Leistungseffekte des Zentralabiturs? Eine kritische Auseinandersetzung mit bildungsökonomischen Interpretationen zu den Effekten der Prüfungsorganisation auf der Basis von PISA-E-2003-Daten [Einzelbeitrag]

Rainer Block, Esther Dominique Klein, Isabell van Ackeren und Svenja Mareike Kühn

Bildungsökonomisch ausgerichtete Sekundäranalysen von PISA-Datensätzen kommen zu dem Schluss, dass mit dem Zentralabitur in der BRD ausgeprägte Leistungseffekte einhergehen. Sie sind Ausgangspunkt bildungspolitischer Empfehlungen zur Prüfungsorganisation. Die theoretischen und methodischen Grundlagen solcher Analysen erweisen sich gleichwohl als diskussionswürdig. Im Rahmen eigener Sekundäranalysen der PISA-E-2003-Daten konnten generelle und substanzielle Leistungseffekte des Zentralabiturs nicht repliziert werden.

1. Einleitung

Mit einer Ausnahme knüpfen alle deutschen Bundesländer mit der Einführung zentraler Prüfungen am Ende der Sekundarstufe II an internationale Entwicklungen an. Die Standardisierung von Prüfungsverfahren wird dabei als eine zentrale Maßnahme zur Sicherung von Qualitätsstandards betrachtet. Dies bezieht sich sowohl auf die erreichten Kompetenzen als auch auf die Qualität des Unterrichts. Erwartet wird, dass Themen aufgrund der auch für die Lehrkräfte unbekannten Themenstellung hinreichend breit abgedeckt werden, innovative Curricula und Aufgabenformate schneller in der Breite durchgesetzt werden, die Leistungsbereitschaft von Schüler/innen und Lehrkräften im Sinne extrinsischer Motivierung erhöht und diagnostische Kompetenzen sowie die Anwendung eines kriterialen Bewertungsmaßstabes eher gefördert werden könnten (Bishop & Wößmann 2004). Diesen positiven und intendierten Steuerungserwartungen stehen potenzielle, nicht erwünschte Wirkungen gegenüber: Ver-

mutet werden eine thematische Engführung des Unterrichts in der Erwartung bestimmter Aufgabenstellungen, die Förderung reproduktiven Lernens sowie die Marginalisierung verständnisorientierter und kreativer Arbeitsformen, die kaum von zentraler Stelle mit der erforderlichen Tiefe der Aufgabenstellung abfragbar seien; weiterhin fehlende Möglichkeiten, aktuelle Themen, lokale Bedingungen sowie eigene und schülerbezogene Interessen zu berücksichtigen (auch im Sinne einer deprofessionalisierenden Wirkung) sowie das Außerachtlassen unterschiedlicher Bedingungen der Leistungserbringung, etwa hinsichtlich der Frage, ob ein bestimmtes Thema bereits in allen Facetten im Unterricht besprochen wurde (van Ackeren 2007).

Bemerkenswert ist, dass sich die o.g. Erwartungen hinsichtlich der Steuerungswirkungen zentraler Abschlussprüfungen weiterhin auf vergleichsweise wenig empirische Forschung stützen können, welche die Komplexität schulischer Bildungsprozesse berücksichtigt. Eine Durchsicht der Forschungsbefunde (z.B. Bishop 1998; Baumert & Watermann 2000; Jürges, Schneider & Büchel 2003; Fuchs & Wößmann 2007; Hanushek & Wößmann 2007; Wößmann 2005, 2007, 2008), die insbesondere auf Reanalysen von TIMSS und PISA-Datensätzen beruhen, deuten tendenziell Vorteile zentraler Prüfungen für das Erreichen hoher Standards an. Über die Studien hinweg erweist sich das Bild jedoch als inkonsistent und variiert fach- und altersgruppenspezifisch (vgl. Maag Merki 2008). Exemplarisch herausgehoben seien an dieser Stelle die Analysen im Kontext von TIMSS zum Zusammenhang von Prüfungsform sowie individuellen und institutionellen Lernprozessen (Baumert & Watermann 2000). Dabei zeigt sich, dass die Ergebnisse nach Fach (Mathematik und Physik) und Kursniveau differenziell zu betrachten sind: Die Leistungsstreuung ist in Ländern mit Zentralabitur (ZA) im Grundkurs Mathematik reduziert; für Physik lassen sich auf dieser Datenbasis keine entsprechenden Effekte nachweisen. „Ebenso wenig lassen sich Hinweise finden, dass die Leistungsunterschiede zwischen Schulen bei einem Zentralabitur kleiner werden" (Baumert & Watermann 2000, S. 346). Dennoch wirkt das ZA offenbar im untersten Leistungsbereich in beiden Domänen standardsichernd und führt auf den höchsten Kompetenzstufen zu leicht erhöhten Schüleranteilen gegenüber der Verteilung in Ländern mit dezentraler Prüfungsorganisation. Insgesamt zeigen die Analysen aber kein konsistentes Bild.

In diesem thematischen Kontext haben in jüngerer Zeit bildungsökonomisch geprägte Analysen zur Leistungswirksamkeit so genannter zentraler Abschlussprüfungen in der BRD besondere Aufmerksamkeit erfahren (z.B. Wößmann 2007, 2008), die – insbesondere auf der Grundlage der Reanalyse von PISA-

2003-Daten – mit bildungspolitischen Schlussfolgerungen verknüpft sind: „Über vier internationale Schülerleistungsvergleiche [...] hinweg, zeigen umfassende Regressionsanalysen der Schülerindividualdaten, dass zentrale Abschlussprüfungen im internationalen Vergleich mit wesentlich besseren Schülerleistungen einhergehen [...]. Der gleiche Zentralprüfungseffekt findet sich in statistisch nicht zu unterscheidender Größenordnung auch im Vergleich der deutschen Bundesländer (Wößmann 2008, S. 824). Somit ist „die Tendenz in der deutschen Schulpolitik, dass immer weitere Bundesländer zentrale Abschlussprüfungen einführen, begrüßenswert [ist]. Die hier berichtete internationale Evidenz gibt klare Hinweise für die politische Diskussion über bundeseinheitliche Abschlussprüfungen in Sekundarabschlüssen wie dem Zentralabitur" (ebd. S. 824f). Die Analysen werden nachfolgend zum Ausgangspunkt für den Bericht und die Diskussion eigener Auswertungen zu den Leistungseffekten des ZA in der BRD gemacht.

2. Bildungsökonomische Interpretationen bildungssystemischer Effekte

Die knapp skizzierten bildungsökonomischen Analysen in Bezug auf die Effekte des ZA für das mittlere Leistungsniveau in der BRD basieren auf OLS-Regressionen auf der Aggregatebene der 16 Bundesländer anhand von PISA-E-2003-Daten für die Alterskohorte der 15-Jährigen über alle Schulformen hinweg, wobei die Bruttoleistungswerte allein für den Kompetenzbereich der Mathematik modelliert werden. Zur Erklärung der Bruttoleistungswerte wird auf ein Set von Prädiktoren (auf der Aggregatebene der einzelnen Bundesländer) zurück gegriffen, das aus einer Mischung von Kontextmerkmalen und aggregierten Individualmerkmalen von Schülerinnen und Schülern (wie z.B. socio-economic background, parental education und father employment) besteht. Die Form der Abschlussprüfungen am Ende der Sekundarstufe II wird als binäres Merkmal (ZA: ja oder nein) kodiert. Die angenommenen Leistungsvorteile der Bundesländer mit ZA gegenüber solchen ohne ZA werden schließlich u.a. anhand von deskriptiven – inferenzstatistisch ausdrücklich nicht abgesicherten – Mittelwertunterschieden der einzelnen Bundesländer abgeleitet. Ein solcher Analyseansatz kann gleichwohl unter mehreren Gesichtspunkten diskutiert werden.

Generelle oder singuläre domänenspezifische Leistungseffekte. Bei den Analysen bleibt ungeklärt, ob es sich beim angenommenen Leistungsvorsprung der Bundesländer mit ZA im Kompetenzbereich Mathematik um einen singulären Ef-

fekt handelt, der für andere Leistungsdomänen (wie Lesekompetenz und naturwissenschaftliche Kompetenz) möglicherweise so nicht zu beobachten ist. Würde es sich lediglich um einen domänenspezifischen Effekt handeln, dann wäre eine Verallgemeinerung der Leistungswirksamkeit so genannter Zentralprüfungen nicht zulässig. Zudem wäre ein singulärer Effekt ein deutlicher Hinweis darauf, dass die Leistungsunterschiede möglicherweise eher das Ergebnis unbeobachteter Heterogenität als die kausale Wirkung einer zentralen Prüfungsorganisation sind.

