf verzichtet, zwischen den verschiedenen Typen von testbasiertem Rechenschaftsdruck zu differenzieren. Quantitative Metaanalysen sind jedoch nur dann sinnvoll, wenn die Experimentalbedingung bzw. die unabhängige Variable relativ einheitlich operationalisiert und vergleichbare Effekte (abhängige Variable) untersucht werden können (Davies, 2000).

Aus diesen Gründen wird in dieser Arbeit eine narrative Forschungsbefundsynthese erstellt, die auf einer systematischen Literaturrecherche basiert und in der die Studien sowohl hinsichtlich ihres Forschungsdesigns (threats-to-validity) als auch hinsichtlich der Form testbasierter Rechenschaftslegung operationalisiert werden. Im folgenden Abschnitt wird der Begriff testbasierter Rechenschaftsdruck definiert und genauer eingegrenzt, sowie zwei theoretische Modelle zur Erklärung von Effekten von Rechenschaftsdruck eingeführt, die zur Einordnung der Forschungsbefunde herangezogen werden.

3. Vorgehensweise bei der systematischen, narrativen Forschungsbefundsynthese

Grundlage dieser Literaturübersicht ist eine systematische Literaturrecherche. Zunächst wurden einschlägige, in der Regel englischsprachige Übersichtsartikel zu den Effekten testbasierter Rechenschaftslegung, speziell zu den Auswirkungen auf Schule, Unterricht und Schülerleistungen gesichtet (z. B. Cheng & Curtis, 2004; Herman, 2004; Rea-Dickins & Scott, 2007; Stecher, 2002; Watanabe, 2004). Von besonderer Relevanz war die quantitative Metaanalyse von Lee (2008), in der US-Vergleichsstudien zu Effekten testbasierter Rechenschaftslegung auf Schülerleistungen synthetisiert wurden. 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 2010 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.

Eine große Menge an Suchergebnissen ließ sich durch die Festlegung von Selektionskriterien sehr schnell auf eine überschaubare Menge an Studien reduzieren. Genauer analysiert wurden Studien, die testbasierten Rechenschaftsdruck als unabhängige Variable in einem Staaten- oder Ländervergleich quasi-experimentell variieren und Effekte auf Schülerleistungen als abhängige Variable untersuchen. Ausgeschlossen wurde eine große Anzahl von Studien, in denen z. B. negative Effekte von „high stakes testing“ auf Unterricht oder auch positive „washback“-Effekte auf Schülerlernen untersucht wurden. Ebenfalls ausgeschlossen wurden Studien zu Effekten testbasierter Rechenschaftslegung in nur einem Bundesstaat (z. B. Finnigan & Gross, 2007; Parke, Lane & Stone, 2006) oder ländervergleichende Studien, die Effekte auf Lehrereinschätzungen (z. B. Opfer, Henry & Mashburn, 2008) oder Unterrichtsmethoden untersuchen (Firestone, Mayrowitz & Fairman, 1998; Firestone, Winter & Fitz, 2000; McDonnell & Choisser, 1997; Vogler, 2008). Diese Studien geben zwar wertvolle Einblicke in die Mikroprozesse der Umsetzung testbasierter Schulreformen. Es gibt jedoch eine Fülle weiterer, eher qualitativer Studien zu dieser Forschungsfrage, die keinem ländervergleichenden Ansatz folgen und hier nicht berücksichtigt werden können.

Die ausgewählten Studien bzw. Teilstudien wurden entlang folgender Fragen analysiert (siehe Tabelle 1 und 2 im Anhang):

- Wie wurde testbasierter Rechenschaftsdruck operationalisiert („student accountability“ vs. „school accountability“)?
- Wie wurde Schulleistung operationalisiert und welche weiteren abhängigen Variablen werden in der Studie untersucht?
- Welches Forschungsdesign wurde gewählt (Quasi-experimentelle Gruppenvergleiche mit oder ohne Pretest vs. Regressions- oder Kovarianzanalysen)?
- Welche Kontextmerkmale bzw. welche relevanten Prädiktoren für Schulleistung wurden in den Analysen auf welchen Schulsystemebenen berücksichtigt?

Die Analyse des Forschungsdesigns und der berücksichtigten Kontextmerkmale bilden die Grundlage für die Bewertung der Validität der Ergebnisse. Abschließend wurden die von den jeweiligen Studien berichteten Effekte testbasierter Rechenschaftslegung auf Schülerleistungen aufgelistet.

4. Ergebnisse der Forschungsbefundsynthese

Insgesamt wurden 20 Studien mit einem ländervergleichenden Ansatz zu Effekten testbasierter Rechenschaftslegung auf Schülerleistungen gefunden. Größere Publikationen enthielten mehrere Teilstudien, die gesondert betrachtet wurden (z. B. Bishop, 1995). Ebenso wurden Publikationen zusammengefasst, wenn in mehreren Aufsätzen die gleichen Befunde zu Effekten testbasierter Rechenschaftslegung erwähnt wurden (z. B. Wößmann, 2007, 2008; Wößmann & Fuchs, 2007). Die Studien ließen sich entlang der Operationalisierung von testbasierter Rechenschaftslegung in zwei Gruppen teilen. Die eine Hälfte der Studien (siehe Tabelle 1 im Anhang) beschäftigte sich mit den Auswirkungen zentraler, curriculumbasierter Abschlussprüfungen in der Sekundarstufe II („student accountability"). Die andere Hälfte der Studien (siehe Tabelle 2 im Anhang) basiert auf Indizes für testbasierten Rechenschaftsdruck als unabhängige Variable für Regressionsanalysen oder quasi-experimentellen Gruppenbildungen. Diese Indexvariablen erfassen sowohl Instrumente zu „student accountability" als auch zu „school accountability".

4.1 Effekte von testbasiertem Rechenschaftsdruck auf Schülerebene

Eine erste US-Vergleichsstudie mit positiven Effekten von „high school graduation exams" auf NAEP-Daten in den 1980er Jahren legte Frederikson (1994) vor. Maßgebend in diesem Bereich sind allerdings die bildungsökonomischen Analysen von Bishop (1995, 1998), der in einer Serie von Teilstudien (USA, Kanada, international) der Frage nachgeht, wie sich curriculumbasierte, zentrale Abschlussprüfungen am Ende der Sekundarstufe II auf Schülerleistungen auswirken. Er fokussierte die Effekte von sog. „curriculum-embedded high school graduation exams", d. h. zentrale Tests am Ende der Highschool oder nach einzelnen Kursen, deren Bestehen über den Schulabschluss entscheidet und die sich direkt auf das gelehrte Curriculum beziehen. Um sein bildungsökonomisches Modell prüfen zu können, untersucht Bishop Effekte auf zwei Gruppen von abhängigen Variablen: Schülerleistungen (internationale „large scale assessments" der 1990er Jahre und nationale Leistungsvergleiche in den USA) und Anstrengungen, die das Bildungssystem unternimmt, damit Schüler die zentralen Abschlusstests bestehen können (z. B. Anforderungen an den Lehrerberuf, Lehrergehälter, Ausstattung von Schulen, Bildungsausgaben pro Schüler, etc.).

Die Befunde der Forschungsserie von Bishop bestätigen allesamt sehr eindrücklich seine theoretischen Annahmen. In Zeiten bzw. Regionen, in denen die Zulassung zum College stark von den Leistungen in der Highschool abhängt, wählen Schüler anspruchsvollere Kurse. In Ländern bzw. Staaten mit zentralen, curriculumbasierten Abschlussprüfungen in der Sekundarstufe II sind Mindeststandards

für das Lehramt höher, werden höhere Lehrergehälter gezahlt, erzielen die Schüler höhere Werte in „large scale assessments“, sind Bildungsausgaben und Schulausstattung im Highschool-Bereich höher, werden eher fachlich ausgebildete Mathematik- und Naturwissenschaftslehrkräfte beschäftigt und an den Schulen herrscht eher ein leistungsförderliches Lernklima (mehr Hausaufgaben, häufigere Leistungskontrollen, mehr Ansporn durch Eltern). Die Befunde werden in späteren Studien bestätigt. Beispielsweise nutzt auch Rindermann (2008) die Datensätze internationaler „large scale assessments“ (TIMSS 1995–2003, PISA 2000, 2003, PIRLS 2001, IGLU 2001), um Effekte von Bildungssystemmerkmalen auf Schülerleistungen zu schätzen. Auch er findet einen positiven Zusammenhang zwischen Schulleistung und Zentralprüfungen, der sich bei der Herauspartialisierung weiterer Schulsystem- oder Ländermerkmale (z. B. Bruttoinlandsprodukt) sogar noch erhöht.

