

Die Auswirkungen zentraler Abschlussprüfungen auf die Schulleistung – Quasi-experimentelle Befunde aus der deutschen TIMSS-Stichprobe*

Von Felix Büchel**, Hendrik Jürges*** und Kerstin Schneider****

Zusammenfassung: Die Third International Mathematics and Science Study (TIMSS) zeigt, dass die Schulleistungen in Bundesländern mit zentralen Abschlussprüfungen deutlich besser sind als in den anderen Bundesländern. Schulabschlüsse der Sekundarstufe werden in Deutschland in Abhängigkeit vom Bundesland mit oder ohne zentrale Abschlussprüfungen vergeben. Diese im internationalen Vergleich seltene institutionelle Variation kann für die Schätzung des kausalen Effekts zentraler Abschlussprüfungen auf die Schulleistung genutzt werden. Wir stellen einen Differenzen-in-Differenzen-Schätzer vor, bei dem wir uns den quasi-experimentellen Charakter der deutschen TIMSS-Stichprobe für die Sekundarstufe I zunutze machen. Es stellt sich heraus, dass die Leistungsunterschiede zwischen den Ländern nur zum Teil auf die zentralen Abschlussprüfungen selbst zurückzuführen sind. Die Ergebnisse unserer Studie quantifizieren den positiven Effekt zentraler Prüfungen auf die Schulleistung auf mindestens ein Drittel des Äquivalents eines Schuljahres.

Summary: This paper makes use of the regional variation in schooling legislation within the German secondary education system to estimate the causal effect of central exit examinations on student performance. We propose a difference-in-differences framework that exploits the quasi-experimental nature of the German TIMSS middle school sample. The estimates show that students in federal states with central exit examinations clearly outperform students in other federal states, but that only part of this difference can be attributed to central exit examinations themselves. Our results suggest that such examinations increase student achievement by about one third school year equivalent.

1 Einleitung

Das schlechte Abschneiden der deutschen Schüler bei TIMSS und PISA löste intensive bildungspolitische Debatten über die Notwendigkeit einer Reform des deutschen Schulsystems aus. Reformen könnten zum einen in der besseren finanziellen Ausstattung des Schulsystems, zum anderen in einer Veränderung der institutionellen Rahmenbedingungen bestehen. Mehr finanzielle Mittel allein scheinen jedoch wenig erfolgversprechend zu sein, wie die Diskussion um die Leistungseffekte der Klassengröße zeigt (Hanushek 1999, Hoxby 2000a, Wößmann und West 2002, Jürges und Schneider 2003): Zwar lassen sich kleine positive Effekte nachweisen, doch sind allgemeine Verkleinerungen der Klassengröße sehr kostenintensiv.

* Wir danken Dr. Rainer Watermann, Max-Planck-Institut für Bildungsforschung, Berlin, für die Ermöglichung des administrativen und technischen Zugangs zu den dieser Studie zugrunde liegenden regionalen Daten sowie Prof. Dr. Jürgen Baumert für seine Sondergenehmigung zur Verwendung dieser Daten. Die regionalen TIMSS-Daten unterliegen derzeit noch qualifizierten Datenschutzbestimmungen und werden interessierten Forschern erst nach ihrer faktischen Anonymisierung zugänglich sein.

** Max-Planck-Institut für Bildungsforschung, Berlin, E-Mail: buechel@mpib-berlin.mpg.de

*** MEA, Universität Mannheim, E-Mail: juerges@mea.uni-mannheim.de

**** Universität Dortmund, E-Mail: kerstin.schneider@wiso.uni-dortmund.de

Institutionelle Rahmenbedingungen des Schulsystems dagegen scheinen schon mehr zur Erklärung der Schulleistung beizutragen. Veränderungen bestimmter Merkmale des Schulsystems, mit denen leistungsfördernde Anreize geschaffen werden, könnten mehr bewegen als die bloße Erhöhung der dem Bildungssektor zugewiesenen Ressourcen.¹ Eine der Fragen, die in diesem Zusammenhang diskutiert wurden, ist Wettbewerb zwischen Schulen (Hoxby 2002b). Ein weiterer Aspekt, und dies ist das Thema dieses Beitrags, sind zentral festgelegte Standards in Form zentraler Abschlussprüfungen.

Die wirtschaftswissenschaftliche Literatur kommt tendenziell zu dem Schluss, dass zentrale Abschlussprüfungen die Schulleistung erhöhen und sogar die allgemeine Wohlfahrt steigern können (Costrell 1997, Effinger und Polborn 1999). Zentrale Prüfungen setzen eher als dezentrale die richtigen Anreize für Schülerinnen und Schüler,² Lehrkräfte und Schulen (z. B. Bishop 1997, 1999). Die Schüler z. B. profitieren, weil Ergebnisse von zentralen Abschlussprüfungen auf dem Arbeitsmarkt schon aufgrund ihrer breiteren Vergleichbarkeit bessere Signale sind als Ergebnisse von nicht zentralen Prüfungen. Außerdem entfällt, wenn zum Abschluss der Schullaufbahn externe Standards erfüllt werden müssen, der Anreiz zu einer Kartellbildung in der Klasse, bei der sich die Schüler, womöglich mit stillschweigendem Einverständnis der Lehrkräfte, auf ein niedrigeres Leistungsniveau einstellen. Testergebnisse könnten auch zur regelmäßigen Kontrolle der Unterrichtsqualität und somit der Lehrkräfte herangezogen werden. Schließlich könnte gar der Ruf ganzer Schulen von den Leistungen ihrer Schüler abhängig gemacht werden, so dass – vorausgesetzt, die aggregierten Ergebnisse der zentralen Abschlussprüfungen sind öffentlich zugänglich – gute Schulen bessere Schüler anziehen.

Mit TIMSS-Daten durchgeführte, internationale vergleichende Studien zum Leistungseffekt zentraler Abschlussprüfungen kommen zu der Aussage, dass das Vorhandensein zentraler Prüfungen die Schulleistung signifikant in der Größenordnung etwa eines Schuljahrs erhöht (Bishop 1997, 1999). Allerdings ist der Befund eines positiven Leistungseffekts in internationalen Vergleichen auf Grundlage von TIMSS nicht sonderlich robust (Jürges und Schneider 2003). Studien von Leistungsunterschieden in Ländern mit der notwendigen institutionellen Variation scheinen erfolgversprechender zu sein. Bishop (1997, 1999) legt neben internationalen Vergleichen auch noch Ergebnisse aus kanadischen Mikrodaten vor. Ähnlich wie in Deutschland gibt es in Kanada ein gemischtes System, bei dem in manchen Provinzen die Highschool mit zentralen Abschlussprüfungen endet und in anderen nicht. Bishops Schätzungen zufolge lässt sich der Leistungseffekt zentraler Abschlussprüfungen in Kanada je nach Fach mit drei bis vier Fünfteln des Äquivalents einer Jahrgangsstufe angeben.