‚Fernwirkung' des Zentralabiturs auf die Alterskohorte 15-Jähriger. Die analysierten PISA-2003-Daten beziehen sich auf das Kompetenzniveau von 15-jährigen Schüler/innen. Diese Kompetenzwerte wiederum werden regressionsanalytisch auf eine zentrale oder dezentrale Prüfungsorganisation im Abitur, d.h. am Ende der Sekundarstufe II, zurückgeführt. Eine solche ‚Fernwirkung' der Prüfungsorganisation im Abitur auf den vorgelagerten Unterricht von 15-jährigen Alterskohorten erscheint u.E. wenig plausibel.

Leistungseffektivität für Schüler/innen aller Schulformen. Die Regressionsanalysen modellieren den Zusammenhang zwischen der Prüfungsorganisation im Abitur und den mittleren Mathematik-Leistungswerten der einzelnen Bundesländer für alle Schulformen zusammen. Es erscheint allerdings kaum nachvollziehbar, dass die Art der Prüfungsorganisation im Abitur auch Einfluss nehmen könnte auf die Lernleistungen von 15-jährigen Hauptschülern und das Lehrverhalten von Hauptschullehrkräften. Dem Gegenstandsbereich angemessener wäre wohl eine Modellierung der Kompetenzwerte bezogen auf die Gymnasien, zumal die Rangplätze der Bundesländer in Bezug auf die Mathematikkompetenz der Schüler/innen aller Schulformen und hinsichtlich der Mathematikkompetenz der 15-jährigen Gymnasiast/innen mit rho = .87 bei weitem nicht deckungsgleich sind (vgl. PISA-Konsortium Deutschland 2005, S. 60, 77). Bei einer solchen inhaltlich notwendigen Passung der Indikatoren für das ZA und der zu erklärenden Zielgröße der Gymnasialkompetenzen liegt die Schwierigkeit allerdings darin, dass ein Großteil der verwendeten aggregierten Individualmerkmale (wie z.B. parental education und father employment) als Prädiktoren für die Gruppe der 15-jährigen Gymnasiasten im herangezogenen Datensatz nicht verfügbar war.

Modellierung mittlerer Leistungswerte bzw. unterschiedlicher Leistungsquantile. Im Rahmen der bildungsökonomischen Konzeptualisierung zentraler Prüfungsorganisation wird eine generelle Anreizfunktion des ZA im Sinne steigender extrinsischer Motivation unterstellt. Damit wird ein gleichmäßiger Leistungseffekt über alle Bereiche der Leistungsverteilung hinweg vorausgesetzt. Folge-

richtig beschränken sich die OLS-Regressionen auf die Modellierung durchschnittlicher Leistungswerte (des konditionalen Mittelwertes).

In der bildungspolitischen Diskussion werden dem s.g. ZA im Wesentlichen die Funktionen der Vergleichbarkeit und Standardsicherung zugewiesen (vgl. z.B. Schulministerium NRW 2009). Standardsicherung meint dabei zweierlei: Zum einen die Sicherung von Mindestnormen, die sich auf die Qualität des Prüfungsverfahrens beziehen, und zweitens die Sicherung der Output-Qualität im Sinne der erzielten Leistungsergebnisse. Einem solchen Verständnis zufolge sollte das ZA speziell in den unteren Leistungsbereichen Effekte im Sinne einer Mindeststandardsicherung bewirken. Dies würde forschungspraktisch eher eine Modellierung unterer Leistungsquantile mittels Quantilregressionen nahe legen. Die berichteten spärlichen und nicht durchweg konsistenten empirischen Befunde wiederum deuten zudem auf mögliche Leistungseffekte des ZA sowohl im unteren als auch oberen Leistungsbereich hin (vgl. Baumert & Watermann 2000). Dies wiederum würde u.E. eine regressionsanalytische Modellierung oberer und unterer Leistungsquantile – nicht aber zwingend des konditionalen Mittelwertes – implizieren.

Die Aggregatebene der Bundesländer. Die Analyseebene der Bundesländer erweist sich als problematisch. Insbesondere für Schülerleistungsuntersuchungen gilt: Es gibt kaum überwindbare Interpretationsprobleme korrelativer Zusammenhänge auf der Aggregatebene der Länder, „ ...da fast alle strukturellen und institutionellen, aber auch kulturellen Kontextvariablen untereinander und mit Drittvariablen konfundiert sind. Ergebnisse auf dieser Ebene haben in der Regel den Wert von bestenfalls anregenden Vermutungen" (Baumert, Carstensen & Siegle 2005, S.326). Auch das weiter oben referierte bildungsökonomische Produktionsmodell impliziert letztlich Handlungsanreize für Individuen. Will man aber Effekte auf der individuellen Ebene – wie ein hohes Leistungsniveau – erklären, dann ist eine Analysestrategie, die unterschiedliche erklärende Individualmerkmale – wohlgemerkt: keine Kontextmerkmale – auf Länderebene aggregiert (wie z.B. socio-economic background, parental education und father employment), wegen der Gefahr des ökologischen Fehlschlusses (Robinson 1950) u.E. nicht geeignet. Die Frage, inwieweit die länderspezifisch unterschiedliche Verteilung leistungsrelevanter Individualmerkmale für Leistungsunterschiede zwischen Ländern verantwortlich ist, lässt sich nur auf der Basis einer Individualdatenanalyse adäquat klären (vgl. auch Baumert, Carstensen & Siegle 2005).

Der Mehrebenen-Charakter der Leistungsdaten. Eine Auswertungsstrategie, die sich auf die Aggregatebene der Bundesländer beschränkt, wird dem Mehrebe-

nencharakter der Leistungsdaten nicht gerecht. Bei Analysen auf der Ebene der Bundesländer gehen die Varianzen innerhalb der einzelnen Länder verloren. Ein solches Vorgehen erscheint nur dann plausibel und vertretbar, wenn die Zwischenvarianz die zentrale Quelle der Varianzaufklärung der Leistungswerte darstellt und die Binnenvarianzen vernachlässigbar sind. Dies müsste im Vorfeld entsprechender Analysen zunächst geklärt werden. Dass die Varianzaufklärung der Kompetenzwerte wesentlich auf Effekte der Länderebene – und eben nicht auf Effekte der Schul- oder Klassenebene – begründet ist, ist jedoch angesichts der Befundlage der internationalen Schuleffektivitätsforschung wenig wahrscheinlich. Der Forschungsstand dokumentiert eher das Gegenteil: Die Klassenebene stellt mit 38 bis 55 Prozent erklärter Leistungsvarianz die wichtigste Varianzkomponente dar, gefolgt von der Schulebene mit 7 bis 20 Prozent Varianzaufklärung (vgl. z.B. den Überblick bei Hill 2001; Scheerens & Bosker 1997).

Leistungsunterschiede zwischen einzelnen Ländern. In den o.g. bildungsökonomischen Analysen wird anhand der deskriptiven Gegenüberstellung der mittleren Bruttoleistungswerte der Bundesländer in der Mathematikkompetenz auf substanzielle Leistungseffekte eines zentral organisierten Abiturs geschlossen. Wie oben bereits aufgeführt, erscheint es wenig stringent, die mittleren Testleistungen über alle Schulformen unmittelbar mit der Indikatorvariablen für das ZA zu vergleichen. Setzt man hingegen die mittlere gymnasiale Testleistung der Länder in Relation zur Existenz zentraler oder dezentraler Abschlussprüfungen, dann zeigt sich ein deutlich abgeschwächter deskriptiver Zusammenhang: Ein Bundesland ohne ZA weist einen höheren Bruttoleistungswert als drei Länder mit ZA auf, ein weiteres Bundesland ohne ZA ist besser als zwei Länder mit ZA und weitere zwei Länder ohne ZA sind nominal besser als ein Bundesland mit ZA (Korrelation r= .66 zwischen gymnasialen Leistungswerten und der Indikatorvariablen ZA). Eine allein deskriptive Gegenüberstellung verkennt den Stichprobencharakter der PISA-Daten. Die Frage, ob sich die mittleren Schülerleistungen in der Grundgesamtheit der einzelnen Bundesländer unterscheiden, lässt sich adäquat nur anhand von Mittelwertvergleichen der Bundesländer auf Individualdatenbasis in Form paarweiser multipler Signifikanztests (mit alpha-Adjustierung) beantworten. Mit der Feststellung möglicher signifikanter Unterschiede wäre aber gleichwohl noch keine Aussage über die praktische, substanzielle Bedeutsamkeit solcher Unterschiede getroffen.