Einige Teilstudien von Bishop müssen jedoch aufgrund des Forschungsdesigns relativiert werden. Bishop führt in der Regel Zwei-Gruppen-Vergleiche ohne Vortest durch. Auch relevante Kontextmerkmale der Schulsysteme (z. B. Bildungsausgaben, Bildungsniveau der Bevölkerung, etc.) werden nicht in allen Teilstudien systematisch berücksichtigt. Ebenso werden keine Interaktionen mit weiteren Merkmalen testbasierter Rechenschaftslegung (z. B. Berichterstattung von Daten auf Schulebene, Sanktionen und Unterstützungsmaßnahmen auf Schulebene) erfasst. Einige Studien haben zudem eher den Charakter von Fallstudien (New York State vs. US benchmarks) mit Analysen auf einem rein deskriptiven Niveau. Noch relativ überzeugend ist der kanadische Provinzen-Vergleich (Bishop, 1995). In dieser Studie sind die beiden Gruppen mit und ohne zentrale Abschlussexamina ähnlich umfangreich, es handelt sich um einen vergleichbaren kulturellen Kontext und die elterliche Lesekompetenz wird kontrolliert.

Die internationalen Ländervergleiche auf der Basis von TIMSS- und PISA-Datensätzen von Wößmann (2007, 2008) sowie Wößmann und Fuchs (2007) schließen an die bildungsökonomische Theorietradition von Bishop an und differenzieren darüber hinaus zwischen zwei Typen von „student accountability“, die mit unterschiedlich hohem Rechenschaftsdruck assoziiert sind: zentrale Abschlussexamina für die Sekundarstufe II und regelmäßige zentrale Tests in der Sekundarstufe I und II. Zudem werden die Analysen auf Individualebene gerechnet, womit sich die länderinterne Varianz mit der Varianz zwischen den Ländern vergleichen lässt. Ein weiterer Vorteil des PISA-Datensatzes ist, dass eine Fülle von individuellen, familialen und schulischen Kontextvariablen in den Regressionsanalysen mitberücksichtigt werden können. Von Nachteil ist allerdings, dass z. B. PISA nur Querschnittsdaten zur Verfügung stellt und somit keine Effekte auf individuelle Leistungsentwicklungen geprüft werden können.

Die Bishop-Studien werden grundsätzlich bestätigt. Schüler in Schulsystemen mit zentralen Abschlussprüfungen in Mathematik schneiden in den Mathematiktests der „large scale assessments“ besser ab. Dagegen haben zentrale Tests für sich allein genommen keinen signifikanten Einfluss auf die Schülerleistungen. In Schulsystemen ohne zentrale Abschlussprüfungen ist der Einfluss stan-

dardisierter Tests auf die Schulleistung sogar signifikant negativ. Mit zentralen Abschlussexamen 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. Diese Interaktionseffekte gelten auch für das Schulsystemmerkmal Schulautonomie. Bei erhöhter Schulautonomie sind die Zusammenhänge zwischen zentralem Testen und Schülerleistungen höher.

Auch die Studien von Muller und Schiller (2000) sowie Jacob (2001) liefern ähnlich valide Befunde wie Wößmann und Fuchs und damit eine wertvolle Ergänzung zu den von Bishop durchgeführten US-internen Vergleichen. Beide Studien untersuchen Effekte von zentralen „high school graduation exams" auf Leistungsentwicklung anhand echter US-Längsschnittstudien (National Education Longitudinal Study: NELS 88–92, National Longitudinal Study of Schools: NLSS). In beiden Studien werden sowohl relevante Kontextmerkmale der beteiligten US-Staaten als auch Prädiktorvariablen auf Individualebene (SES, race, gender, etc.) berücksichtigt. Bei Jacob (2001) zusätzlich noch Kontextmerkmale auf Schulebene (Sozioökonomischer Status der Familien im Einzugsgebiet einer Schule).

Muller und Schiller (2000) kategorisieren die testbasierten Schulreformen aller 50 US-Staaten nach Testhäufigkeit (Wie oft und in wie vielen Fächern wurden Highschool-Schüler 1994 extern getestet?) und Anzahl bedeutsamer Konsequenzen für Schüler und Schulen. Untersucht werden nicht nur Effekte auf Schülerleistungszuwächse anhand echter Längsschnittstudien (NELS, NLSS), sondern auch auf die „gatekeeping"-Funktion von Lehrkräften. Dabei zeigte sich, dass in Staaten mit relevanten Testkonsequenzen für Schüler (Versetzung, Abschluss) und Schulen (Sanktionen) die Leistungszuwächse höher sind. Allerdings steigt auch die soziale Selektivität (z. B. „black-white achievement gap") bei hohen Testkonsequenzen. Diese Befunde müssten eigentlich in die nachfolgend zu besprechende Gruppe von Studien eingeordnet werden (siehe Tabelle 2 im Anhang). Von Interesse für die Effekte von „student accountability" ist allerdings das Ergebnis, dass die Korrelation zwischen der Einschätzung von Middleschool-Lehrkräften ob ein Schüler den Highschool-Abschluss schafft und dem tatsächlichen Erreichen des Highschool-Abschlusses durch häufiges Testen sinkt. Dies bedeutet, dass häufiges Testen in der Highschool unabhängig von den damit verbundenen Konsequenzen die soziale Selektivität des Lehrerurteils reduziert.

Jacob (2001) untersuchte, ob sich zentrale Abschlusstests zur Überprüfung von Mindestkompetenzen in Lesen und Mathematik (sog. „minimum competency testing in basic skills") auf längsschnittlich gemessene Schülerleistungszuwächse und Nichtversetzungsquoten („drop out rates") zwischen 1988 und 1992 auswirken. Die Analysen der Rohdaten zeigen, dass Staaten mit diesen Tests eine eher bildungsbenachteiligtere Bevölkerung haben, die Lernzuwächse in den Längsschnittstudien jedoch gleich groß sind wie in der Kontrollgruppe ohne Tests. Wenn sämtliche Kontextvariablen in der Regressionsanalyse berücksichtigt werden, ergibt sich kein signifikanter Effekt von „minimum competency testing" auf die Schülerleistungen in Jahrgangsstufe 12. In Staaten ohne Tests verzeichnen jedoch die leistungs-

schwächsten Schüler (unteres Dezentil) einen höheren Zuwachs an Lesekompetenz und haben eine geringere Nichtversetzungsquote.

4.2 Effekte von testbasiertem Rechenschaftsdruck auf Schüler- und Schulebene

In einer zweiten Gruppe können ländervergleichende Studien zusammengefasst werden, die mit Indizes für testbasierten Rechenschaftsdruck auf Schüler- und Schulebene als unabhängige Variable oder Gruppierungskriterium arbeiten (siehe Tabelle 2 im Anhang). Diese Studien haben den Vorteil, dass der in einem US-Bundesstaat aufgebaute Rechenschaftsdruck möglichst umfassend abgebildet wird. Die Indexvariablen bzw. die Gruppierungen berücksichtigen sowohl den Umfang zentralen Testens auf allen Schulstufen als auch die damit für Schüler, Lehrer und Schulen verknüpften Konsequenzen. Es wird somit ein Typ von testbasierter Rechenschaftslegung erfasst, der durch die NCLB-Reform intendiert ist und in zahlreichen Bundesstaaten bereits vor 2002 existierte. Im Gegenzug entsteht der Nachteil, dass diese Studien keine Aussagen über Effekte einzelner Rechenschaftsinstrumente (Testhäufigkeit, Tests auf bestimmten Schulstufen, Belohnungssysteme, Rückmeldeformate, etc.) erlauben.