Allerdings ist es wichtig, zwischen einfacher Korrelation und Kausalität zu unterscheiden. So könnten in Ländern oder Regionen mit zentralen Abschlussprüfungen Bildung und Leistung ganz allgemein höhere Priorität haben. In diesem Falle wären sowohl die hohe durchschnittliche Schulleistung als auch die zentralen Abschlussprüfungen lediglich Ausdruck einer Präferenz der Bevölkerung für gute Schulbildung. Frühere Beiträge haben versucht, diesem Problem Rechnung zu tragen, indem sie der Frage nachgingen, ob sich Länder mit zentralen Abschlussprüfungen von Ländern ohne solche Prüfungen auch noch in anderen

¹ Für entsprechende Überlegungen im Kontext der internationalen TIMSS-Daten vgl. Jürges und Schneider (2003).

² Wir verwenden im Folgenden im Interesse der besseren Lesbarkeit nur die männliche Form.

Dimensionen als nur der mittleren Schulleistung unterscheiden, etwa in der Unterrichtsdisziplin oder im Fernbleibeverhalten der Schüler (Bishop 1997). Die Daten gaben jedoch letztlich keine überzeugende Identifikationsstrategie her, so dass hier wichtige Fragen offen bleiben mussten.

Die föderale Struktur des deutschen Bildungssystems stellt eine besonders geeignete Quelle institutioneller Variation zur Identifikation des Kausaleffekts zentraler Abschlussprüfungen auf die Schulleistung dar. Frühere Studien (Baumert und Watermann 2000) konnten Leistungsunterschiede in der Sekundarstufe II nur bei Schülern in Mathematikgrundkursen zeigen. Die Sekundarstufe II des deutschen Schulsystems ist jedoch nach unserer Auffassung aufgrund der vielfältigen Wahlmöglichkeiten der Schüler und des damit verbundenen Selektivitätsproblems für eine Analyse des Effekts von zentralen Abschlussprüfungen nur eingeschränkt geeignet. Wir konzentrieren uns deshalb auf den Effekt von Abschlussprüfungen am Ende der Sekundarstufe I, die interessanterweise in der öffentlichen Debatte zum Thema Abschlussprüfungen kaum Beachtung finden.

Unser Beitrag ist wie folgt gegliedert: Abschnitt 2 beschreibt die Praxis zentraler Abschlussprüfungen in Deutschland und unsere Datenbasis. Abschnitt 3 behandelt Fragen der Schätzung und Identifikation. Die Ergebnisse werden in Abschnitt 4 dargestellt. Abschnitt 5 enthält knappe Schlussfolgerungen.

2 Die Praxis zentraler Abschlussprüfungen und die TIMSS-Daten

Die Diskussion zentraler Abschlussprüfungen in Deutschland beschränkt sich in der Hauptsache auf das Zentralabitur, also den Abschluss der Sekundarstufe II. Nur selten werden die Abschlüsse der Sekundarstufe I einbezogen, obwohl etwa zwei Drittel der Schüler der Sekundarstufe I Haupt- und Realschulen besuchen. Bei Abschluss der Sekundarstufe I erhalten die Schüler der Haupt- oder Realschule ein Abschlusszeugnis, sofern sie die 9. oder 10. Klasse erfolgreich abgeschlossen haben. In einigen Bundesländern müssen die Schüler eine zentrale Abschlussprüfung ablegen (Tabelle 1 stellt den Status quo für 1995 dar, das Erhebungsjahr der TIMSS-Daten). Gymnasialschüler erhalten am Ende der Sekundarstufe I kein Abschlusszeugnis, erwerben aber die Qualifikation zum Besuch der gymnasialen Oberstufe.

Am häufigsten sind zentrale Abschlussprüfungen am Ende der Sekundarstufe II. So gab es 1995 in sieben der 16 deutschen Bundesländer ein zentrales Abitur (Abschluss der Sekundarstufe II mit Berechtigung zum Hochschulstudium) auf Landesebene. Diese Länder konzentrieren sich im Süden (Baden-Württemberg, Bayern, Saarland) und Osten Deutschlands (Mecklenburg-Vorpommern, Sachsen, Sachsen-Anhalt, Thüringen). In den anderen Bundesländern gab es dezentrale Systeme, bei denen die Aufgaben für die Abschlussprüfung von den Lehrkräften ausgearbeitet und von der Schulaufsichtsbehörde einzeln genehmigt werden. In sechs Ländern gab es zentrale Prüfungen zum Realschulabschluss und in nur vier Ländern auch zum Hauptschulabschluss.

Abschlussprüfungen umfassen nicht immer alle Unterrichtsfächer. Beim Abitur können die Schüler innerhalb eines bestimmten Rahmens drei oder vier Prüfungsfächer wählen. Bei Prüfungen zum Real- und Hauptschulabschluss sind Deutsch und Mathematik immer Pflichtfächer. Zur Bewertung des Leistungseffekts von zentralen Abschlussprüfungen werden wir uns auf die Mathematikleistung an der Haupt- und Realschule konzentrieren: Maße

Tabelle 1

Zentrale Abschlussprüfungen nach Bundesland und Art des Abschlusses sowie Anteil der Schüler nach Schulform/Art des Abschlusses
Stand von 1995

| | Hauptschulabschluss | Realschulabschluss | Abschluss Sekundarstufe II (Abitur) | Schüler in der 8. Klasse/Haupt-/Realschule | Schüler in nicht zum Abitur führenden Bildungsgängen** |
|-----------------------------|---------------------|--------------------|-------------------------------------|--|--|
| | | | | Anteil in % | |
| Baden-Württemberg (BW) | D/M/F/A | D/M/F | J | 71,8 | 71,6 |
| Bayern (BY) | D/M/F/N*/A | D/M/F/N*/A | J | 73,2 | 73,1 |
| Berlin (BE) | | | | 67,2 | 57,2 |
| Brandenburg (BB) | | | | 71,7 | 54,8 |
| Bremen (HB) | | | | 69,1 | 63,9 |
| Hamburg (HH) | | | | 65,6 | 56,7 |
| Hessen (HE) | | | | 67,2 | 61,4 |
| Mecklenburg-Vorpommern (MV) | | D/M/F | J | 72,2 | 70,7 |
| Niedersachsen (NI) | | | | 73,2 | 72,0 |
| Nordrhein-Westfalen (NW) | | | | 69,6 | 64,7 |
| Rheinland-Pfalz (RP) | | | | 72,0 | 70,5 |
| Saarland (SA) | | | J | 71,8 | 66,0 |
| Sachsen (SN) | D/M/F | D/M/N | J | 69,9 | 69,9 |
| Sachsen-Anhalt (ST) | | D/M | J | 67,6 | 67,2 |
| Schleswig-Holstein (SH) | | | | 73,7 | 71,5 |
| Thüringen (TH) | D/M | D/M/F | J | 67,7 | 67,4 |

D = Deutsch; M = Mathematik; F = Fremdsprache (meist Englisch); N = Naturwissenschaften; A = Andere; J = Jedes für die schriftlichen Prüfungen gewählte Fach.