Binäre Kodierung des Zentralabiturs. Eine einfache binäre Kodierung der Abschlussprüfungen am Ende der Sekundarstufe II (mit Zentralabitur vs. ohne Zentralabitur) wird der Vielgestaltigkeit, der Komplexität und dem differenziellen Standardisierungsgrad der realen Prüfungsorganisation von Abschlussprü-

fungen in den einzelnen Bundesländern inhaltlich nicht gerecht. Nationale und internationale Übersichtsstudien zur Prüfungsorganisation in der Sekundarstufe II (zusammengefasst in Klein u.a. 2009) haben genau diese Problematik aufgegriffen. Ein innerdeutscher Vergleich der Prüfungsverfahren im ZA zeigt, dass sich nur vergleichsweise wenige Prüfungselemente als weitgehend konsensfähig unter den Ländern der BRD erweisen und die Unterschiede der Bundesländer in Bezug auf die einzelnen Prüfungselemente überwiegen. Diese Differenzen zeigen sich bei – unter Standardisierungsgesichtspunkten – so zentralen Sachverhalten wie z.B. der Frage, ob in allen oder nur in ausgewählten Fächern die schriftliche Abiturprüfung abgelegt wird, ob alle schriftlichen Prüfungen zentral erfolgen oder in Abhängigkeit von bestimmten Fächern bzw. Fächergruppen und Anforderungsniveaus, ob als Grundlage der Prüfungen die so genannten EPA und Lehrpläne mit oder ohne Berücksichtigung von Schwerpunktthemen dienen, zu welchem Zeitpunkt vor der Prüfung entsprechende Schwerpunktthemen bekannt gegeben werden, ob Auswahlmöglichkeiten bei den Prüfungsthemen für Schüler/innen und/oder die Lehrerschaft bestehen, aus welcher Personengruppe die Zweitkorrektoren stammen (schulintern oder -extern), ab welcher Bewertungsdifferenz zwischen Erst- und Zweitkorrektor ein Drittkorrektor hinzugezogen wird oder welchen Anteil die zentralen Prüfungselemente an der Gesamtbewertung ausmachen.

Unbeobachtete Heterogenität. Der Versuch, die o.g. Ergebnisse der bildungsökonomischen Untersuchungen zur Effektivität des ZA in der BRD durch Modellierungen auf der Ebene von OECD-Staaten abzusichern, erweitert die oben aufgeführten methodischen und interpretativen Probleme anstatt sie zu lösen. Schümer und Weiß (2008, S.45) schlussfolgern:

> „Schon bei der Betrachtung der deutschen Bundesländer wird deutlich, dass der Leistungsvorsprung der Schüler aus Ländern mit zentralen Prüfungen zu einem ganz erheblichen Teil auf unbeobachtete Unterschiede zwischen diesen Ländern und den Bundesländern ohne zentrale Prüfungen zurückzuführen ist. Beim Vergleich der Leistungen von Schülern aus unterschiedlichen Staaten verstärkt sich der Verdacht, dass die den zentralen Prüfungen zugeschriebenen Effekte Wirkungen nicht erfasster Einflussgrößen sind. Im Hinblick auf die großen Unterschiede im Charakter der Prüfungen, in ihrer Relevanz für die untersuchte Altersgruppe und in den Bedingungen, unter denen sie durchgeführt werden, ist gar nicht zu erwarten, dass

> diese zentralen Prüfungen vergleichbare Wirkungen auf die Schüler ausüben und ihre Leistungen steigern. [...] Im Bemühen um generalisierbare Aussagen werden trotzdem über alle Länder hinweg Durchschnittseffekte für die institutionellen Faktoren ermittelt und damit – verschiedener Kontextunterschiede ungeachtet – nationale Politikempfehlungen begründet."

3. Reanalyse von PISA-E-2003-Daten

Im Rahmen eigener Sekundäranalysen der PISA-E-2003-Daten lassen sich generelle und substanzielle Leistungseffekte des ZA nicht aufzeigen. Die Reanalysen wurden im Rahmen eines laufenden DFG-Projektes durchgeführt.[69] Dieses greift die referierten Forschungsdesiderata auf und untersucht die Wirkungen unterschiedlicher Prüfungsmodalitäten (zentral vs. dezentral) auf schulische Arbeitsprozesse in drei Ländern mit jeweils unterschiedlichen Prüfungstraditionen im Abitur (BW, RP, NW). Die Aggregatebene der Länder erweist sich als eine unter methodischen und theoretischen Gesichtspunkten letztlich suboptimale Analyseebene, wenn es darum geht die Wirkungen schulsystemischer Steuerungselemente (wie dem ZA) auf Schülerleistungen abzuschätzen. Potenzielle Leistungseffekte des ZA lassen sich – wie oben ausgeführt – nur mit einem Analysemodell auf Individualdatenbasis untersuchen, das dem Mehrebenencharakter der Leistungsdaten hinreichend Rechnung trägt. Ein solcher Ansatz soll im Folgenden vorgestellt werden.

Methode

Bei der Modellierung der PISA-2003-Querschnittdaten wird auf ein in der internationalen Schuleffektivitätsforschung etabliertes Analysemodell zurück gegriffen, das unter dem Namen ‚unpredicted achievement model' (Hill 2001, Scheerens & Bosker 1997) oder auch ‚(cross sectional) contextualised attainment model' (OECD 2008) firmiert. Bei diesem Ansatz werden die Bruttoleistungswerte um die Effekte ausgesuchter Variablen der Eingangsselektivität (wie soziodemografische Hintergrundmerkmale der Schüler/innen) bereinigt. Gruppenvergleiche (auch mehrebenenanalytischer Art) zur Klassen-, Schul- oder (Bundes-)Ländereffektivität können dann auf der Basis der adjustierten, bereinigten Testleistungswerte durchgeführt werden.

[69] Durchführung im Rahmen der DFG-Forschergruppe und des Graduiertenkollegs Naturwissenschaftlicher Unterricht an der Universität Duisburg-Essen („nwu-essen")

Die Analysen beschränken sich auf die Testleistungen der 15-jährigen Gymnasiast/innen des PISA-E-2003-Datensatzes. Als Fallgewicht wird das Populationsgewicht PISA-EOM, E inkl. Oversampling nach Migrationshintergrund benutzt. Für jedes der fünf Plausible Values der drei Leistungsdomänen (Mathematik, Lesen, Naturwissenschaften) wird eine OLS-Regression auf Individualdatenbasis durchgeführt. Die Bruttotestleistungswerte werden auf die Merkmale Geschlecht, Migrationshintergrund (binär kodiert) und index of economic, social and cultural status – den ESCS-Index – als für das gymnasiale Leistungsniveau bedeutsame Indikatoren der Eingangsselektivität regressionsanalytisch zurückgeführt. Leider müssen die besonders erklärungskräftigen kognitiven Grundfähigkeiten der Schüler als Kovariate unberücksichtigt bleiben, da sie im für interessierte Wissenschaftler/innen zugänglichen PISA-Datensatz – durchaus gut begründet – nicht verfügbar sind.

Die Residuen der fünf Regressionsmodelle je Leistungsdomäne werden anschließend gemittelt und über die Gymnasialschüler/innen aller Länder auf einen Wertebereich mit einem Mittelwert von 500 und einer Standardabweichung von 100 transformiert. Bei diesen Residuen handelt sich um adjustierte, um Effekte der Eingangsselektivität bereinigte Testleistungswerte. Um einer Verwechselung mit den Bruttoleistungswerten gleicher Skalierung vorzubeugen, nennen wir die adjustierten Testleistungswerte im Folgenden Residuen Mathematikkompetenz, Residuen Lesekompetenz und Residuen naturwissenschaftliche Kompetenz.