Zunächst gibt es eine Serie von Studien, die ihren Ausgangspunkt in der von Amrein und Berliner (2002) eingeführten Einteilung in „low-“ und „high stakes testing“-Staaten haben (Amrein & Berliner, 2003; Amrein-Beardsley & Berliner, 2003; Braun, 2004; Rosenshine, 2003). Ein US-Bundesstaat wird der Gruppe der „high stakes testing“-Staaten zugeordnet, wenn die staatlichen Schulleistungstests in den 1990er Jahren mit mindestens einer von bis zu sechs Konsequenzen verknüpft sind (z. B.: Veröffentlichung der Ergebnisse, Abschlüsse abhängig von zentralem Test, reale Konsequenzen für Schulen wie Zuschüsse oder Schließung, finanzielle Anreize für Lehrer). Amrein und Berliner (2002) nutzen in ihrer ersten Studie keine Kontrollgruppe, berücksichtigen keine Kontextvariablen und führen keine inferenzstatistischen Analysen durch. Auf der Grundlage nationaler Leistungsstudien (NAEP: 1996–2000 für Jgs. 4 und 8; „*American College Test*“ ACT und „*Scholastic Aptitude Test*“ SAT für Leistungen auf Highschool-Niveau) werden die Leistungswerte von 18 „high stakes testing“-Staaten mit den entsprechenden US-Mittelwerten deskriptiv verglichen. Das Fazit dieser Analysen ist aus Sicht von Amrein und Berliner (2002) ernüchternd: Die „highstakes testing“(HST)-Staaten verzeichnen einen Rückgang der ACT und SAT-Werte in den 1990er Jahren. Als problematisch wird auch beschrieben, dass die Teilnahme am ACT als Indikator für Interesse der Highschool-Schüler am Studium lediglich in 9 von 18 HST-Staaten steigt. Weniger als die Hälfte der HST-Staaten verzeichneten quasi-längsschnittliche Zuwächse in den NAEP-Mathematiktests. Verbesserungen der HST-Staaten im NAEP-Lesetest führen die Autoren auf Interaktionen mit nationalen Anstrengungen zur Verbesserung der Lesekompetenz zurück.

Rosenshine (2003) reagierte mit einer kritischen Reanalyse dieser Studie durch Hinzunahme einer Kontrollgruppe von 15 „low stakes testing"-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. Aus Anlass der Kritik wiederholten Amrein-Beardsley und Berliner (2003) einen Teil ihrer ersten Studie mit der von Rosenshine (2003) eingeführten Kontrollgruppe. Die Autoren bestätigen die Befunde von Rosenshine (2003), argumentieren allerdings, dass HST-Staaten auch eine signifikant höhere NAEP-Ausschlussrate haben. Die NAEP-Gewinne der HST-Staaten werden von Amrein-Beardsley und Berliner (2003) als Verluste interpretiert, da sie ausschließlich auf Schulausschlüsse zurückzuführen sind. Trotz dieser methodischen Verbesserung bleiben forschungsmethodische Unklarheiten. Die beiden quasi-experimentellen Untersuchungsgruppen umfassen längst nicht alle 50 US-Staaten und auch bei Rosenshine wird die genaue Auswahl und Zuordnung zu den Gruppen nicht exakt definiert. Zudem bleiben relevante Kontextvariablen auf Staaten- und Schulebene komplett unberücksichtigt.

Auch Braun (2004) reanalysierte einen Teil der Befunde von Amrein und Berliner (2002) und nutzte als Indikator für Schülerleistungen die NAEP-Mathematiktests 1992, 1996 und 2000 in den Jahrgangsstufen 4 und 8. Im Vergleich zu Amrein und Berliner (2002) sowie Rosenshine (2003) werden die testbasierten Schulreformen aller US-Staaten genauer analysiert. Da jedoch nicht alle Staaten im betreffenden Zeitraum an allen NAEP-Erhebungen teilgenommen haben, sind für Braun (2004) Ländervergleiche zwischen 15 von 18 als HST klassifizierte Staaten und 18 von 32 als LST klassifizierte Staaten möglich. Ebenfalls im Gegensatz zu den beiden Vorstudien berücksichtigt Braun (2004) den Standardfehler der NAEP-Zuwachsraten für einzelne Bundesstaaten. Es zeigt sich, dass die HST-Staaten bei den querschnittlichen Vergleichen besser abschneiden, auch wenn der prozentuale Anteil der vom Test ausgeschlossenen Schüler kontrolliert wird. Im quasi-längsschnittlichen Vergleich einzelner NAEP-Kohorten (z. B. 1992–1996) schneiden dagegen die LST-Staaten besser ab. Braun (2004) gibt allerdings zu bedenken, dass dieser Effekt durch die nicht regelmäßige Teilnahme von Staaten an einzelnen NAEP-Tests im Zeitraum 1992–2000 zustande kommen könnte.

Eine zweite Serie von Studien greift auf den hoch aggregierten und zum Teil interpretativen „index for accountability pressure" des *Consortium of Policy Research in Education* (CPRE; Goertz & Duffy, 2001) als unabhängige Variable zurück. Der Index basiert auf einer Analyse der von einzelnen Staaten implementierten Rechenschaftsinstrumente. Staaten mit keinen zentralen Testinstrumenten (z. B. Iowa und Nebraska vor NCLB) erhalten eine 0. Wird im Grund- und Sekundarschulwesen lediglich zentral getestet, ohne dass Konsequenzen folgen, erhält ein Staat die 1. Je nach Testkonsequenzen für Schüler, Lehrer und Schulen sind Indexwerte von 2 bis 5 möglich. Der Wert 5 wird vergeben, wenn sowohl für Schüler als auch für Schulen wesentliche Entscheidungen von den staatlichen Testresultaten abhängen.

Carnoy und Loeb (2002) nutzen erstmals diesen kombinierten Index, um in Regressionsanalysen den Effekt testbasierter Rechenschaftslegung auf quasi-längsschnittliche NAEP-Zuwächse und Versetzungsquoten in der Highschool („high school progression rates") zu berechnen. In den Analysen werden relevante Variablen auf Ebene des Bundesstaats kontrolliert (z. B. ethnische Zusammensetzung, Durchschnittseinkommen). Carnoy und Loeb (2002) berichten positive Zusammenhänge zwischen höheren Indexwerten und den NAEP-Mathematiktests in Jahrgangsstufe 8 und dies sowohl im oberen als auch im unteren Leistungsbereich. Dagegen finden sie keine Zusammenhänge mit den NAEP-Mathematiktests in Jahrgangsstufe 4 und den Highschool-Versetzungsquoten als Indikator für den Langzeiteffekt von testbasierter Rechenschaftslegung. Die Befunde weisen darauf hin, dass sich neue Formen von „school accountability" vor allem auf die Sekundarstufe I auswirken. Der nicht vorhandene Effekt auf Highschool-Versetzungsquoten sollte nach Carnoy und Loeb (2002) relativiert werden, weil Daten zu „high school progression rates" immer mit der Bevölkerungsentwicklung eines Staates konfundiert sind und somit mit nicht kontrollierbaren Interaktionseffekten zu rechnen ist.

Carnoy (2005) berechnete in weiteren Regressionsanalysen die Zusammenhänge zwischen Rechenschaftsdruck in 45 US-Staaten und Highschool-Abschlussquoten bzw. Highschool-Versetzungsquoten als Indikatoren für langfristigen Bildungserfolg. Kontrolliert wurden relevante Prädiktoren auf Schulsystemebene, wie die ethnische Zusammensetzung und Größe des Staates, die Bildungsausgaben pro Schüler sowie Versetzungsquoten vorausgehender Schuljahre. Es zeigten sich keine bzw. schwach negative Zusammenhänge zwischen testbasiertem Rechenschaftsdruck und Versetzungsquoten, d. h. der Befund von Carnoy und Loeb (2002) wurde bestätigt.

Auch Lee und Wong (2004) sortierten alle US-Staaten aufgrund der Indexvariable des CPRE in drei Kategorien und untersuchten mit Mehrebenenanalysen deren Einfluss auf die quasi-längsschnittlichen NAEP-Lernzuwächse in den Jahren 1990 bis 2000. Die Studie stellt einen weiteren forschungsmethodischen Fortschritt dar. In Mehrebenenanalysen wurden sozioökonomische Variablen (ethnische Zusammensetzung, Durchschnittseinkommen) und Bildungsausgaben (1990–2000) sowohl auf Ebene des Distrikts bzw. der Schule (Ebene 1) als auch der Bundesstaaten (Ebene 2) kontrolliert. Lee und Wong (2004) fanden keine Effekte von testbasierter Rechenschaftslegung auf Zusammenhänge zwischen Einkommen, Minderheitenstatus und Schulleistung, d. h. der in den USA als Indikator für soziale Selektivität genutzte „achievement gap" zwischen weißen Schülern und Minderheitenschülern wird durch Rechenschaftslegung weder positiv noch negativ beeinflusst. Ebenfalls fanden die Autoren keine Effekte von Rechenschaftslegung auf die Veränderung der Bildungsausgaben (Klassengröße, Lehrerqualifikation, Ausgaben pro Schüler).