* Wahlfach.

** Schüler in der 8. Klasse der Haupt- und Realschule plus zwei Drittel aller Schüler in der 8. Klasse der Gesamtschule.

Quelle: Eigene Auswertungen von TIMSS-Deutschland.

für die Mathematikleistung liegen uns mit den TIMSS-Daten vor. Andere Fächer bei zentralen Abschlussprüfungen sind Fremdsprachen (meist Englisch) oder – weniger häufig – naturwissenschaftliche Fächer.

Unsere Datenbasis, die TIMSS-Deutschland-Stichprobe, enthält Daten zu fast 6 000 Schülern, die im Schuljahr 1994/95 in 7. und 8. Klassen von 137 Schulen erhoben wurden. Die Schüler kommen aus 14 der 16 Bundesländer (Baden-Württemberg und Bremen nahmen nicht teil). Wir verwenden hier eine Unterstichprobe, die nur Schüler in Haupt- und Realschulen (bzw. Mittelschulen) enthält. Diese umfasst insgesamt 3 197 Schüler mit vollständigen Daten aus 84 Schulen bzw. 168 Klassen. Außer den eigentlichen Testergebnissen für Mathematik und Naturwissenschaften enthalten die TIMSS-Daten ein breites Spektrum von Variablen zu sozialem Hintergrund und Einstellungen der Schüler wie auch zu Lehrkräften und Schulen. Trotz der Fülle des vorliegenden Materials werden wir einen eher eng gefassten Ansatz verfolgen und nur eine begrenzte Anzahl Kontrollvariablen zum Hintergrund von Schülern und Schulen mit einem hohen Erklärungswert für die Schulleistung auswählen.

Tabelle 2 enthält Variablendefinitionen und deskriptive Statistiken, aufgeschlüsselt nach der Art der Abschlussprüfung.³ Im Gegensatz zu Veröffentlichungen, bei denen der inter-

³ Zu beachten ist, dass wir im gesamten Beitrag mit Gewichtung der Stichprobe arbeiten.

Tabelle 2

Deskriptive Statistik (gewichtet)

| Variable | Schüler in Ländern ohne zentrale Abschlussprüfungen | | Schüler in Ländern mit zentralen Abschlussprüfungen | |
|--|---|--------------------|---|--------------------|
| | Mittel | Standardabweichung | Mittel | Standardabweichung |
| Leistung in Mathematik (Rasch-Score) | -0,134 | 0,943 | 0,213 | 0,971 |
| Leistung in Naturwissenschaften | -0,114 | 0,966 | 0,216 | 0,969 |
| Geschlecht (1 = Mädchen) | 0,486 | | 0,459 | |
| Bücher zu Hause: 0–10 | 0,183 | | 0,188 | |
| Bücher zu Hause: 11–25 | 0,291 | | 0,333 | |
| Bücher zu Hause: 26–100 | 0,168 | | 0,175 | |
| Bücher zu Hause: 101–200 | 0,234 | | 0,209 | |
| Bücher zu Hause: 200+ | 0,123 | | 0,096 | |
| Immigrantenkind (beide Elternteile im Ausland geboren) | 0,191 | | 0,077 | |
| Schulform (1 = Realschule) | 0,522 | | 0,541 | |
| Klasse (1 = 8. Klasse) | 0,498 | | 0,478 | |
| Mindestens einmal Klasse wiederholt | 0,359 | | 0,246 | |
| Arbeitet zu Hause für Mathematik | 0,877 | | 0,838 | |
| Arbeitet zu Hause für Naturwissenschaften | 0,829 | | 0,784 | |
| Nachhilfestunden Mathematik | 0,149 | | 0,096 | |
| Nachhilfestunden Naturwissenschaften | 0,077 | | 0,058 | |
| Gewöhnlich gut in Mathematik: stimme voll zu | 0,400 | | 0,335 | |
| Gewöhnlich gut in Biologie: stimme voll zu | 0,315 | | 0,358 | |
| Ostdeutschland | 0,042 | | 0,455 | |
| Kumulierte Mathematikstunden (in 1 000) | 1,198 | 0,121 | 1,292 | 0,105 |
| N Beobachtungen | 1 834 | | 1 363 | |

Quelle: Eigene Auswertungen von TIMSS-Deutschland.

nationaler Vergleich von Schulleistungen im Mittelpunkt steht, verwenden wir als abhängige Variablen nicht die international standardisierten Werte für Mathematik (bzw. in manchen Analysen für Naturwissenschaften). Wir nutzen vielmehr, um eine bessere Vergleichbarkeit innerhalb Deutschlands zu gewährleisten, die separat für Deutschland berechneten, von uns noch auf Mittelwert = 0 und Varianz = 1 standardisierten Rasch-Scores. Die Höhe der geschätzten Regressionskoeffizienten wird damit direkt als Standardabweichung interpretierbar sein.

Der auffallendste Unterschied zwischen Schülern mit und ohne zentrale Abschlussprüfungen ist bei den Leistungen in Mathematik und den naturwissenschaftlichen Fächern erkennbar (Tabelle 2). Die Werte von Schülern, die zentral geprüft werden, liegen im Durchschnitt um etwa eine Drittel-Standardabweichung über denen von Schülern ohne zentrale Prüfungen. Diese Unterschiede spiegeln sich nicht in allen Hintergrundvariablen wider. So ist der Bildungshintergrund, gemessen an der Zahl der im Elternhaus vorhandenen Bücher, von Schülern aus Bundesländern mit zentralen Abschlussprüfungen und ohne diese recht ähnlich.⁴ In der Gruppe ohne zentrale Abschlussprüfungen finden sich jedoch weit mehr

⁴ Die Verwendung der Anzahl der Bücher im Elternhaus als Indikator für den Bildungshintergrund von Schülern mag auf den ersten Blick merkwürdig anmuten. Diese Variable hat jedoch einen mehr als doppelt so hohen Erklärungsgehalt für die Leistungen der Schüler als der höchste formale Bildungsabschluss der Eltern.

Kinder von Immigranten als in der Gruppe mit zentralen Abschlussprüfungen. Dies erklärt sich vor allem dadurch, dass es in Ostdeutschland, wo die meisten Bundesländer zentrale Abschlussprüfungen haben (ein Erbe des früheren DDR-Bildungssystems), relativ wenige Immigranten gibt.