Die dergestalt generierten Residuen auf Individualebene können auf unterschiedlichen Aggregatebenen – wie der der Schulen oder Bundesländer – als Effektivitätswerte zusammen gefasst werden und ermöglichen so eine flexible Modellierung bereinigter, adjustierter Testleistungswerte inklusive hierarchischer, mehrebenenanalytischer Random Effect Modelle (vgl. zu diesem Vorgehen auch Watermann & Stanat 2004, und in Bezug auf Längsschnittdaten OECD 2008; Doran & Izumi 2004).

Ergänzend werden ausgesuchte mathematikbezogene Unterrichts- bzw. Schülervariablen untersucht, um die eingangs genannten Hypothesen zu den unterrichtlichen Effekten der Prüfungsorganisation zu prüfen. Dabei handelt es sich um die Skalen (vgl. PISA-Konsortium Deutschland 2006)

- ‚instrumental motivation' in Mathematik (Beispielitem „Ich gebe mir in Mathematik Mühe, weil es mir in meinem späteren Job weiterhelfen wird"),

- ‚mathematic self-efficacy' (Beispielitem „Wie sicher glaubst du folgende Mathematikaufgaben lösen zu können?"),
- ‚memorisation' (Beispielitem „Wenn ich für Mathematik lerne, lerne ich so viel wie möglich auswendig"),
- ‚competitive learning' (Beispielitem „In Mathematik wäre ich gern der/die Beste"),
- ‚cooperative learning' (Beispielitem „In Mathematik arbeite ich gerne in Gruppen mit Mitschülerinnen und Mitschülern zusammen"),
- ‚interest and enjoyment' (Beispielitem „Ich freue mich auf meine Mathematikstunden"),
- teacher support (Beispielitem „Unser Lehrer/unsere Lehrerin interessiert sich für den Lernfortschritt jedes einzelnen Schülers/jeder Schülerin") und
- ‚student-teacher relation' (Beispielitem „Die meisten meiner Lehrer/Lehrerinnen interessieren sich für das, was ich zu sagen habe").

Mit dem hier vorgestellten Modell auf Individualdatenebene wird ein Großteil der weiter oben benannten Schwierigkeiten der o.g. Aggregatdatenanalyse in Bezug auf die PISA-E-2003-Daten berücksichtigt. Einige problematische Konzeptualisierungen – wie die der Rückbindung einer lediglich binär kodierten Prüfungsorganisation im ZA auf die mittleren Testleistungen 15-Jähriger – werden hingegen fortgeschrieben. Dies geschieht mit der Intention, anhand der PISA-E-2003-Individualdaten und auf der Basis eines etablierten (mehrebenenanalytischen) Value-Added-Effektivitätsmodells die behaupteten Leistungseffekte (ebenso wie mögliche Unterrichtseffekte) zentraler Abiturprüfungen in der BRD sekundäranalytisch prüfen zu können.

Befunde

Bei den adjustierten Leistungswerten (im Sinne regressionsanalytisch ermittelter Residuen) zeigen sich – unter Anwendung robuster Varianzschätzer, die der mehrstufigen Klumpenstichprobe von PISA Rechnung tragen (Taylor Series Expansion, vgl. American Institutes for Research & Cohen 2009) – statistisch signifikante, aber insgesamt wenig prägnante Unterschiede der mittleren Testleistungen in den drei Leistungsdomänen zwischen Ländern mit und ohne ZA, mit Leistungsvorteilen für die zuerst genannte Ländergruppe (vgl. Tabelle 1). Die Mittelwerte unterscheiden sich in einer Größenordnung von maximal rund einer drittel Standardabweichung (Residuen Mathematik d=.35, Residuen Lesen d=.20, Residuen Naturwissenschaften d=.28).

In der empirischen Bildungsforschung ist es – in Anlehnung an Cohen (1988, 1992) – gängige Praxis, Effektstärken in der Größenordnung von d>=.2 als „kleine", d>=.5 als „mittlere" und d>=.8 als „große" Effekte zu bezeichnen (vgl. PISA-Konsortium Deutschland 2008, S.59). Darüber hinaus findet sich in den Veröffentlichungen des Deutschen PISA-Konsortiums das Vorgehen, die praktische Relevanz von Leistungsunterschieden zudem durch eine Umrechnung in durchschnittliche Lernzeiten zu veranschaulichen. Demnach entsprechen Leistungsunterschiede von 25, 30 oder auch 40 Punkten – u.a. in Abhängigkeit von der Leistungsdomäne und dem Zeitpunkt der PISA-Erhebung – einem praktisch relevanten Kompetenzzuwachs von rund einem Schuljahr (vgl. PISA-Konsortium Deutschland 2008, S.59, PISA-Konsortium Deutschland 2005, S.38). Für den Bereich der mathematischen Kompetenz entspricht dieser Lernzuwachs eines Schuljahres wiederum einer Effektgröße von d=0.33 (ebd.).

Die Gruppendifferenzen der Residual-Leistungspunkte lassen sich in Rahmen unseres Modellansatzes nicht mehr sinnvoll in durchschnittliche Lernzeiten rückübersetzen. Zugleich erachten wir es für plausibel, im Rahmen unserer Modellierungen erst ab einer mittleren Effektgröße (im Sinne Cohens) von praktisch relevanten, substanziellen Mittelwertunterschieden auszugehen, und zwar aus folgenden Gründen: Zum einen haben Monte Carlo-Simulationen gezeigt (vgl. Barnette 2006), dass kleine Effektgrößen – in Abhängigkeit vom Stichproben- und Gruppenumfang der Zufallsstichproben – auch dann gehäuft auftreten, wenn zwischen den Gruppen realiter keinerlei Mittelwertunterschied besteht. Zum anderen werden von bildungspolitischer und mehr noch bildungsökonomischer Seite starke Steuerungsimpulse vom Zentralabitur in Bezug auf das Leistungsniveau erwartet bzw. postuliert (s.o.). Vor diesem Hintergrund halten wir kleinere Effektgrößenunterschiede für ein unzureichendes Relevanzkriterium. Schließlich ist daran zu erinnern, dass bei der Berechnung der adjustierten Schülerleistungen die Grundintelligenz der Schüler – als ein zentraler Prädiktor der Eingangsselektivität – mangels Verfügbarkeit nicht berücksichtigt werden konnte. Auch deshalb sollten kleine Mittelwertunterschiede nicht überinterpretiert werden.

Die Bewertung der Ergebnisse ändert sich nicht, unabhängig davon, ob man ab einer drittel oder erst ab einer halben Standardabweichung Differenz von praktisch relevanten, substanziellen Mittelwertunterschieden sprechen möchte: Die mittleren Testleistungen in den drei Leistungsdomänen zwischen Ländern mit und ohne Zentralabitur unterscheiden sich in einer statistisch signifikanten aber insgesamt eher kleinen Größenordnung. Ein deskriptiver Befund, der sich

so auch schon bei Baumert und Watermann (2000, S. 351) in Bezug auf die TIMSS/III-Erhebung findet.

	Bundesländer ohne ZA			Bundesländer mit ZA			t-Test	Mittelwert-Differenz Effektgröße
Merkmal	**MW**	**SD**	**SE**	**MW**	**SD**	**SE**	**p>t**	**d**
Residuen Mathematik	485	98.3	2.2	519	98.7	2.9	0.000	-0.35
Residuen Lesen	491	98.9	2.6	511	100.3	3.1	0.000	-0.20
Residuen Naturwissenschaft	488	99.6	2.3	516	98.3	2.6	0.000	-0.28
Skala Instrumental Motivation	-0.06	1.01	0.02	-0.16	1.01	0.03	0.003	0.10
Skala Self-Efficiacy	0.37	0.94	0.02	0.40	0.92	0.02	0.296	-0.03
Skala Memorisation	-0.11	0.96	0.01	-0.33	0.94	0.02	0.000	0.23
Skala Competitive Learning	-0.20	0.95	0.02	-0.26	0.94	0.03	0.035	0.07
Skala Cooperative Learning	-0.01	0.94	0.02	-0.10	0.91	0.02	0.001	0.10
Skala Interest and Enjoyment	-0.01	0.98	0.02	-0.13	0.97	0.02	0.001	0.12
Skala Teacher Support	-0.17	0.97	0.02	-0.24	0.95	0.03	0.071	0.08
Skala Student-Teacher Relation	-0.12	0.89	0.03	-0.03	0.88	0.03	0.036	-0.09

Tabelle 1: Mittelwertdifferenzen zwischen Bundesländern ohne bzw. mit Zentralabitur nach ausgesuchten Merkmalen

Die Frage wiederum, inwieweit sich einzelne Länder in den bereinigten durchschnittlichen Testleistungen überzufällig unterscheiden, lässt sich adäquat anhand multipler Mittelwertvergleiche[70] beantworten. Für die adjustierten Testleistungen der Gymnasien finden sich im multiplen, paarweisen Ländervergleich nur noch wenige signifikante Unterschiede (s. Abbildung 1 bis 3).