In einem weiteren Schritt berücksichtigt Lee (2006) auch die staatlichen Inputvorgaben als unabhängige Variable und klassifiziert die US-Staaten entlang von zwei Dimensionen: (1) ergebnisorientierter Rechenschaftsdruck und

(2) Ressourcengarantie für Schulen (Ausgaben pro Schüler, Klassengröße, Lehrerqualifikation). Analog zu Lee und Wong (2004) werden Effekte auf quasilängsschnittlichen Lernzuwachs in der Sekundarstufe (NAEP 1990–2000) berechnet und relevante Faktoren auf beiden Ebenen des Mehrebenenmodells kontrolliert. Lee (2006) findet keinen Zusammenhang zwischen den beiden unabhängigen Variablen: testbasiertem Rechenschaftsdruck und Ressourcengarantie für Einzelschulen durch den Staat. Es gibt allerdings einen Zusammenhang zwischen Ressourcengarantie und Schülerleistungen, jedoch keine Zusammenhänge zwischen reinem Rechenschaftsdruck und Schülerleistungen bzw. Erhöhung der sozialen Disparitäten. Von besonderem Interesse ist der positive Interaktionseffekt zwischen testbasiertem Rechenschaftsdruck und Ressourcengarantie auf Schülerleistungszuwächse.

Sowohl Hanushek und Raymond (2005) als auch Dee und Jacob (2009) greifen auf die Gruppierungen bzw. Indizes der vorausgehenden Studien zurück, verschränken jedoch den Ländervergleich mit einer längsschnittlichen Betrachtung. D. h. es wird berücksichtigt, zu welchem Zeitpunkt ein Staat den testbasierten Rechenschaftsdruck auf Schüler, Schulen und Lehrer erhöht hat. Analog zu Amrein und Berliner (2002) arbeiten Hanushek und Raymond (2005) mit einer sehr groben Einteilung der US-Staaten in sog. „report card states" (Schulleistungsdaten werden lediglich publiziert) und „consequential states", in denen bereits in den 1990er Jahren Schulen mit realen Testkonsequenzen rechnen mussten. Hanushek und Raymond (2005) schätzen mit ihren Regressionsgleichungen sowohl den querschnittlichen Effekt von Rechenschaftslegung im Vergleich zu Staaten ohne Rechenschaftslegung als auch die Differenz vor und nach Einführung des Rechenschaftssystems. Berücksichtigt werden die Veränderungen bei NAEP-Testausschlüssen (z. B. „special needs students") in einzelnen Staaten und die ethnische Zusammensetzung der Schülerschaft.

Die Analysen von Hanushek und Raymond (2005) zeigen, dass Staaten, die während der 1990er Jahre „consequential accountability systems" eingeführt haben, größere NAEP-Leistungszuwächse haben. Für „report card"-Staaten unterscheiden sich die Zuwachsraten nicht signifikant von Null. NAEP-Testausschlüsse von „special needs students" wirken sich immer positiv auf Testwertsteigerungen aus. Allerdings gibt es keinen systematischen Zusammenhang zwischen der Einführung von testbasierter Rechenschaftslegung und Erhöhung der Ausschlussraten. Nach ethnischen Gruppen ergeben sich höhere Effekte von „consequential accountability" auf die Leistungszuwächse von Hispanics und weißen Schülern. Die Lücke zwischen weißen und schwarzen Schülern („black-white achievement gap") wird durch die Einführung von „consequential accountability" allerdings noch vergrößert. Hanushek und Raymond (2005) verfeinern ihre Analyse weiter, indem sie verschiedene qualitative Bewertungen der staatlichen Bildungsstandards als Grundlage der testbasierten Rechenschaftslegung in Indexvariablen überführen und in ihre Regressionsanalysen einfügen. Die Ergebnisse zeigen zumindest der Tendenz nach, dass „consequential accountability" Staaten mit anspruchsvol-

leren Bildungsstandards (z. B. North Carolina, Alabama, California) höhere NAEP-Zuwachsraten haben.

Dee und Jacob (2009) beanspruchen sogar, mit ihren Analysen den Effekt der NCLB-Reform abschätzen zu können. Hierfür nutzen sie die jüngsten NAEP-Datenreihen (2003, 2005, 2007) nach der Einführung von NCLB im Jahr 2002. Weiterhin teilen sie die 39 US-Staaten, für die vollständige NAEP-Datenreihen vorliegen, in „early vs. late adopters of accountability" auf Basis der bekannten Kategorisierungen (Carnoy & Loeb, 2002; Hanushek & Raymond, 2005; Lee & Wong, 2004). Für eine längsschnittliche und quasi-experimentelle Perspektive werden nur Staaten betrachtet, die mindestens zweimal vor Einführung von NCLB an den NAEP-Tests teilgenommen haben. Mit einem quasi-experimentellen Zeitreihenvergleich können die Autoren berechnen, wie sich die Testwerte der Staaten, die vor NCLB noch keinen hohen Rechenschaftsdruck hatten (Quasi-Experimentalgruppe) im Vergleich zu den Testwerten der Staaten, die vor NCLB bereits Testsysteme mit hohem Rechenschaftsdruck hatten (Quasi-Kontrollgruppe), ändern.

Obwohl Dee und Jacob (2009) NAEP-Daten des darauf folgenden Jahrzehnts untersuchen, decken sich die Befunde mit denen von Hanushek und Raymond (2005) weitestgehend. Die Bundesstaaten in der Quasi-Experimentalgruppe („late adopters") verzeichnen signifikante Leistungssteigerungen im NAEP Mathematiktest für Jahrgangsstufe 4, vor allem bei weißen Schülern, Hispanics und sozioökonomisch benachteiligten Schülern. Ebenfalls ergaben sich moderat positive, signifikante Effekte auf den NAEP-Mathematiktest für Jahrgangsstufe 8, vor allem bei sozioökonomisch benachteiligten Schülern und Schülern mit geringem Leistungsniveau. Der Zeitreihenvergleich konnte jedoch keine Effekte von erhöhtem Rechenschaftsdruck auf die Leseleistungen nachweisen, weder in Jahrgangsstufe 4 noch in Jahrgangsstufe 8. Dee und Jacob (2009) gehen davon aus, dass die Effekte von Rechenschaftsdruck auf Einzelschulen in diesen Analysen eventuell unterschätzt wurden. Es ist unklar, inwiefern schulische Kompositionseffekte die Resultate des Zeitreihenvergleichs beeinträchtigen. Zum Beispiel könnte es durchaus sein, dass durch NCLB bestimmte Schulen stärker öffentlich angeprangert werden und Eltern ihre Kinder eher auf Privatschulen schicken können. Dieser Effekt kann nicht kontrolliert werden und könnte zu einer Unterschätzung der durch NCLB erzeugten Leistungseffekte führen.

5. Zusammenfassung und Diskussion

Zusammenfassend lässt sich sagen, dass die konsistent positiven Effekte zentraler Abschlussexamina in den Studien von Bishop durch methodisch zuverlässigere Studien relativiert, jedoch nicht grundsätzlich revidiert werden müssen. Wößmann (2007, 2008) sowie Wößmann und Fuchs (2007) zeigen, dass die Effekte zentraler Abschlussexamina auf Schülerleistungen im Zusammenhang mit anderen

Schulsystemmerkmalen (z. B. Schulautonomie) zu sehen sind. Methodisch überzeugend sind die Studien von Muller und Schiller (2000) sowie Jacob (2001). Beide Vergleiche basieren auf echten Längsschnittdatensätzen aller US-Staaten. Die Studien zeigen, dass zentrale Prüfungen in der Sekundarstufe II differenziert zu betrachten sind. Bei relevanten Testkonsequenzen für Schüler und Schulen steigen die Leistungen, allerdings steigt auch die soziale Selektivität. Häufigeres Testen konnte dagegen die soziale Selektivität des Lehrerurteils in der Sekundarstufe I reduzieren. Jacob (2001) zeigte, dass die Prüfung von Mindeststandards in der Highschool bestenfalls einen neutralen Effekt hat, d. h. wirkungslos ist. Dieser Befund lässt sich ebenfalls mit dem bildungsökonomischen Modell erklären: Nur wenn anspruchsvolle Standards geprüft werden (Hanushek & Raymond, 2005) und Schüler mit Konsequenzen rechnen müssen, steigt die Lernmotivation und damit die Schulleistung.