Hervorzuheben sind noch zwei weitere Unterschiede zwischen Schülern mit und Schülern ohne zentrale Prüfungen: Erstens ist die kumulierte Anzahl der Mathematikstunden – berechnet nach den amtlichen Stundenplänen aller Bundesländer (Frenck 2001) – bei Schülern in Ländern ohne zentrale Abschlussprüfungen erheblich niedriger als in anderen. Zweitens stimmen Schüler ohne zentrale Abschlussprüfungen häufiger uneingeschränkt der Aussage zu: „Ich bin gewöhnlich gut in Mathematik“, schneiden jedoch – gemessen an ihren Leistungen in TIMSS – schlechter ab als ihre Mitschüler mit zentralen Prüfungen. Diese Form der systematisch verzerrten Selbsteinschätzung könnte durch höhere Leistungsstandards erklärt werden, an denen sich Schüler in Ländern mit zentralen Abschlussprüfungen messen müssen.

3 Identifikation und Schätzung

Anhand der deutschen TIMSS-Daten, ergänzt um Informationen zu Bundesland und Schulform, ist eine Identifikation des Kausaleffekts von zentralen Abschlussprüfungen auf die Schulleistung auf verschiedene Weise möglich. Diese beruhen jeweils auf unterschiedlichen Grundannahmen und weisen entsprechend unterschiedliche Stärken und Schwächen auf.

Der simpelste Ansatz besteht in der Berechnung *einfacher Differenzen* zwischen der Durchschnittsleistung in Ländern mit und in Ländern ohne zentrale Abschlussprüfungen, bei Kontrolle des Schülerhintergrunds und anderer einschlägiger Variablen, z. B. der Gesamtunterrichtszeit, die auf Mathematik und Naturwissenschaften verwendet wird.⁵ Einfache Differenzen sind allerdings nur von deskriptivem Wert, weil zwei potentiell konfundierende Effekte außer Betracht bleiben. Der eine ist ein Zusammensetzungseffekt: Wie Tabelle 1 zeigt, besuchen in Ländern mit zentralen Abschlussprüfungen im Durchschnitt mehr Schüler die Haupt- und Realschule und weniger Schüler das Gymnasium als in Ländern ohne zentrale Abschlussprüfungen. Da die Zuweisung der Schüler zu den Sekundarschulen hauptsächlich aufgrund ihrer Leistungen in der Primarschule erfolgt, ergibt sich der interessante Effekt, dass für jede Schulform die durchschnittliche Schulleistung in Ländern mit zentralen Abschlussprüfungen schon deshalb höher ist, weil sich bei ihnen in allen Schulformen relativ mehr gute Schüler befinden. Wir werden diesen Zusammensetzungseffekt in unseren Analysen statistisch kontrollieren (Kovariate „Alpha“).⁶ Wichtiger als die genannte unterschiedliche Zusammensetzung der Schüler in den einzelnen Schulformen ist ein weiterer Konfundierungseffekt, zu dem es kommt, wenn sich Länder mit zentralen Abschlussprüfungen in nicht beobachteten Aspekten von Ländern ohne zentrale Abschlussprüfungen unterscheiden. So könnten sich beispielsweise die Einstellungen der Eltern zu Schulbildung und Schulleis-

⁵ Unter dieser Gesamtzeit verstehen wir nicht nur die durchschnittliche Stundenzahl im jeweiligen Schuljahr, sondern die über alle Schuljahre bis zum Prüfungszeitpunkt hinweg akkumulierte Stundenzahl.

⁶ Die je nach Bundesland unterschiedliche Zusammensetzung der Schülerschaft in den deutschen Sekundarschulen wird als Ergebnis unterschiedlicher Schwellenwerte α für den Stand der Fähigkeiten interpretiert, nach denen die Aufteilung der Schüler auf die Schulformen erfolgt. Als Näherungsvariable für α werden wir $\Phi^{-1}(1-\alpha)$ benutzen, wobei Φ der Standardnormalverteilung und α dem Anteil der Gymnasiasten (plus einem Drittel aller Gesamtschüler) an der Gesamtzahl der Achtklässler entspricht.

tung in den Bundesländern systematisch unterscheiden. Eine zentrale Abschlussprüfung könnte somit auch einfach ein Indikator für die hohe Priorität sein, die Bildung in der dortigen Bevölkerung genießt. Insofern wären zentrale Abschlussprüfungen als im statistischen Sinne (zumindest langfristig) endogene Variable zu betrachten. Einfache Differenzen werden daher immer dem Einwand unterliegen, dass sie lediglich eine Korrelation und keinen Kausalzusammenhang abbilden. Zentrale Prüfungen könnten sogar kurzfristig endogen sein, wenn Eltern von einem Bundesland in ein anderes umziehen, um ihre Kinder in Schulen mit zentralen Abschlussprüfungen zu schicken. Dies dürfte jedoch kaum in so umfangreichem Maße geschehen, dass dies einen wesentlich Einfluss auf unsere Ergebnisse hätte.

Eine weitere Möglichkeit zur Einschätzung der Wirkung von zentralen Abschlussprüfungen sind *Vorher-Nachher-Vergleiche*. Unter dem Eindruck der intensiven öffentlichen Debatte, die auf die Bekanntgabe der Ergebnisse von TIMSS und vor allem PISA folgte, haben einige Länder zentrale Abschlussprüfungen eingeführt oder deren Einführung geplant. Pläne zur Abschaffung zentraler Abschlussprüfungen gibt es dagegen in keinem Bundesland. Da PISA als wiederholte Querschnittserhebung angelegt ist, könnten die PISA-Daten für eine Schätzung des Leistungseffekts dieser schulpolitischen Veränderungen herangezogen werden. Bei einem solchen Ansatz wird es allerdings noch immer schwer sein, die Kausaleffekte von zentralen Abschlussprüfungen zu identifizieren, weil es als Reaktion auf TIMSS und PISA auch noch zu anderen Veränderungen in der Bildungspolitik, den Einstellungen der Eltern zur Bildung usw. gekommen sein könnte. Einstweilen verfügen wir jedoch nur über Daten aus einer einzigen Querschnittserhebung, so dass wir noch nicht in der Lage sind, Vorher-Nachher-Vergleiche durchzuführen.