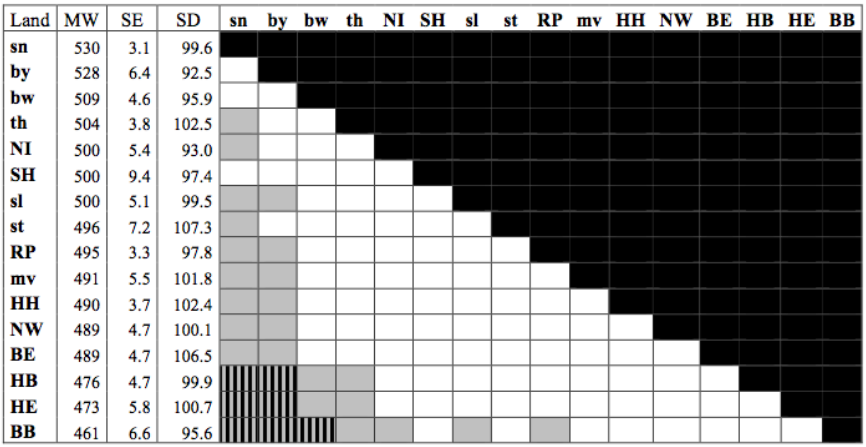
Aus den Abbildungen ist ersichtlich, ob ein Bundesland in der Tabellenspalte statistisch signifikant (grau unterlegt) besser ist als ein Land in der Tabellenzeile. Ergänzend ist aufgeführt, ob sich ein Bundesland mit mindestens mittlere Effektgröße von einem anderen Land unterscheidet (gemusterte Zelle). Bundesländer mit ZA sind in kleinen Lettern angegeben.

[70]Bei den Analysen wurden Bonferroni-Adjustierungen bei einem multiplen Testniveau von alpha=0.05 durchgeführt.

Ein Muster, wonach die Länder mit ZA systematisch signifikant besser wären als Länder ohne ein solches, ist auf der Grundlage dieser Analysen nicht zu beobachten. Zudem zeigen sich tendenziell domänenspezifische Ausprägungen, wobei in diesem Zusammenhang daran zu erinnern ist, dass die mit den Kompetenzbereichen korrespondierenden Fächer in sehr unterschiedlicher Form in die zentralen Abiturprüfungsverfahren eingebunden sind. Als besonders gering erweisen sich die Differenzen im Bereich Lesen: So unterscheiden sich allein fünf Länder ohne ZA (zum Testzeitpunkt, von insgesamt neun Län*dern) statistisch nicht signifikant von allen sieben Ländern mit ZA. Im Hinblick* auf die praktische Signifikanz fallen die Unterschiede noch geringer aus. Nur die Länder Bayern und Baden-Württemberg – als Vertreter der Gruppe mit ZA – differieren in ihrer mittleren Leseleistung mit mindestens mittlere Effektgröße von Brandenburg.

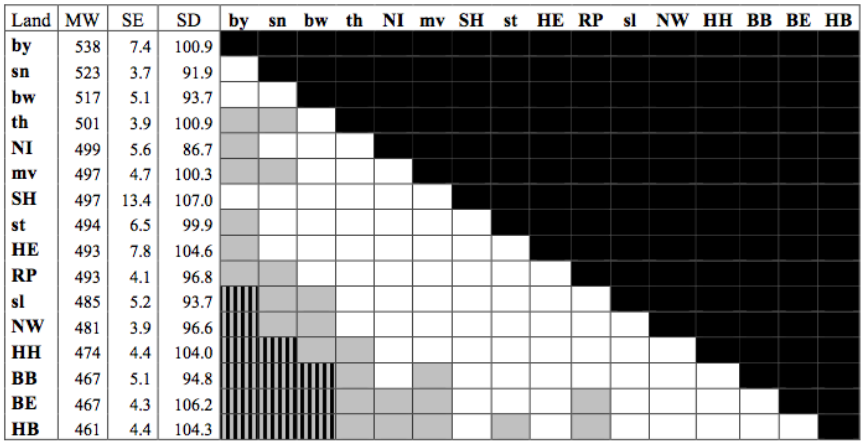
Land	MW	SE	SD	by	bw	SH	sn	RP	th	HH	NI	sl	NW	st	HE	HB	mv	BE	BB
by	523	8.0	100.4																
bw	518	5.1	91.6																
SH	508	8.1	98.4																
sn	507	4.1	95.3																
RP	501	3.7	96.9																
th	501	4.6	113.4																
HH	501	3.6	111.7																
NI	497	4.4	88.8																
sl	497	3.7	92.8																
NW	494	5.8	97.8																
st	487	7.6	110.3																
HE	484	6.6	99.9																
HB	482	5.2	101.7																
mv	479	5.7	105.0																
BE	479	4.0	98.6																
BB	465	5.5	111.5																

Abbildung 1: Multiple Mittelwertvergleiche der Bundesländer – Residuen Lesekompetenz



Land	MW	SE	SD	sn	by	bw	th	NI	SH	sl	st	RP	mv	HH	NW	BE	HB	HE	BB
sn	530	3.1	99.6																
by	528	6.4	92.5																
bw	509	4.6	95.9																
th	504	3.8	102.5																
NI	500	5.4	93.0																
SH	500	9.4	97.4																
sl	500	5.1	99.5																
st	496	7.2	107.3																
RP	495	3.3	97.8																
mv	491	5.5	101.8																
HH	490	3.7	102.4																
NW	489	4.7	100.1																
BE	489	4.7	106.5																
HB	476	4.7	99.9																
HE	473	5.8	100.7																
BB	461	6.6	95.6																

Abbildung 2: Multiple Mittelwertvergleiche der Bundesländer – Residuen naturwissenschaftliche Kompetenz



Land	MW	SE	SD	by	sn	bw	th	NI	mv	SH	st	HE	RP	sl	NW	HH	BB	BE	HB
by	538	7.4	100.9																
sn	523	3.7	91.9																
bw	517	5.1	93.7																
th	501	3.9	100.9																
NI	499	5.6	86.7																
mv	497	4.7	100.3																
SH	497	13.4	107.0																
st	494	6.5	99.9																
HE	493	7.8	104.6																
RP	493	4.1	96.8																
sl	485	5.2	93.7																
NW	481	3.9	96.6																
HH	474	4.4	104.0																
BB	467	5.1	94.8																
BE	467	4.3	106.2																
HB	461	4.4	104.3																

Abbildung 3: Multiple Mittelwertvergleiche der Bundesländer – Residuen Mathematikkompetenz

In den Naturwissenschaften unterscheiden sich Sachsen und Bayern mit (mindestens) mittlerer Effektgröße von Brandenburg, Hessen und Hamburg. Daneben erweist sich noch Baden-Württemberg deutlich besser als Brandenburg. Ansonsten finden sich beim paarweisen Ländervergleich keine weiteren Leis-

tungsunterschiede dieser Größenordnung bei den adjustierten naturwissenschaftlichen Testleistungen. Bei den Residuen der Mathematikkompetenz sind es immerhin drei Länder mit ZA (BY, SN, BW), die mit mindestens mittlerer Effektgröße leistungsstärker als drei bis hin zu fünf Länder ohne zentrale Abschlussprüfung sind. In den Residuen der Mathematikkompetenz zeigen sich somit die deutlichsten Leistungsunterschiede der Länder mit unterschiedlichen Prüfungsorganisationen im Abitur.