Problematisch scheint dagegen Hypothese 2 im bildungsökonomischen Modell von Bishop (1995) zu sein: Staaten mit zentralen Abschlussprüfungen investieren mehr in die Ausstattung der Schulen und die Qualifikation der Lehrkräfte. Gerade dieser Teil der Theorie wird durch neuere Befunde widerlegt. Lee (2006) zeigte, dass Ressourcengarantie als Input und hoher Rechenschaftsdruck als Outputkontrolle nicht miteinander korrelieren. Dagegen ist die Ressourcengarantie eine Voraussetzung für die effektive Umsetzung testbasierter Schulreformen. Es spricht also einiges dafür, die bildungsökonomische Kausalkette umgekehrt zu lesen: Weil bestimmte Länder hohen Wert auf Bildung legen, investieren sie in Bildung, erwarten hohe Anforderungen von Lehramtskandidaten, statten ihre Schulen gut aus und verlangen von den Schülern über zentrale Examina auch entsprechende Leistungen. Damit wären die Befunde von Bishop gerade ein Hinweis darauf, dass nicht die zentrale Abschlussprüfung allein, sondern eher die Kombination von Bildungsausgaben und zentralen Tests zu überdurchschnittlichen Leistungen in internationalen Vergleichsstudien führt.

Die Studien mit einer groben Einteilung in HST-Staaten und LST-Staaten produzieren inkonsistente und nur eingeschränkt valide Ergebnisse. Vor allem die kritischen Analysen von Amrein und Berliner (2002) eignen sich kaum, um die bildungsökonomische Theorie von Testeffekten auf Schülermotivation und Schülerleistungen in Frage zu stellen. Von besonderer Bedeutung sind jüngere Studien auf Basis von Indexvariablen, die aufgrund von Mehrebenenanalysen und der Berücksichtigung relevanter Prädiktorvariablen relativ valide zeigen, dass Rechenschaftsdruck auf Schüler- und Schulebene zu neutralen bis positiven Effekten auf Schülerleistungen führen kann. Diese positiven Effekte beschränken sich jedoch immer wieder auf einzelne Fächer (Mathematik) oder Schulstufen (z. B. Sekundarstufe). Die Hinweise auf eine Reduktion oder eine Erhöhung der sozialen Selektivität sind insgesamt uneinheitlich. Beide Befunde widersprechen nicht grundsätzlich dem bildungsökonomischen Erklärungsmodell, stützen es allerdings auch nicht mehr so konsistent und eindrücklich wie bei Bishop (1995).

Wie lassen sich die Befunde vor dem Hintergrund des system- und professionstheoretischen Modells von O'Day (2002, 2004) einordnen? Es gibt Hinweise,

dass testbasierte Schulreformen dann effektiv sind, wenn sie Informationen bereitstellen, die Lehrer und Schüler auf Unterricht und Lernen aufmerksam machen. Beispielsweise gilt dies in besonderem Maße für die von Bishop untersuchten curriculumbasierten zentralen Abschlussprüfungen. Die Prüfungen bezogen sich direkt auf den unterrichteten Lehrstoff und Lehrkräfte können die Testergebnisse direkt mit ihrem Unterricht in Verbindung setzen. Auch das Ergebnis von Muller und Schiller (2000), dass durch häufiges Testen die soziale Selektivität des Lehrerurteils reduziert wird, kann dahingehend interpretiert werden. Erst ab einer gewissen Dichte an externen Testrückmeldungen können Lehrkräfte pädagogisch relevante Prognosen ableiten.

Die hier dargestellten Studien geben allerdings keine Hinweise darauf, inwiefern Lehrkräfte durch testbasierte Schulreformen zum aktiven Umgang mit Testrückmeldungen, d. h. zu einer ergebnisorientierten Weiterentwicklung des eigenen Unterrichts angeregt wurden. Dies liegt zunächst daran, dass in dieser Arbeit lediglich Studien mit Schülerleistungen als abhängige Variable untersucht wurden. In der Literatur finden sich zahlreiche Studien, die genau dieser Frage nachgehen, allerdings nicht mit einem ländervergleichenden Ansatz arbeiten. Und wenn Effekte testbasierter Rechenschaftslegung auf Unterricht ländervergleichend untersucht werden, fehlt die Verknüpfung mit Leistungseffekten (z. B. Firestone, Mayrowitz & Fairman, 1998; Firestone, Winter & Fitz 2000; McDonnell & Choisser, 1997). In der deutschsprachigen Diskussion wird aber genau dieser Zusammenhang postuliert. Über zentrale Vergleichsarbeiten wird eine ergebnisorientierte Schul- und Unterrichtsentwicklung möglich, die sich dann positiv auf die Kompetenzentwicklung der Schüler niederschlägt. Für diese Kausalkette fehlen sowohl empirische Belege als auch praktikable Forschungsansätze.

Die hier besprochenen Forschungsarbeiten belegen dagegen sehr gut, dass leistungssteigernde Effekte testbasierter Rechenschaftslegung nur dann zu erwarten sind, wenn es Anreizstrukturen gibt, die eine informationsbasierte Ressourcenverteilung ermöglichen. Diese vierte Bedingung im Rahmenmodell nach O'Day (2002, 2004) wird gestützt durch Befunde von Wößmann und Fuchs (2007), die zeigen, dass in Schulsystemen mit höherer Schulautonomie auch die Effekte zentraler Tests und Abschlussprüfungen auf Schülerleistungen höher sind. Oder durch den Befund von Lee (2006), dass testbasierter Rechenschaftsdruck nur in Kombination mit staatlicher Ressourcengarantie positiv wirkt.

Wie lassen sich nun die Ergebnisse der narrativen Forschungsbefundsynthese mit den Entwicklungen im deutschsprachigen Raum in Verbindung bringen? Ausgangspunkt war die Überlegung, dass externer, testbasierter Rechenschaftsdruck ein konstitutives Element testbasierter Schulreformen ist und eine weitere Senkung des ohnehin nicht hohen Rechenschaftsdrucks auf Lehrer und Schulen zu einer Marginalisierung von Bildungsstandards, Vergleichsarbeiten und weiteren Monitoringinstrumenten führen würde. Welche empirisch begründeten Weiterentwicklungsvarianten wären also denkbar? Eine Übertragung der vor allem US-amerikanischen Befundlage auf den deutschen Kontext ist natürlich hoch spekulativ. Ungeachtet aller kulturellen und schulstrukturellen Differenzen sollte

dennoch festgehalten werden, dass es keine stichhaltigen empirischen Belege dafür gibt, dass sich testbasierter Rechenschaftsdruck negativ auf Schülerleistungen auswirkt oder soziale Disparitäten im Bildungssystem erhöht. Es gibt aber wohl empirische Belege dafür, dass nicht jede Form des testbasierten Rechenschaftsdrucks zu einer Steigerung der Bildungsqualität beiträgt. In Deutschland sollte man deshalb mit einer Erhöhung des testbasierten Rechenschaftsdrucks vorsichtig experimentieren, bevor zentrale Leistungsmessungen in den Schulen ad acta gelegt werden. Gleichzeitig müssen aber auch die Test- und Rückmeldeformate so weiterentwickelt werden, dass Lehrkräfte feinkörnige und für ihr Handeln relevante Daten zur Verfügung haben. Beide Theoriemodelle bieten zumindest einen groben Rahmen für diese Entwicklungsarbeit.

Literatur

Amrein, A.L. & Berliner, D.C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, *10*(18). Zugriff am 31.10.2002 unter http://epaa.asu.edu/epaa/v10n18/

Amrein, A.L. & Berliner, D.C. (2003). The effects of highstakes testing on student motivation and learning. *Educational Leadership*, *60*, 32–38.