Wenden wir uns nun der von uns bevorzugten Identifikationsstrategie zu. Tabelle 1 zeigt nicht nur, welche Bundesländer zentrale Abschlussprüfungen in welcher Schulform haben, sondern auch, welche Fächer geprüft werden. Zu beachten ist, dass es zentrale Abschlussprüfungen an Haupt- und Realschulen (falls überhaupt) nur in Deutsch, Mathematik und einer Fremdsprache (meist Englisch) gibt. Die naturwissenschaftlichen Fächer sind bei zentralen Prüfungen nicht vertreten – mit zwei Ausnahmen: In Sachsen sind die Naturwissenschaften Pflichtfach für den Realschulabschluss, und in Bayern sind die Naturwissenschaften Wahlfach. Dort werden jedes Jahr rund 40 % aller Schüler, die einen Hauptschulabschluss anstreben, in Biologie, Chemie und Physik geprüft, d. h. in allen Fächern, die auch Gegenstand des TIMSS-Tests für die naturwissenschaftlichen Fächer sind. 20 bis 25 % der Schüler, die einen Realschulabschluss anstreben, legen eine schriftliche Prüfung allerdings nur in Physik ab (etwa ein Drittel der naturwissenschaftlichen Items in TIMSS stammt aus der Physik). Ist Mathematik ein Prüfungsfach bei zentralen Abschlussprüfungen, die Naturwissenschaften aber nicht, sollte der Leistungseffekt von zentralen Abschlussprüfungen im Mathematikunterricht größer ausfallen als in den naturwissenschaftlichen Fächern. Da in TIMSS Testergebnisse sowohl für Mathematik als auch für die naturwissenschaftlichen Fächer vorliegen, können wir die Differenzen-in-Differenzen zwischen Fächern schätzen. Man beachte, dass bei den vorliegenden Daten (Ergebnisse für Mathematik und Naturwissenschaften für ein und denselben Schüler) Zusammensetzungseffekte kein Problem mehr darstellen. Der größte Vorteil dieses Schätzers ist, dass jeder einzelne Schüler gewissermaßen seine eigene Kontrollgruppe ist, das heißt, dass wir in der Lage sind, Heterogenität – z. B. Befähigung allgemein – auf individueller Ebene zu kontrollieren.⁷ Die Interpretation

⁷ Eine verbleibende Quelle von Heterogenität sind natürlich Unterschiede zwischen den relativen (angeborenen) Fähigkeiten in Mathematik und in den Naturwissenschaften.

von Differenzen-in-Differenzen zwischen Fächern als kausaler Leistungseffekt zentraler Abschlussprüfungen impliziert zwei identifizierende Annahmen:

1. Es gibt keinerlei externe Effekte von der Mathematik auf die Naturwissenschaften.⁸
2. Es gibt zwischen Bundesländern mit und ohne zentrale Prüfungen keine Unterschiede in der relativen Präferenz für die Mathematik gegenüber den Naturwissenschaften.

Externe Effekte können positiv oder negativ sein. Sie sind positiv, wenn gute mathematische Kompetenzen eine Voraussetzung für gute Leistungen in den naturwissenschaftlichen Fächern oder – genauer gesagt – in den naturwissenschaftlichen TIMSS-Items wären. In diesem Falle würden Differenzen-in-Differenzen zwischen Fächern den Leistungseffekt von zentralen Abschlussprüfungen *unterschätzen*. Unserer Ansicht nach dürfte der externe Effekt einer guten mathematischen Bildung auf die Leistungen im naturwissenschaftlichen TIMSS-Test jedoch sehr gering sein. Um die Wahrscheinlichkeit eines solchen externen Effekts abzuschätzen, haben wir die (freigegebenen) TIMSS-Items (IEA TIMSS 1998) analysiert. Dabei handelt es sich um 87 Items, von denen lediglich vier recht elementare mathematische Kompetenzen wie etwa das Dividieren durch einen Bruch voraussetzen.

Ein negativer externer Effekt von der Mathematik auf die Naturwissenschaften wäre gegeben, wenn die Schüler die naturwissenschaftlichen Fächer vernachlässigen und sich stärker auf die Mathematik konzentrieren würden, weil diese nach einem externen Leistungsstandard geprüft wird. Träte ein solcher Verschiebungseffekt ein, so würde man mit dem Ansatz Differenzen-in-Differenzen zwischen Fächern den Leistungseffekt allgemeiner zentraler Abschlussprüfungen überschätzen. Unser Schätzer gäbe dann den *partiellen* Effekt einer Einführung zentraler Prüfungen wieder. Zur Einschätzung des Ausmaßes eines Verschiebungseffekts haben wir Selbstberichte zum Schülerverhalten untersucht. Zu diesem Zweck schätzten wir einige bivariate Probitschätzungen, bei denen die Wahrscheinlichkeit, dass zu Hause für Mathematik und die Naturwissenschaften gelernt wird, und die Wahrscheinlichkeit von Nachhilfeunterricht in Mathematik und in den naturwissenschaftlichen Fächern die abhängigen Variablen sind. Läge ein Verschiebungseffekt vor, müsste bei Schülern in Ländern mit zentralen Abschlussprüfungen die relative Wahrscheinlichkeit, dass sie zu Hause für Mathematik lernen, größer sein als die Wahrscheinlichkeit, dass sie für die naturwissenschaftlichen Fächer lernen, bzw. die Wahrscheinlichkeit, dass sie Nachhilfestunden in Mathematik erhalten, größer sein als die Wahrscheinlichkeit von Nachhilfestunden in den naturwissenschaftlichen Fächern. Dies ist jedoch, wie unsere Ergebnisse zeigen, nicht der Fall. Im Gegenteil, bei Schülern in Ländern mit zentralen Abschlussprüfungen ist die Wahrscheinlichkeit, dass sie Nachhilfestunden in Mathematik erhalten, sogar geringer als bei anderen, und zwar sowohl absolut, was überrascht, als auch relativ zu den naturwissenschaftlichen Fächern, was gegen einen Verschiebungseffekt spricht. Schüler in Bundesländern mit zentralen Abschlussprüfungen lernen zu Hause eher für die Mathematik als für die naturwissenschaftlichen Fächer, was auf eine gewisse Verschiebung hindeutet, doch ist die Differenz nicht signifikant (Ergebnisse nicht tabellarisch dokumentiert).

⁸ Externe Effekte von den naturwissenschaftlichen Fächern auf die Mathematik bleiben außer Betracht, da wir Effekte des spezifischen Wissens und der spezifischen Kompetenzen (Naturwissenschaften) auf die allgemeinen Kompetenzen (Mathematik) nicht erwarten.

Die zweite identifizierende Annahme ist, dass die relativen Präferenzen in Bundesländern mit und in Bundesländern ohne zentrale Abschlussprüfungen gleich sind. Die Tatsache, dass in den meisten Ländern mit zentralen Abschlussprüfungen zwar Mathematik, nicht aber die naturwissenschaftlichen Fächer zu den Prüfungsfächern gehören, deutet darauf hin, dass mathematische Kompetenzen allgemein höher bewertet werden als naturwissenschaftliche Kompetenzen. Sie lässt jedoch nicht den Schluss zu, dass die *relative* Präferenz für Mathematik in den Ländern mit zentralen Abschlussprüfungen größer ist als in den anderen Ländern. Mathematik ist in allen Ländern ein Kernfach, woraus sich auch erklärt, dass in der Primarschule rund ein Fünftel der Unterrichtszeit und in der Sekundarstufe I rund ein Siebtel der Unterrichtszeit auf Mathematik verwendet werden. Die Unterschiede zwischen den relativen Unterrichtszeiten in Ländern mit und ohne zentrale Abschlussprüfungen sind jedoch nur sehr gering. In Ländern mit zentralen Abschlussprüfungen beträgt der Anteil der Mathematikstunden an allen Stunden 14,3 % in der Hauptschule und 13,7 % in der Realschule. In Ländern ohne zentrale Abschlussprüfungen liegen die entsprechenden Zahlen bei 14,6 und 13,7 %, das heißt, hier ist der durchschnittliche Anteil der Mathematikstunden sogar noch etwas höher (Frenck 2001).