Gerade diese tendenziell domänenspezifisch differierenden Effekte lassen sich aber möglicherweise als ein Hinweis darauf hin interpretieren, dass weniger die Prüfungsorganisation als vielmehr unbeobachtete Drittvariablen für die Leistungsunterschiede verantwortlich zeichnen. Die bisherigen Erklärungsversuche fach- und kursniveauspezifischer Effekte zentraler Abschlussprüfungen können nicht wirklich überzeugen. So identifizieren Baumert und Watermann (2000) kleine Effekte in der Leistungsdomäne Mathematik (nicht aber in Physik), die sich nur in Grundkursen, nicht aber in Leistungskursen finden. Die Autoren vermuten, dass sich die potentiell standardsichernde Wirkung des Zentralabiturs im unteren Leistungsbereich nur in obligatorischen, nicht selektiv wählbaren Kursen entfalten könne (ebd., S.350). Dem gegenüber eruieren Maag Merki u.a. (2010) – allerdings mit dem Fokus auf die Unterrichtsqualität gerichtet – Effekte der Prüfungsorganisation in Grund- und Leistungskursen, und zwar beschränkt auf die Fächer Englisch und Mathematik (nicht aber in Deutsch und Biologie). Die Diskussion möglicher fach- und kursniveauspezifischer Effekte des Zentralabiturs steht erst am Anfang und bedarf der Vertiefung durch fachdidaktische Analysen.

Insgesamt zeigen sich bei den o.g. multiplen Mittelwertvergleichen über die drei Leistungsdomänen hinweg keine systematischen Lageunterschiede zwischen Ländern mit bzw. ohne zentrale Abschlussprüfungen im Abitur, die auf einen generellen Zentralprüfungseffekt schließen ließen. Dieser Eindruck verstärkt sich bei der Betrachtung der mittleren Leistungswerte der einzelnen Schulen in den jeweiligen Ländern (s. dazu die Abbildung 4 bis 6). Die Ergebnisse werden grafisch in Form von Strip Charts mit hinterlegten Dichteschätzern wieder gegeben. Auf der horizontalen Achse sind die Länder in zwei Gruppen unterteilt: solche mit ZA (in kleinen Lettern) und diejenigen ohne ZA (in großen Lettern). Innerhalb dieser beiden Gruppen sind die Länder nach der Höhe der mittleren Leistungswerte in absteigender Reihenfolge geordnet. Die Streifen repräsentieren die mittleren Leistungswerte der einzelnen untersuchten Gymnasien innerhalb der Länder. Das Leistungsniveau wird an der vertikalen Achse der Grafik abgetragen. Da sich die mittleren Leistungswerte einzelner Schulen

z.T. überlappen, wird die Dichteverteilung der Schulen je Bundesland ergänzend abgebildet.

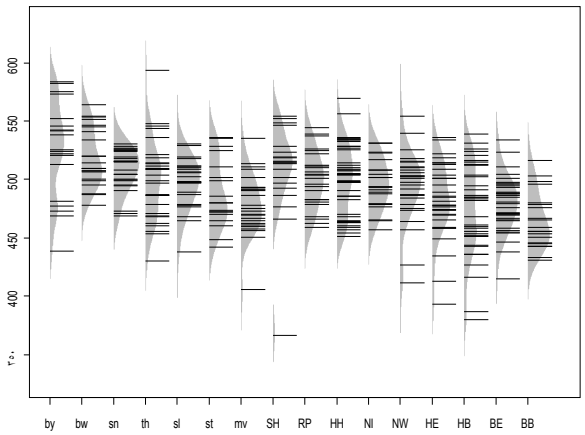
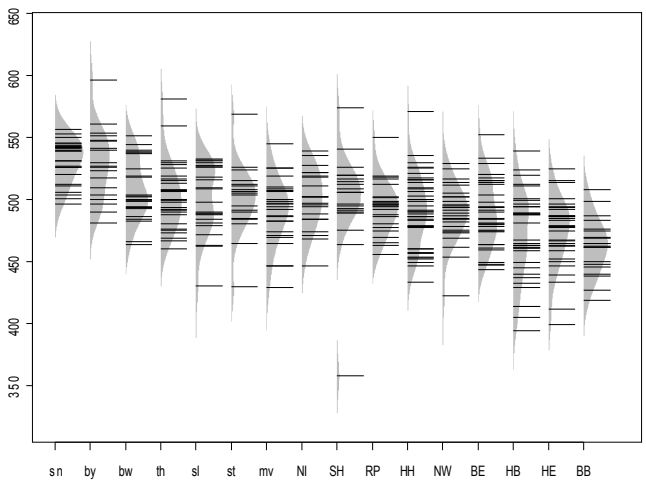


Abbildung 4: Stripchart Residuen Lesekompetenz – Arithmetisches Mittel der einzelnen Schulen nach Bundesländern

Abbildung 5: Stripchart: Residuen naturwissenschaftliche Kompetenz – Arithmeti-



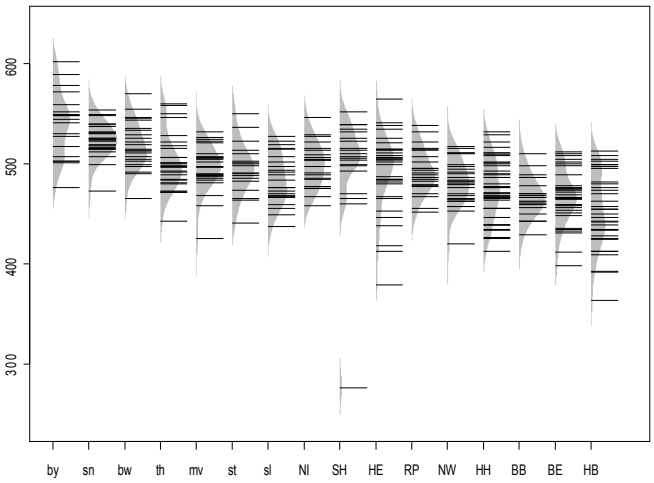


Abbildung 6: Stripchart: Residuen Mathematikkompetenz – Arithmetisches Mittel der einzelnen Schulen nach Bundesländern

Würde das ZA einen substanziellen Steuerungseffekt auf das mittlere Leistungsniveau haben, dann wäre zu erwarten, dass sich systematische Lageunterschiede in Bezug auf die bereinigten Testleistungswerte nicht nur zwischen Ländern, sondern gerade auch zwischen den Schulen der Länder mit bzw. ohne ZA zeigen. Wie die Strip Charts verdeutlichen, ist dies gerade nicht der Fall. Es lassen sich über die Leistungsdomänen hinweg keine systematischen Lageunterschiede auf der Schulebene beobachten, im Gegenteil: Die Überlappungen der Leistungsverteilungen sind erheblich. Allerdings deuten sich auch hier wieder tendenziell domänenspezifische Unterschiede an: Während sich bei den Residuen der Lesekompetenz selbst an den Rändern der Leistungsverteilung – d.h. zwischen den Schulen der besten Länder mit ZA und den Schulen der leistungsschwächsten Länder ohne ZA – erheblich Überschneidungen finden, sind diese (Kontrast-) Gruppen bei den Residuen der naturwissenschaftlichen Kompetenz und mehr noch bei den Residuen der Mathematikkompetenz deutlicher separiert. Insgesamt aber finden sich für das Gros der Bundesländer erheblich Überlappungen des mittleren Leistungsniveaus auf Schulebene. Gleiches gilt

im Übrigen – und dies sei an dieser Stelle nur nachrichtlich mitgeteilt – für die relative Leistungsstreuung auf Schulebene: Auch bei den Variationskoeffizienten der Residuen der einzelnen Schulen zeigen sich keine systematischen Lageunterschiede zwischen den Schulen der Länder mit bzw. ohne ZA in dem Sinne, dass sich erstere durchgängig durch ein homogeneres Leistungsniveau auszeichnen würden. Die starken Überlappungen des Leistungsniveaus auf Schulebene, unabhängig von der Prüfungsorganisation, stützen die plausible Einschätzung, dass die (Einzel-)Schule die zentrale Handlungsebene für die Leistungserstellung darstellt, die administrativen Steuerungsstrategien offensichtlich wenig zugänglich ist.