Amrein-Beardsley, A.L. & Berliner, D.C. (2003). Re-analysis of NAEP math and reading scores in states with and without high-stakes tests: Responses to Rosenshine. *Education Policy Analysis Archives*, *11*(25). Zugriff am 24.08.2003 unter http://epaa.asu.edu/epaa/v11n25/

Baumert, J. (2001). Vergleichende Leistungsmessung im Bildungsbereich. In J. Oelkers (Hrsg.), *Zukunftsfragen der Bildung* (S. 13–36). Weinheim: Beltz. (Zeitschrift für Pädagogik, 43. Beiheft).

Bishop, J.H. (1995). The impact of curriculum-based external examinations on school priorities and student learning. *International Journal of Educational Research*, *23*, 653–752.

Bishop, J.H. (1998). The effect of curriculum-based external exit systems on student achievement. *Journal of Economic Education*, *29*, 171–182.

Braun, H. (2004). Reconsidering the impact of high-stakes testing. *Education Policy Analysis Archives*, *12*(1). Zugriff am 10.03.2004 unter http://epaa.asu.edu/epaa/v12n1/

Campbell, D. & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Carnoy, M. (2005). Have state accountability and high-stakes tests influenced student progression rates in high school? *Educational Measurement: Issues and Practice*, *24*(4), 19–31.

Carnoy, M. & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, *24*, 305–331.

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

Cheng, L. & Curtis, A. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. In L. Cheng, Y. Watanabe & A. Curtis (Hrsg.), *Washback in language testing. Research contexts and methods* (S. 3–17). Mahwah, NJ: Lawrence Erlbaum.

Cook, T.D. & Campbell, D.T. (1976). The design and conduct of quasi-experiments and true experiments in Field Settings. In M. Dunnette (Hrsg.), *Handbook of industrial and organizational research* (S. 223–326). Chicago, IL: Rand McNally.

Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record*, *106*, 1047–1085.

Davies, P. (2000). The relevance of systematic reviews to educational policy and practice. *Oxford Review of Education*, *26*, 365–378.

Dee, T. & Jacob, B. (2009). *The impact of No Child Left Behind on student achievement*. (NBER Working Paper 15531). Cambridge, MA: National Bureau of Economic Research.

Finnigan, K.S. & Gross, B. (2007). Do accountability policy sanctions influence teacher motivation? Lessons from Chicago's low-performing schools. *American Educational Research Journal*, *44*, 594–630.

Firestone, W.A., Mayrowitz, D. & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis, 20*, 95–113.

Firestone, W.A., Winter, J. & Fitz, J. (2000). Different assessments, common practice? Mathematics testing and teaching in the USA and England and Wales. *Assessment in Education*, *7*, 13–37.

Frederikson, N. (1994). *The influence of minimum competency tests on teaching and learning*. Princeton, NJ: Educational Testing Services, Policy Information Center.

Goertz, M. E. & Duffy, M. E. (2001). *Assessment and accountability systems in the 50 states: 1999–2000* (Research Rep. No. RR-046). Philadelphia, PA: Consortium for Policy Research in Education.

Hamilton, L.S. & Koretz, D.M. (2002). Tests and their use in test-based accountability systems. In L.S. Hamilton, B.M. Stecher & S.P. Klein (Hrsg.), *Making sense of test-based accountability in education* (S. 13–49). Santa Monica, CA: RAND Education.

Hanushek, E.A. & Raymond, M.E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, *24*, 297–327.

Herman, J.L. (2004). The effects of testing on instruction. In S.H. Fuhrman & R.F. Elmore (Hrsg.), *Redesigning accountability systems for education* (S. 141–166). New York: Teachers College Press.

Hursh, D. (2007). Assessing No Child Left Behind and the rise of neoliberal education policies. *American Educational Research Journal*, *44*, 493–518.

Jacob, B.A. (2001). Getting tough? The impact of high school graduation exams. *Educational Evaluation and Policy Analysis*, *23*, 99–121.

Kellaghan, T., Madaus, G.F. & Airasian, P.W. (1982). *The effects of standardized testing*. London: Kluwer-Nijhoff Publishing.

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

Lee, J. (2006). Input-guarantee versus performance-guarantee approaches to school accountability: Cross-state comparisons of policies, resources, and outcomes. *Peabody Journal of Education*, *81*, 43–64.

Lee, J. (2008). Is test-driven external accountability effective? Synthesizing the evidence from cross-state causal-comparative and correlational studies. *Review of Educational Research*, *78*, 608–644.

Lee, J. & Wong, K.K. (2004). The impact of accountability on racial and socioeconomic equity: Considering both school resources and achievement outcomes. *American Educational Research Journal*, *41*, 797–832.

Linn, R.L. (2004). Accountability models. In S.H. Fuhrman & R.F. Elmore (Hrsg.), *Redesigning accountability systems for education* (S. 73–95). New York: Teachers College Press.

McDonnell, L.M. & Choisser, C. (1997). *Testing and teaching: Local implementation of new state assessments* (CSE Tech. Rep. No. 442). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Muller, C. & Schiller, K.S. (2000). Levelling the playing field? Students' educational attainment and states' performance testing. *Sociology of Education*, *73*, 196–218.

Opfer, V.D., Henry, G.T. & Mashburn, A.J. (2008). The district effect: Systemic responses to high stakes accountability policies in six southern states. *American Journal of Education*, *114*, 299–332.

O'Day, J.A. (2002). Complexity, accountability, and school improvement. *Harvard Educational Review*, *72*, 293–329.

O'Day, J.A. (2004). Complexity, accountability, and school improvement. In S.H. Fuhrman & R.F. Elmore (Hrsg.), *Redesigning accountability systems for education* (S. 15–43). New York: Teachers College Press.

Parke, C.S., Lane, S. & Stone, C.A. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation*, *12*, 239–269.

Popham, W.J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan*, *68*, 679–682.

Porter, A.C. (1994). National standards and school improvement in the 1990s: Issues and promise. *American Journal of Education*, *102*, 421–449.

Rea-Dickins, P. & Scott, C. (2007). Washback from language tests on teaching, learning and policy: Evidence from diverse settings. *Assessment in Education*, *14*, 1–7.

Rindermann, H. (2008). International vergleichende Schulleistungs- und Intelligenzstudien: Worauf sind die Unterschiede zwischen Staaten zurückführbar? Versuch einer Erklärung unter ausschließlicher Berücksichtigung von Bildungsmerkmalen. *Empirische Pädagogik*, *22*, 17–48.

Rosenshine, B. (2003). High-stakes testing: Another analysis. *Education Policy Analysis Archives*, *11*(24). Zugriff am 11.01.2009 unter http://epaa.asu.edu/epaa/v11n24/

Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

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

Vogler, K.E. (2008). Comparing the impact of accountability examinations on Mississippi and Tennessee social studies teachers' instructional practices. *Educational Assessment*, *13*, 1–32.

Watanabe, Y. (2004). Methodology in washback studies. In L. Cheng, Y. Watanabe & A. Curtis (Hrsg.), *Washback in language testing. Research contexts and methods* (S. 19–36). Mahwah, NJ: Lawrence Erlbaum.

Wößmann, L. (2007). International evidence on school competition, autonomy and accountability: A review. *Peabody Journal of Education*, *82*, 473–497.

Wößmann, L. (2008). Zentrale Abschlussprüfungen und Schülerleistungen: Individualanalysen anhand von vier internationalen Tests. *Zeitschrift für Pädagogik*, *54*, 810–826.

Wößmann, L. & Fuchs, T. (2007). What accounts for international differences in student performance? A Re-examination using PISA data. *Empirical Economics*, *32*, 433–464.