Wie aus den oben beschriebenen Zusatzdaten hervorgeht, erscheinen unsere identifizierenden Annahmen beim Ansatz der Differenzen-in-Differenzen zwischen Fächern als plausibel. Es bleibt jedoch noch ein potentielles Problem: Da in Sachsen Mathematik und die naturwissenschaftlichen Fächer generell Bestandteil der Abschlussprüfungen sind, schließen wir Sachsen aus diesem Teil unserer Analyse aus.⁹ In Bayern wiederum legen 25 bis 40 % aller Schüler zentrale Abschlussprüfungen in den Naturwissenschaften ab. Vieles spricht für die Annahme, dass Schüler, die in den naturwissenschaftlichen Fächern gut sind, diese auch als Fächer für ihre Abschlussprüfung wählen; dennoch können wir nicht mit Sicherheit sagen, wer am Ende in welchem Fach geprüft wird. Eine Möglichkeit, mit diesem Problem umzugehen, wäre der Ausschluss aller bayerischen Schüler aus unserer Analyse. Diesen Weg haben wir aus zwei Gründen nicht gewählt: Erstens hat Bayern von allen Ländern mit zentralen Abschlussprüfungen die umfangreichste Stichprobe und trägt rund 60 % zu allen Beobachtungen für zentrale Abschlussprüfungen bei. Zweitens lägen alle verbleibenden Länder mit zentralen Abschlussprüfungen in Ostdeutschland, nähme man Bayern aus den Daten heraus. Da vier von fünf Bundesländern in Ostdeutschland zentrale Abschlussprüfungen haben, wäre es bei einem Ausschluss Bayerns aus unserer Stichprobe nicht mehr möglich, den Effekt „zentrale Abschlussprüfungen“ eindeutig vom Effekt „ehemalige DDR“ zu unterscheiden. Dies aber ist wichtig, weil es an den Schulen in der ehemaligen DDR eine etwas andere Tradition des naturwissenschaftlichen Unterrichts zu geben scheint als im übrigen Deutschland.¹⁰

Im Zusammenhang mit dieser Problematik werden wir eine einfache Robustheitsprüfung vorlegen, indem wir diejenigen bayerischen Schüler aus der Stichprobe herausnehmen, bei denen die größte Wahrscheinlichkeit besteht, dass sie bei der Abschlussprüfung ein naturwissenschaftliches Fach wählen. Dabei handelt es sich um Hauptschüler, die der Aussage, in Biologie und/oder Chemie/Physik seien sie „gewöhnlich gut“, uneingeschränkt zustimmen, und um Realschüler, die der entsprechenden Aussage für Chemie/Physik uneinge-

⁹ Die Alternative wäre, alle sächsischen Schüler mit einer Dummy-Variable zu versehen, was nur zu geringfügigen Veränderungen unserer Ergebnisse führt; die Ergebnisse dieser Variante werden nicht tabellarisch dokumentiert.

¹⁰ Neuere Analysen von PISA zeigen, dass die naturwissenschaftlichen Leistungen der Schüler in Ostdeutschland bei spezifisch deutschen Items signifikant besser ausfallen als bei den internationalen Items. In Westdeutschland sind keine solchen Unterschiede zu beobachten (Baumert et al. 2002).

schränkt zustimmen. Müssten einige bayerische Schüler in unserer Stichprobe am Ende doch eine zentrale Abschlussprüfung in den naturwissenschaftlichen Fächern ablegen, hätte dies natürlich Auswirkungen auf unsere Schätzungen des Effekts zentraler Abschlussprüfungen mit dem Ansatz der Differenzen-in-Differenzen zwischen Fächern. Es ergäbe sich aber allein eine Verzerrung nach unten. Entsprechend sind wir in der Lage, eine Art Untergrenze für den Kausaleffekt von zentralen Abschlussprüfungen anzugeben.

4 Ergebnisse

Unsere Schätzergebnisse sind in Tabelle 3 aufgeführt. Spalte (1) enthält einen Schätzwert für die einfache Differenz in der Mathematikleistung von Schülern mit und Schülern ohne zentrale Abschlussprüfung in Mathematik. Die Differenz beträgt immerhin 0,472 Standardabweichungen, mehr als das Äquivalent eines ganzen Schuljahrs (0,372 Standardabweichungen). Man beachte, dass dabei für Herkunft- und Zusammensetzungseffekte kontrolliert wird. Bei Bishop (1997) beträgt der Effekt zentraler Abschlussprüfungen bei seinem Vergleich der kanadischen Provinzen mit und ohne zentrale Abschlussprüfungen wie bereits berichtet nur rund zwei Drittel des Äquivalents einer Jahrgangsstufe.

Alle unsere Hintergrundvariablen haben die erwarteten Effekte auf die Mathematikleistungen der Schüler. Da Realschüler aufgrund ihrer Leistungen in der Grundschule ihrer Schulform zugewiesen werden, liegen ihre Leistungen weit über denen der Hauptschüler. Die Zahl der zu Hause vorhandenen Bücher wird als Stellvertretervariable für den Bildungshintergrund der Eltern benutzt. Tatsächlich beträgt die Leistungsdifferenz zwischen Schülern aus Haushalten mit weniger als zehn Büchern und Schülern aus Haushalten mit über 200 Büchern mehr als ein Schuljahr. Immigrantenkinder schneiden etwas schlechter ab als andere, ebenso Kinder, die schon einmal eine Klasse wiederholt haben. Mädchen erbringen im Durchschnitt schlechtere Leistungen in Mathematik als Jungen,¹¹ westdeutsche Schüler bessere als ostdeutsche Schüler.

Die in Spalte (1) wiedergegebene Korrelation zwischen zentraler Abschlussprüfung und Schulleistung könnte auch auf nicht beobachtbare Variablen zurückzuführen sein, die mit zentralen Abschlussprüfungen korrelieren. Um diese Korrelation von einem Kausalzusammenhang zu unterscheiden, wenden wir uns nun der von uns bevorzugten Schätzfunktion zu, nämlich Differenzen-in-Differenzen zwischen Fächern. Wie bereits erwähnt, hat diese Schätzfunktion den großen Vorteil, dass jeder Schüler quasi seine eigene Kontrollgruppe ist, weil die zentrale Prüfung nur in Mathematik, nicht aber in den Naturwissenschaften abgelegt wird. Wir haben diese Schätzfunktion implementiert, indem wir für jeden Schüler die Differenz zwischen seinen Testwerten in Mathematik und Naturwissenschaften berechnet und diese Differenz einer Regressionsanalyse mit den üblichen erklärenden Variablen unterzogen haben. Die einzige Änderung in den Kovariaten liegt darin, dass wir statt der Zahl der Mathematikstunden das Verhältnis zwischen den kumulierten Mathematikstunden und einer Stellvertretervariablen für die Stunden in den naturwissenschaftlichen Fächern benutzen.¹² Die Schätzwerte sind in Spalte (2) von Tabelle 3 aufgeführt. Dabei fällt

¹¹ Zur detaillierten Analyse geschlechtstypischer Unterschiede in der Lösung von TIMSS-Aufgaben vgl. Mullis et al. (2000).