Die Erkenntnisse der grafischen Exploration mittels Stripcharts lassen sich durch mehrebenenanalytische Regressionen numerisch exakter fassen (vgl. Tab. 2). Die Ergebnisse mehrebenenanalytischer Random-Intercept-Only-Modelle (mit den Schulen als Ebene 2 und den Bundesländern mit bzw. ohne ZA als Ebene 3) zur Schätzung der Varianzkomponenten stützen die kritischen Befunde zu den angenommenen Leistungseffekten des ZA. Die Länderebene (mit/ohne ZA) erklärt maximal 4 Prozent der Varianz der adjustierten Testleistungen in Mathematik, 2.6 Prozent der Leistungen in den Naturwissenschaften und 0.8 Prozent der adjustierten Leseleistung[71]. Die maximale Varianzaufklärung durch die Prüfungsform bewegt sich unter dem Gesichtspunkt der praktischen Signifikanz damit in einer maximal kleinen bis trivialen Größenordnung. Der Steuerungseffekt durch das ZA in Bezug auf das mittlere Leistungsniveau scheint somit wenig ausgeprägt. Gemessen an der maximalen Varianzaufklärung erweist sich die einzelschulische Ebene als deutlich leistungsrelevanter.

Auch bei der Analyse der Unterrichts- bzw. Schülervariablen (u.a. Skala ‚student-teacher relation', Skala ‚competitive learning', Skala ‚cooperative learning', Skala ‚teacher support', Skala ‚memorisation') zeigen sich auf der Ebene der Länder mit und ohne ZA zunächst insgesamt signifikante, aber überwiegend erwartungswidrige und triviale Unterschiede (siehe Tab. 1). So weisen z.B. die Länder ohne ZA insgesamt tendenziell höhere Skalenwerte in den Bereichen ‚instrumental motivation', ‚memorisation' und ‚competitve learning' auf

[71] Die angegebenen t-Werte der Zwischenvarianzen lassen sich nicht eindeutig interpretieren (und sind nur der Vollständigkeit halber aufgeführt), da die mehrebenenanalytischen Analysen – aus Vergleichsgründen – mit den o.g. Populationsgewichten durchgeführt wurden. Tendenziell sind die t-Werte also deutlich niedriger anzusetzen und die Zwischenvarianzen der Ebene 3 (Länder mit vs. ohne Zentralabitur) somit in hohem Maße als nicht signifikant einzuschätzen.

als die Länder mit ZA. Dies widerspricht den weiter oben berichteten, hypothetisierten Steuerungswirkungen des ZA auf das Unterrichtsgeschehen.

Bei den paarweisen Ländervergleichen hinsichtlich dieser Variablen – dies sei hier nur nachrichtlich mitgeteilt – lässt sich ebenso kein systematisches Muster erkennen. Die maximale Varianzaufklärung der Länderebene (mit/ohne ZA) hinsichtlich der Unterrichtsvariablen schließlich ist sehr gering (fast durchweg unter einem Prozent, vgl. Tab. 2).

	Ebene 2: 412 Gymnasien		**Ebene 3: Länder mit/ohne Zentralabitur**	
Merkmal	**max. Varianz-aufklä-rung (ICC in %)**	**random inter-cept t-Wert**	**max. Varianz-aufklärung (ICC in %)**	**random in-tercept t-Wert**
Residuen Mathematik	10.4	13.82	4.0	0.70
Residuen Lesen	10.1	13.79	0.8	0.66
Residuen Naturwissenschaften	8.5	13.70	2.6	0.70
Skala Instrumental Motivation	-	-	-	-
Skala Mathematics Self-Efficiacy	6.2	13.46	0.1	0.64
Skala Memorisation	4.2	13.04	1.3	-
Skala Competitive Learning	5.5	13.29	0.1	0.86
Skala Cooperative Learning	5.0	13.33	0.2	0.55
Skala Interest and Enjoyment of Mathemat-ics	5.5	13.29	0.0	0.00
Skala Teacher Support	-	-	-	-
Skala Student-Teacher Relation	8.8	13.68	0.4	1.35

Tabelle 2: Random Intercept Only-Modelle zur Schätzung der Varianzkomponenten einzelner Merkmale[72]

Die Art der Prüfungsorganisation führt offensichtlich nicht zu (erwartungstreuen und substanziellen) differenziellen Unterrichtseffekten (in Bezug auf die hier untersuchten Unterrichtsmerkmale). Insofern erscheint fraglich, dass und vor

[72] Anmerkung: Bei Zellen mit – hat der Algorithmus keine stabilen Schätzungen ermöglicht.

allem wie und über welche Vermittlungsebenen das ZA Leistungseffekte induziert. Die Analysen zu den Unterrichtsvariablen stützen u.E. die Einschätzung, dass es sich bei Leistungsunterschieden – soweit sich diese zwischen einzelnen Ländern mit unterschiedlicher Prüfungsorganisation finden lassen – möglicherweise eher um das Ergebnis unbeobachteter Heterogenität als um eine kausale Wirkung verschiedener Abschlussprüfungen handelt.

4. Zusammenfassung und Diskussion

Die Ergebnisse der Sekundäranalyse der PISA-E-2003-Daten zeigen, dass sich generelle und substanzielle Effekte des ZA nicht hinreichend und erwartungskonform beobachten lassen, weder in Bezug auf die untersuchten Leistungsvariablen noch hinsichtlich ausgesuchter Unterrichtsvariablen. Dort, wo sich noch signifikante Unterschiede zeigen, sind diese entweder erwartungswidrig, von der Größenordnung her eher klein oder domänenspezifisch variierend. Angesichts der befragten Altersgruppe kann die Schlussfolgerung gezogen werden, dass generelle und substanzielle Effekte des ZA auf die Gruppe der 15-Jährigen offensichtlich nicht zu konstatieren sind und es somit keine ‚Fernwirkung' des ZA auf jüngere Alterskohorten im gymnasialen Bildungsgang gibt. Die Reanalysen machen deutlich, dass sich mittels der PISA-E-Daten von 2003 die Wirksamkeit zentraler Abschlussprüfungen nicht überzeugend belegen lässt. Damit bleibt die Hypothese der generellen und substanziellen Leistungseffektivität des ZA für den bundesrepublikanischen Diskussionszusammenhang weiterhin unbestätigt und die Befunde entziehen sich klaren bildungspolitischen Empfehlungen. Mögliche Leistungseffekte differenzierter Prüfungsorganisationsformen bleiben ein Forschungsdesiderat, national wie international.

Der Bildungsproduktionsfunktionsansatz erscheint als theoretisches Fundament in diesem Zusammenhang weniger geeignet. Das Modell unterstellt einen direkten kausalen Wirkungszusammenhang zwischen unabhängigen Variablen (wie Ressourcen, institutionelle Kontexte) und der abhängigen Zielgröße, z.B. in Form von Testleistungen. Bei diesem Erklärungsansatz von Leistungsergebnissen handelt es sich um eine Handlungstheorie ohne Handlungsakteure, denn der eigentliche Vermittlungsprozess der Handlungssubjekte im Unterrichtsgeschehen bleibt als Black Box unberücksichtigt. Zudem wird unterstellt, dass sich systemische Steuerungsintentionen, wie z.B. das ZA als institutioneller Kontext, quasi ‚eins zu eins', ohne substanzielle Reibungsverluste oder Brechungen durch die beteiligten Akteursgruppen, in Leistungseffektivität umsetzen lassen.

Dabei gehört es gerade zu den wenigen konsensfähigen Postulaten der Schulwirksamkeitsforschung, dass das konkrete Unterrichtshandeln (im Rahmen der Klasse) den zentralen Prädiktor der Leistungseffektivität darstellt – oder statistisch formuliert: Die Unterrichtsebene (in Form von Klassen) erklärt die höchsten Varianzanteile von Leistungsmerkmalen. Insofern ist Schuleffektivität nur unter Berücksichtigung dieser zentralen Handlungsebene sinnvoll zu modellieren.

Hier bietet der Educational Governance-Ansatz (vgl. z.B. Altrichter, Brüsemeister & Wissinger 2007) konzeptionell die Möglichkeit, die Subjekte oder unmittelbaren schulischen Handlungsakteure ins Zentrum der Schuleffektivitätsmessung zu setzen. Dieser Ansatz geht davon aus, dass Steuerungshandeln auf der Bildungssystemebene auf unterschiedliche institutionelle und individuelle Handlungsbedingungen hin adaptiert wird. Das Konstrukt Schule wird in der Educational Governance als Mehrebenensystem gesehen, in welchem individuelle und korporative Akteure auf unterschiedlichen Ebenen (z.B. administrative Ebene, Schulebene, Unterrichtsebene, Individualebene) jeweils auf ihre an den eigenen Zielvorstellungen und Deutungsmustern orientierten Handlungsmuster hin selbstreferenziell adaptieren.