Tabelle 1: Zentrale, curriculumbasierte Abschlussprüfungen (Sekundarstufe II) als quasi-experimentelles Treatment bzw. unabhängige Variable

Studie / Autoren	Operationalisierung von „accountability pressure“	Abhängige Variablen	Design	Ergebnisse	Einschränkungen der Validität (threats-to-validity)
Frederikson (1994)	US-Bundesstaaten mit high school graduation exams	NAEP 1978–1986	Längsschnitt	+ Höherer Leistungszuwachs in Staaten mit high stakes + Effekte bei Schülern in Jgs. 9 größer als bei älteren Schülern	- keine Kontrolle des Vorwissens - keine Kontrolle weiterer Schülervariablen - Vermutlich Kontexteffekte
Bishop (1995)	Länder mit zentralen Abschlussexamina für die Sek II in den 1980er Jahren (Australien, Finnland, Deutschland, Japan, u. a. vs. Schweden, USA)	Anforderungen Lehramt, Lehrergehälter, Schülerleistungen (IAEP 1991)	Zwei-Gruppen-Design; Internationaler Vergleich	+ Mindeststandards für Sek-II-Lehramt höher + Höhere Lehrergehälter + Bei IAEP 1991 höhere Mathematik- und Geografieleistungen	- Inferenzen mit anderen Kontextmerkmalen der einzelnen Länder sind sehr wahrscheinlich, werden jedoch nicht systematisch auf die Effekte bezogen
Bishop (1995)	Anspruchsvolle „high school regents-level courses“ mit zentralen Abschlusstests in NY State vs. nationale Indikatoren	Bildungsausgaben, Lehrergehälter, Leistungen im SAT	Zwei-Gruppen-Design; hohe Kontraste; ähnlicher kultureller Kontext	+ Hohe Bildungsausgaben im Highschool-Bereich + SAT in NY State besser als in Vergleichsgruppe	- Treatmentgruppe besteht aus einem Extremfall + Bei den SAT-Analysen werden Kontextvariablen systematisch berücksichtigt
Bishop (1995)	Kanadische Provinzen mit zentralen, curriculumbasierten Abschlussexamen in der Sek II (Highschool)	Schülerleistungen in LSA, Unterrichtsmerkmale	Zwei-Gruppen-Design; hohe Kontraste; ähnlicher kultureller Kontext	+ Bessere Leistungen in Math. und NaWi in kanadischen Provinzen mit externem Abschlussexamen + eher fachlich ausgebildete Mathematik- und Naturwissenschaftslehrkräfte + bessere Schülerlabore + mehr Hausaufgaben + häufigere Leistungskontrollen + Mehr Ansporn durch Eltern - Keine kleineren Klassen	+ Treatment- und Kontrollgruppe ähnlich umfangreich + vergleichbarer kultureller Kontext + Berücksichtigung der elterlichen Lesekompetenz - Keine Berücksichtigung weiterer, relevanter Faktoren auf individueller und schulischer Ebene

Studie / Autoren	Operationalisierung von „accountability pressure“	Abhängige Variablen	Design	Ergebnisse	Einschränkungen der Validität (threats-to-validity)
Bishop (1995)	Länder mit zentralen, curriculumbasierten Examina in der Sek II in 1980ern (Frankreich, Niederlande vs. GB, USA)	IEA 1983, IAEP 1991: Mathematik und Naturwissenschaften	Zwei-Gruppen-Vergleich	+ Bessere Math. und NaWi-Leistungen in Frankreich und Niederlande + Mögliche Kontextmerkmale der Bildungssysteme zur Erklärung des Effekts: Reputation der Schule gekoppelt an Testergebnis, Schulwahlfreiheit, Lehrergehälter, Mastery Learning	- Fallstudie - keine systematische Kontrolle weiterer Faktoren
Bishop (1997, 1999)	Länder mit externen Abschlussexamina (Sek II)	TIMSS 1995 (Jgs. 7/8)	Zwei-Gruppen-Design	+ höhere Leistungen in Mathematik, NaWi, Geographie, Lesen + Höhere Standards für Lehrerbildung + Bessere Schulausstattung (z. B. Labore) + Unterricht: mehr Quiz, mehr Experimente, weniger memorieren	
Wößmann (2007); Wößmann & Fuchs (2007); Wößmann (2008)	Länder mit: - externen Abschlussexamina (Sek II) - zentrale Tests einmal pro Schuljahr (Sek I und II)	TIMSS 1995, TIMSS-Repeat, PISA 2000 und PIRLS 2001	Regressionsanalysen auf Individualebene mit zwei Dummy-Variablen als UV	+ Pos. Effekt zentraler Abschlussprüfungen in Mathematik auf Mathematikleistungen - Kein Effekt stand. Tests auf Mathematikleistung unter Berücksichtigung zentraler Examina - Interaktionseffekt: neg. Effekt zentraler Tests in Ländern ohne zentrale Abschlussexamen + Effekt curricularer Schulautonomie auf Schülerleistungen nur in Ländern mit externen Abschlussprüfungen	+ Umfangreiche Länderstichprobe + Berücksichtigung individueller, familialer, schulischer Kontextvariablen in den Regressionsanalysen - Querschnittsdaten

Studie / Autoren	Operationalisierung von „accountability pressure“	Abhängige Variablen	Design	Ergebnisse	Einschränkungen der Validität (threats-to-validity)
Rinder-mann (2008)	Zentrale Abschlussprüfungen (Sek II)	Indizes gebildet aus IEA-Reading 1991, TIMSS 1995-2003, PISA 2000, 2003 und IGLU 2001	Makrosoziologische Analysen: Partial-korrelationen	+ Pos. Zusammenhang zwischen Schulleistung und Zentralprüfungen (r = .15 sowie Partialkorrelation von r = .36)	+ BIP herauspartialisiert + Gesellschaftliches Bildungsniveau herauspartialisiert
Muller & Schiller (2000)	Alle US-Staaten: - Wie oft und in wie vielen Fächern wurden Highschool-Schüler 1994 extern getestet? - Anzahl bedeutsamer Konsequenzen für Schüler und Schulen	- Anteil Highschool-Absolventen - Anzahl der „advanced math course credits“ in der Highschool - „Gate-keeping“-Funktion der Lehrkräfte	Regressions- und Korrelations-analysen - National Education Longitudinal Study (NELS 88-92) - National Longitudinal Study of Schools (NLSS)	+ Positive Leistungszuwächse für alle Schüler bei Testkonsequenzen für Schüler + In Staaten mit häufigen Tests sinkt die Korrelation zwischen der Einschätzung von Middleschool-Lehrkräften ob Schüler Highschool-Abschluss schafft und dem tatsächlichen Erreichen des Highschool-Abschlusses, d. h. soziale Selektivität sinkt durch häufiges Testen - Soziale Selektivität steigt bei hohen Testkonsequenzen für Einzelschulen	+ Kontrollvariablen Schüler: SES, Geschlecht, Ethnie, Leistung in der Middleschool + Mehrebenenanalyse (students within states)
Jacob (2001)	High school graduation exam (minimum competencies) in Mathematik u. Lesen im Jahr 1992	- Schülerleis-tung in Mathe-matik und Lesen in Jgs. 12 (NELS 1992) - Dropout-Raten NELS 1992	- 15 US- Staaten mit high school graduation exam vs. Staaten ohne Test - eine NELS-Kohorte	- Staaten mit MCT haben eher bildungsbenachteiligte Schüler, jedoch gleiche Lernzuwächse - Höhere Dropout-Raten in Staaten mit Test - Kein signifikanter Testeffekt auf Schülerleistungen 12 - Ausnahme Schüler im unteren Dezentil: Negativer Testeffekt auf Lesekompetenz und Dropout-Risiko	+ Echter Längsschnitt (NELS) + Kontextvariablen auf Schülerebene (grade point average, race, gender, SES, age, etc.), Schulebene (SES-Variablen des Einzugsgebiets) und Ebene des Bundesstaates (Bildungsausgaben pro Kopf, high school requirements, SES-Daten)

Tabelle 2: Indexvariablen für testbasierten Rechenschaftsdruck auf Schüler- und Schulebene