¹² Diese Stellvertretervariable besteht aus der Gesamtzahl der Unterrichtsstunden minus Mathematik- und Deutschstunden. Informationen nur zu den Unterrichtsstunden in den naturwissenschaftlichen Fächern lagen nicht vor.

Tabelle 3

Effekte zentraler Abschlussprüfungen auf Schulleistung in Mathematik

| | Einfache Differenzen, Zusammen- setzungseffekt berücksichtigt ¹ | Diff-in-Diff zwischen Fächern I ² | Diff-in-Diff zwischen Fächern II ^{2,3} (ver- änderte Population) |
|---------------------------------------|---|--|--|
| | (1) | (2) | (3) |
| Zentrale Abschlussprüfung | 0,472 (6,78)** | 0,123 (2,93)** | 0,174 (3,87)** |
| 8. Klasse | 0,372 (4,27)** | -0,005 (0,14) | -0,002 (0,06) |
| Alpha ⁴ | 2,131 (3,47)** | | |
| Realschule | 0,715 (12,88)** | 0,138 (3,27)** | 0,116 (2,70)** |
| 11–25 Bücher | 0,129 (2,25)* | -0,116 (2,10)* | -0,122 (2,11)* |
| 26–100 Bücher | 0,293 (4,82)** | -0,132 (2,25)* | -0,112 (1,85) |
| 101–200 Bücher | 0,384 (5,97)** | -0,132 (2,00)* | -0,129 (1,87) |
| 200+ Bücher | 0,446 (7,69)** | -0,160 (2,72)** | -0,137 (2,31)* |
| Immigrantenkind | -0,142 (2,47)* | 0,275 (4,92)** | 0,294 (5,17)** |
| Klasse wiederholt | -0,130 (3,32)** | -0,066 (1,68) | -0,065 (1,58) |
| Mädchen | -0,264 (7,00)** | 0,090 (2,58)* | 0,091 (2,57)* |
| Osten | -0,489 (4,63)** | -0,294 (3,68)** | -0,342 (4,16)** |
| Kumulierte Mathematikstunden | -0,410 (0,96) | | |
| Relative kumulierte Mathematikstunden | | 0,590 (0,60) | 0,926 (0,95) |
| Konstante | -1,404 (3,10)** | -0,176 (0,71) | -0,261 (1,08) |
| Beobachtungen | 3 197 | 2 995 | 2 727 |
| R-squared | 0,35 | 0,04 | 0,04 |

Robuste z-Werte (korrigiert um den Effekt der Stichprobenklumpung auf Klassenebene) in Klammern.

* Signifikant auf dem 5%-Niveau.

** Signifikant auf dem 1%-Niveau.

1 Abhängige Variable: Leistung in Mathematik.

2 Abhängige Variable: Differenz der Leistungen in Mathematik und Naturwissenschaften.

3 Unter Ausschluss der bayerischen Schüler, die angeben, sie seien gut in den Naturwissenschaften.

4 Für die Kodierung der Variablen „Alpha“ vgl. Fußnote 6, S. 243.

Quelle: Eigene Auswertungen von TIMSS-Deutschland.

als Erstes auf, dass der Koeffizient für zentrale Abschlussprüfungen positiv und signifikant bleibt. Er fällt jedoch von 0,472 Standardabweichungen bei der Schätzfunktion für die einfache Differenz auf 0,123 Standardabweichungen bzw. auf ein Drittel des Äquivalents eines Schuljahres. Die Schüler in Ländern mit zentralen Abschlussprüfungen weisen in Mathematik relativ bessere Leistungen auf als in den naturwissenschaftlichen Fächern. Dies ist konsistent mit der Hypothese, dass zentrale Prüfungen die Schulleistung verbessern.

Nimmt man die bayerischen Schüler mit starkem naturwissenschaftlichen Interesse (und einer entsprechend hohen Wahrscheinlichkeit, dass sie für die zentrale Abschlussprüfung ein naturwissenschaftliches Fach wählen) aus der Stichprobe heraus, erhöht sich der Schätzwert für den Effekt von zentralen Abschlussprüfungen (Spalte (3)). Dies stimmt mit unseren Erwartungen überein, da wir diejenigen Schüler aus der Stichprobe herausnehmen, bei denen der Einfluss der zentralen Abschlussprüfung auf die *relative* Leistung geringer sein dürfte als beim Durchschnitt.

Obwohl unser Interesse vorrangig der Variablen „Zentrale Abschlussprüfung“ gilt, sind auch die Schätzwerte für andere Parameter in den Spalten (2) und (3) bemerkenswert. Die Dummyvariable „Realschule (vs. Hauptschule)“ ist positiv und signifikant, also weisen Realschüler relativ zu Hauptschülern in Mathematik relativ bessere Leistungen auf als in den Naturwissenschaften. Die „Anzahl Bücher“-Koeffizienten sind negativ und nehmen mit der Zahl der im Elternhaus vorhandenen Bücher leicht an Effektstärke zu. Somit erbringen Schüler aus Haushalten mit höherem Bildungsstand in den Naturwissenschaften relativ bessere Leistungen als andere. Dies überrascht nicht, da vieles von dem, was TIMSS im Bereich Naturwissenschaften abfragt, auch als Allgemeinwissen bezeichnet werden kann. Kinder aus Immigrantenfamilien sind in Mathematik relativ besser als deutsche, was auf eine häufig unzureichende Beherrschung der deutschen Sprache zurückzuführen sein dürfte, die sie in Mathematik weniger einschränkt als in den Naturwissenschaften.

Ein letzter möglicher Einwand gegen eine Berechnung der Differenzen zwischen den standardisierten Rasch-Scores für Mathematik und Naturwissenschaften ist, dass sie nicht auf derselben Skala gemessen werden und daher keine direkt vergleichbaren Maße darstellen. Wir haben deshalb die Robustheit unserer Schätzwerte untersucht, indem wir die bundesweiten Rasch-Scores in exakte Quantile umgewandelt und deren Differenzen als abhängige Variable verwendet haben. Selbst der quantitative Charakter der Ergebnisse ist den oben vorgestellten Ergebnissen sehr ähnlich. Die Quantil-Differenz zwischen Mathematik und Naturwissenschaften ist bei einem durchschnittlichen Schüler in Ländern mit zentralen Abschlussprüfungen signifikant höher (um rund 3,3 Perzentile bzw. wiederum um rund ein Drittel des Äquivalents eines Schuljahrs der jeweiligen Jahrgangsstufe; nicht tabellarisch dokumentiert).