Vor diesem Hintergrund scheint eine empirische Schuleffektivitätsforschung, die sich der Konzeptualisierungen und Konstrukte des Governance-Ansatzes annimmt, geeignet, das Wirkungsgefüge von Steuerungsintentionen (wie dem ZA) und faktischen Leistungseffekten realitätsnah zu modellieren. Dies setzt freilich eine gänzlich andere, bedeutend komplexere Indikatorisierung voraus. Erste empirische Befunde im Kontext einer DFG-geförderten Untersuchung zur Einführung des ZA in Bremen und Hessen mit dem Titel „Implementation und Auswirkungen neuer Steuerungsstrukturen im Schulwesen am Beispiel zentraler Abiturprüfungen" (Maag Merki, Klieme & Holmeier 2008; Maag Merki 2008) deuten auf die Tragfähigkeit einer Governance basierten empirischen Schulforschung und Schuleffektivitätsforschung hin.

Autoren

Dr. Rainer Block
Referent für Begleitforschung
Stiftung „Haus der kleinen Forscher" Berlin
(vormals Universität Duisburg-Essen)
E-Mail: rainer.block@haus-der-kleinen-forscher.de, rainer.block@arcor.de

Esther Dominique Klein
Universität Duisburg-Essen, Fakultät für Bildungswissenschaften
Arbeitseinheit Bildungssystem-und Schulentwicklungsforschung
E-Mail: dominique.klein@uni-due.de

Prof. Dr. Isabell van Ackeren
Universität Duisburg-Essen, Fakultät für Bildungswissenschaften
Leiterin der Arbeitseinheit Bildungssystem-und Schulentwicklungsforschung
E-Mail: isabell.van-ackeren@uni-due.de
URL: http://www.uni-due.de/bifo/vanackeren.php

Dr. Svenja Mareike Kühn
Universität Duisburg-Essen, Fakultät für Bildungswissenschaften
Arbeitseinheit Bildungssystem-und Schulentwicklungsforschung
E-Mail: svenja.kuehn@uni-due.de

Literatur

Ackeren, I. van (2007). Zentrale Abschlussprüfungen. Entstehung, Struktur und Steuerungsperspektiven. Pädagogik, 59, 3, 12-15

Altrichter, H., Brüsemeister, T. & Wissinger, J. (Hrsg.) (2007). Educational Governance. Handlungskoordinaten und Steuerung im Bildungssystem. Wiesbaden: Verlag für Sozialwissenschaften

American Institutes for Research & Cohen, J. (2009). AM Statistical Software Version 10.06.03. Washington. Zugriff am 10. April 2010. http://am.air.org/

Barnette, J.J.(2006). Effect Size and Measures of Association. School of Public Health University of Alabama at Birmingham. Zugriff am 10. April 2010. http://www.eval.org/SummerInstitute/06SIHandouts/SI06.Barnette.TR2.Online.pdf

Baumert, J. & Watermann, R. (2000). Standardsicherung durch die Abiturprüfung. Zentralabitur oder dezentrale Prüfungsorganisation? In Baumert, J., Bos, W. & Lehmann, R. (Hsrg.). TIMSS III. Dritte internationale Mathematik- und Natur-

wissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn, Band 2, 341-351. Opladen: Leske und Budrich
Baumert, J., Carstensen, C.H. & Siegle, T. (2005). Wirtschaftliche, soziale und kulturelle Lebensverhältnisse und regionale Disparitäten des Kompetenzerwerbs. In PISA-Konsortium Deutschland (Hrsg.). PISA 2003. Der zweite Vergleich der Länder in Deutschland – Was wissen und können Jugendliche? 323-365.Münster: Waxmann
Bishop, J. (1998). The Effect of Curriculum-based External Exit Exams on Student Achievement. Journal of Economic Education, 29, 2, 172-182.
Bishop, J. & Wößmann, L. (2004). Institutional Effects in a Simple Model of Educational Production. Education Economics 12, 1, 17-38
Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. New York: Erlbaum
Cohen, J. (1992). A Power Primer. Psychological Bulletin, 112, 155-159
Doran, H.C. & Izumi, L.T. (2004). Putting Education to the Test: A Value-Added Model for California. San Francisco: Pacific Research Institute
Fuchs, T. & Wößmann, L. (2007). What accounts for International Differences in Student Performance? A Re-examination using PISA data. Empirical Economics 32, 2, 433-464
Hanushek, E.A. & Wößmann, L. (2007). The Role of Education Quality in Economic Growth. Washington: World Bank
Hill, P. (2001). Perspectives on Education. Teaching and School Effectiveness. Victoria: Department of Education, Employment and Training.
Jürges, H., Schneider, K. & Büchel, F. (2003). The Effects of Central Examinations on Student Achievement: Quasi-experimental Evidence from TIMSS Germany. München: ifo Working Paper No. 939
Klein, D., Kühn, S., Ackeren, I. van & Block, R. (2009). Wie zentral sind zentrale Prüfungen? Abschlussprüfungen am Ende der Sekundarstufe II im nationalen und internationalen Vergleich. Zeitschrift für Pädagogik 55, 4, 596-621
Maag Merki, K. (2008). Die Einführung des Zentralabiturs in Bremen – Eine Fallanalyse. Die Deutsche Schule 100, 3, 357-368
Maag Merki, K., Klieme, E. & Holmeier, M. (2008). Unterrichtsgestaltung unter den Bedingungen zentraler Abiturprüfungen. Zeitschrift für Pädagogik 54, 6, 791-808
Maag Merki, K., Holmeier, M., Jäger, D., Oerke, B. (2010). Die Effekte der Einführung zentraler Abiturprüfungen auf die Unterrichtsgestaltung in Leistungskursen in der gymnasialen Oberstufe. Unterrichtswissenschaft 38, 2, 173-192
OECD (2008). Measuring Improvements in Learning Outcomes. Best Practices to Assess the Value-Added of Schools. Paris: OECD Publications
PISA-Konsortium Deutschland (Hrsg.) (2005). PISA 2003. Der zweite Vergleich der Länder in Deutschland – Was wissen und können Jugendliche? Münster: Waxmann
PISA-Konsortium Deutschland (Hrsg.) (2006). PISA 2003. Dokumentation der Erhebungsinstrumente. Münster: Waxmann

PISA-Konsortium Deutschland (Hrsg.) (2008). PISA 2006 in Deutschland. Die Kompetenzen der Jugendlichen im dritten Ländervergleich. Münster: Waxmann

Robinson, W.S. (1950). Ecological Correlations and Behaviour of Individuals. American Sociological Review, 15, 351-357

Scheerens, J., Bosker, R.J. (1997). The Foundations of Educational Effectiveness. Oxford: Elsevier Science Ltd.

Schulministerium NRW (2009). Fragen und Antworten zum Zentralabitur. Zugriff am 10. April 2010. http://www.standardsicherung.schulministerium.nrw.de/abitur/abitur-gymnasiale-oberstufe/fragen-und-antworten/

Schümer, G. & Weiß, M. (2008). Bildungsökonomie und Qualität der Schulbildung. Kommentar zur bildungsökonomischen Auswertung von Daten aus internationalen Schulleistungsstudien. Frankfurt a. M.: GEW

Watermann, R. & Stanat, P. (2004). Schulrückmeldungen in PISA 2000: Sozialnorm- und kriteriumsorientierte Rückmeldeverfahren. In Kohler, B. & Schrader, F.W. (Hrsg.), Ergebnisrückmeldung und Rezeption. Empirische Pädagogik 18/1 Themenheft (S.40- 61). Landau:Verlag Empirische Pädagogik

Wößmann, L. (2005). Ursachenkomplexe der PISA-Ergebnisse: Untersuchungen auf der Basis der internationalen Mikrodaten. München: ifo-Arbeitspapier Nr.16

Wößmann, L. (2007). Fundamental Determinants of School Efficiency and Equity: German States as a Microcosm for OECD Countries. München: CESIFO Working Paper No. 1981

Wößmann, L. (2008). Zentrale Abschlussprüfungen und Schülerleistungen. Zeitschrift für Pädagogik 54, 6, 810-826

Online zugänglich unter:

Rainer Block, Esther Dominique Klein, Isabell van Ackeren, Svenja Mareike Kühn (2011). Leistungseffekte des Zentralabiturs. In: bildungsforschung, Jahrgang 8, Ausgabe 1, URL: http://www.bildungsforschung.org/