Studie / Autoren	Operationalisierung von „accountability pressure“	Abhängige Variablen	Design	Ergebnisse	Einschränkungen der Validität (threats-to-validity)
Amrein & Berliner (2002)	18 HST-Staaten mit: externen Tests in den 1990er Jahren mit 1 bis 6 Konsequenzen: Veröffentlichung der Ergebnisse, high school graduation exams, reale Konsequenzen für Schulen (Zuschüsse, Schließung), Lehrer	ACT, SAT (High school, College) NAEP cohort data (4th to 8th grade) 1996–2000	Ein-Gruppen Plan mit zwei Messzeitpunkten	- Rückgang ACT und SAT - ACT Teilnahme steigt in 9 von 18 HST-Staaten - Weniger als die Hälfte der HST-Staaten hatten Zuwächse bei NAEP-4 (1992–1996) - NAEP Quasilängsschnitt: Rückgang der Leistungen bei über der Hälfte der HST-Staaten + Zuwächse bei NAEP Lesetests; evtl. aber Interaktion mit nationalen Anstrengungen zur Verbesserung der Lesekompetenz	- Keine Kontrollgruppe (Ansatzpunkt der Kritik) - Keine Kontrollvariablen - Analyse auf deskriptivem Niveau - Willkürliche Auswahl der HST-Staaten (z. B. ohne Texas und NC)
Rosenshine (2003)	Einteilung vgl. Amrein und Berliner (2002) in eindeutige HST-Staaten; fügt eindeutige LST-Staaten hinzu	NAEP Mathe 4 (1996–2000); NAEP Mathe 8 (1996–2000); NAEP Lesen 4 (1994–1998)	Zwei-Gruppen Plan mit zwei Messzeitpunkten	+ Moderater Effekt NAEP 4 Mathematik + Substanzieller Effekt NAEP 8 Mathematik + Substanzieller Effekt NAEP 4 Lesen	+ Kontrollgruppe - Auswahl der Staaten in der Kontrollgruppe unklar - Keine Kontrollvariablen - Kohortenvergleiche - Einzelne HST-Staaten haben unterdurchschnittliche NAEP-Ergebnisse: Umsetzung der HST-Reformen genauer analysieren
Amrein-Beardsley & Berliner (2003)	siehe Rosenshine 2003	NAEP Mathe 4 (1996–2000); NAEP Mathe 8 (1996–2000); NAEP Lesen 4 (1994–1998)	Zwei-Gruppen Plan mit zwei Messzeitpunkten	- Bestätigung der Rosenshine-Befunde, jedoch ungültig, weil höhere Testausschlussquoten in HST-Staaten	+ Kontrollgruppe - Auswahl der Staaten in der Kontrollgruppe unklar - Effekt der Kontrollvariable Testausschluss nicht systematisch berechnet - Kohortenvergleiche

Studie / Autoren	Operationalisierung von „accountability pressure“	Abhängige Variablen	Design	Ergebnisse	Einschränkungen der Validität (threats-to-validity)
Braun (2004)	15 HST Staaten vs. 32 LST Staaten (nach Amrein & Berliner 2002)	NAEP 1992–2000; Durchschnitte einzelner Staaten	Zwei-Gruppen Vergleich Quer- und Quasilängsschnitt	+ Querschnitt: HST-Staaten besser - Quasi-längsschnittlicher Vergleich LST besser	+ Kontrollgruppe + Kontrolle Anteil vom Test ausgeschlossener Schüler + Berücksichtigung Standardfehler + Veränderung NAEP-Teilnehmerstaaten
Carnoy & Loeb (2002)	Index Rechenschaftsdruck für 37 US-Staaten aufgrund Daten (vor 2000) des Consortium of Policy Research in Education (CPRE; Goertz & Duffy 2001): 5, wenn Tests auf allen Schulstufen, Konsequenzen für Schulen und Schüler (high school exit test)	Zuwachs NAEP 1996–2000 - Highschool Versetzungsquoten (high school survival rates)	Regressionsanalysen	+ Positiver Effekt auf NAEP Mathematik 8 + Sowohl im oberen als auch im unteren Leistungsbereich - Keine Effekte auf NAEP Mathematik 4 - Keine Effekte auf „high school progression rates“ (Langzeiteffekt von accountability)	+ Berücksichtigen relevante Kontrollvariablen (Höherer Rechenschaftsdruck in größeren Staaten mit höherem Minderheitenanteil und niedrigen Testwerten für weiße Schüler) - Daten zu „high school progression rates“ wenig zuverlässig (konfundiert mit Bevölkerungsentwicklung)
Carnoy (2005)	Index Rechenschaftsdruck für 45 US-Staaten vgl. Carnoy & Loeb (2002)	- Highschool Abschlussquoten - Highschool Versetzungsquoten	Regressionsanalysen, differenziert nach ethnischen Gruppen	- Keine bzw. schwach negative Zusammenhänge zwischen testbasiertem Rechenschaftsdruck und Versetzungsquoten	+ Kontrollvariablen: Ethn. Zusammensetzung, Größe des Staates, Bildungsausgaben pro Schüler + Kontrolle der Versetzungs-quoten vorausgehender Jahre
Lee & Wong (2004)	US-Staaten kategorisiert in drei Gruppen basierend auf Carnoy & Loeb (2002) und weiteren „policy indices“	Quasi-längsschnittl. Lernzuwachs (NAEP 1990–2000); Bildungsausgaben nach nationalen Schulstatistiken 1990-2000	Mehrebenenanalysen separat für alle 4 AV; Lehrer/Schulen oder Distrikt (Ebene 1) in Staaten (Ebene 2)	- Kein Effekt auf Zusammenhänge zwischen Einkommen, Minderheitenstatus und Schulleistung (achievement gaps) - Kein Effekt auf Bildungsausgaben (Klassengröße, Lehrerqualifikation, Ausgaben pro Schüler)	+ Kontrollvariablen: SES, ethnische Herkunft

Studie / Autoren	Operationalisierung von „accountability pressure“	Abhängige Variablen	Design	Ergebnisse	Einschränkungen der Validität (threats-to-validity)
Lee (2006)	siehe Lee & Wong (2004)	Quasi-längsschnittl. Lernzuwachs (NAEP 1990–2000); Bildungsausgaben nach nationalen Schulstatistiken 1990-2000	Mehrebenen-analysen	- Repliziert Befund: Kein Zusammenhang zwischen Rechenschaftsdruck und Bildungsausgaben der Staaten - Zusammenhang zwischen Ressourcen-garantie und Schülerleistungen - Kein Zusammenhang zwischen Druck und Schülerleistungen bzw. sozialer Disparität - Leistungszuwächse bei Kombination von Rechenschaftsdruck und Ressourcengarantie	+ Kontrollvariablen: SES, ethnische Herkunft + Berücksichtigung des staatlichen Inputs: Ressourcengarantie (Ausgaben pro Schüler, Klassengröße, Lehrerqualifikation) + Zweifaktorieller Plan: Prüfung von Interaktionseffekten
Hanushek & Raymond (2005)	Dichotome Kategorisierung: „report card states“ vs. „consequential accountability states“ (Antizipation von NCLB)	NAEP Zuwächse 1990–2000	Regressions-analysen für inter- und intra-Staaten-vergleiche	+ Höhere NAEP-Zuwächse bei „consequential accountability“ + Kein Zusammenhang zwischen Testausschlüssen und Einführung von „consequential accountability“ + Höhere Effekte bei weißen Schülern und Hispanics - „black-white achievement gap“ wird größer + Interaktion mit Qualität und Anspruchsniveau der staatlichen Bildungsstandards	+ Effekt der Einführung wird für jeden Staat extra geschätzt (unterschiedliche Zeitpunkte) + Zuwachsraten vor Einführung werden berücksichtigt + Berücksichtigung der Testausschlussraten + Sozioökonomische Inputvariablen (SES, Bildungsausgaben) + Regressionsanalysen differenziert nach ethnischen Gruppen
Dee & Jacob (2009)	Erhöhung des Rechenschaftsdrucks durch NCLB (school accountability); teilen 39 US-Staaten in „early vs. late adopters of accountability“ auf Basis von Carnoy und Loeb (2002), Lee und Wong (2004), Hanushek und Raymond (2005)	Zeitreihen-vergleiche: NAEP 2003, 2005, 2007	Quasi-Experiment: 2 Gruppen, Pre-Posttest	+ Sign. NAEP-Zuwächse Mathematik, Jgs. 4 (v.a. Weiße, Hispanics, low-SES Schüler) + schwache Effekte auf NAEP 8 Mathematik (low-SES und leistungsschwache Schüler) - Kein Effekt auf Leseleistungen (4, 8)	+ Quasi-experimentell + Quasi-längsschnittlich + Staaten, die mind. zweimal vor NCLB am NAEP teilgenommen haben (31–38) - evtl. Unterschätzung der Effekte aufgrund schulischer Kompositionseffekte (Schülerströme hin zu Privatschulen)