5 **Resümee**

In diesem Beitrag geht es um die Frage, ob zentrale Standards in Form zentraler Abschlussprüfungen positive Auswirkungen auf die Schulleistung haben. Theoretisch sind positive Effekte von zentralen Prüfungen zu erwarten. Empirisch jedoch ist der entsprechende Kausaleffekt nicht ohne weiteres zu identifizieren. Anders als frühere Studien nutzen wir die institutionelle Variation zwischen den deutschen Bundesländern, um eine Identifikationsstrategie für die Schätzung des Kausaleffekts von zentralen Abschlussprüfungen auf die Schulleistung zu entwickeln. Im deutschen Schulsystem haben nur einige Länder zentrale Abschlussprüfungen, meist in den Kernfächern Deutsch und Mathematik. Wir verwenden Daten aus der Third International Mathematics and Science Study (TIMSS) von Haupt- und Realschülern in der Sekundarstufe I (Klassenstufen 7 und 8), um den Kausaleffekt von zentralen Abschlussprüfungen auf die Schulleistung in Mathematik zu untersuchen. Beim einfachen Vergleich von Testergebnissen schneiden Schüler in Bundesländern mit zentralen Abschlussprüfungen deutlich besser ab als in Bundesländern ohne solche

Prüfungen. Der Unterschied beträgt, wenn wichtige Hintergrundvariablen berücksichtigt werden, etwa eine halbe Standardabweichung bzw. umgerechnet mehr als ein ganzes Schuljahr (in der entsprechenden Jahrgangsstufe). Ein solcher Unterschied besteht jedoch, in etwas geringerem Maße, auch für die Testergebnisse in naturwissenschaftlichen Fächern, die meist kein Bestandteil zentraler Abschlussprüfungen sind. Unsere Schätzfunktion „Differenzen-in-Differenzen zwischen Fächern“ interpretiert die in TIMSS beobachtbaren Unterschiede (zwischen Ländern mit und ohne Zentralprüfung) in der Differenz zwischen Leistungen in Mathematik und Naturwissenschaften als den Kausaleffekt der fachspezifischen Abschlussprüfungen auf die Schulleistung. Unter der Annahme, dass es keine externen Effekte zwischen den Leistungen in Mathematik und den naturwissenschaftlichen Fächern gibt, wird ein durchschnittlicher Kausaleffekt von zentralen Abschlussprüfungen auf die Mathematikleistung von rund 0,12 Standardabweichungen oder einem Drittel Schuljahr ermittelt.

Die Abweichung zwischen den zunächst ermittelten Leistungsunterschieden zwischen Ländern mit und Ländern ohne zentrale Abschlussprüfungen und dem, was wir als den Kausaleffekt solcher Abschlussprüfungen identifizieren, ist recht groß. Eine Interpretation der beobachteten Differenzen zwischen Ländern mit und ohne zentrale Abschlussprüfungen als Effekt allein der Prüfungsform erscheint entsprechend problematisch. Die Korrelation zwischen zentralen Abschlussprüfungen und Schulleistung dürfte zu einem erheblichen Teil (wenn auch nicht vollständig) auf unbeobachtete Unterschiede zwischen den Bundesländern zurückgeführt werden. Dennoch legen unsere empirischen Befunde den Schluss nahe, dass die Einführung zentraler Abschlussprüfungen zu einem signifikanten Anstieg der Schulleistung führen würde. Auch wenn damit die beobachtete Differenz zwischen Ländern mit und Ländern ohne zentrale Abschlussprüfungen wohl nicht vollständig aufgehoben werden kann, erscheint die Einführung zentraler Abschlussprüfungen sinnvoll, um damit den Schülern wie den Lehrkräften in ihren Ländern Anreize zur Orientierung an höheren Leistungsstandards zu bieten. Verglichen mit anderen ebenfalls diskutierten Maßnahmen zur Steigerung der Schulleistung, etwa einer Verringerung der Klassengröße, ist die Einführung zentraler Abschlussprüfungen erheblich kostengünstiger.

Literaturverzeichnis

- Baumert, J., C. Artelt, E. Klieme et al. (2002): *PISA 2000 – Die Länder der Bundesrepublik Deutschland im Vergleich*. Opladen, Leske + Budrich.
- Baumert, J. und R. Watermann (2000): Institutionelle und regionale Variabilität und die Sicherung gemeinsamer Standards in der gymnasialen Oberstufe. Kap. 8. In: J. Baumert, W. Bos und R. Lehmann (Hrsg.): *TIMSS – Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe*. Opladen, Leske + Budrich, 317–372.
- Bishop, J. H. (1997): The Effect of National Standards and Curriculum-Based Exams on Achievement. *American Economic Review*, 87, 260–264.
- Bishop, J. H. (1999): Are National Exit Examinations Important for Educational Efficiency? *Swedish Economic Policy Review*, 6, 349–401.
- Costrell, R. M. (1994): A Simple Model of Educational Standards. *American Economic Review*, 84, 956–971.
- Costrell, R. M. (1997): Can Educational Standards Raise Welfare? *Journal of Public Economics*, 65, 271–293.

- Effinger, M. R. und M. K. Polborn (1999): A Model of Vertically Differentiated Education. *Journal of Economics*, 69, 53–69.
- Hanushek, E. A. (1999): The Evidence on Class Size In: S. E. Mayer und P. Peterson (Hrsg.): *Earning and Learning: How Schools Matter*. Washington, D. C., Brookings, 131–168.
- Hoxby, C. M. (2000a): The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *Quarterly Journal of Economics*, 115, 1239–1285.
- Hoxby, C. M. (2000b): Does Competition among Public Schools Benefit Students and Taxpayers. *American Economic Review*, 90, 1209–1239.
- Frenck, I. (2001): *Studentafeln der Primar- und Sekundarstufe I im Ländervergleich – Eine empirische Studie am Beispiel der Fächer Deutsch und Mathematik*. Universität Essen (unveröffentlichte Magisterarbeit).
- IEA TIMSS (1998): *TIMSS Science Items: Released Set for Population 2* (seventh and eighth grades). Download unter: www.timss.bc.edu/timss1/timsspdf/bsitems.pdf [1. März 2002].
- Jürges, H. und K. Schneider (2003): International Differences in Student Achievement: An Economic Perspective. *German Economic Review* (im Erscheinen).
- Mullis, I. V. S., M. O. Martin, E. G. Fierros, A. L. Goldberg und S. E. Stemler (2000): *Gender Differences in Achievement*. International Study Center, Lynch School of Education, Boston College.
- Wößmann, L. und M. R. West (2002): *Class-Size Effects in School Systems Around the World: Evidence from Between-Grade Variation in TIMSS*. IZA Discussion Paper, No. 485. Bonn.