

Wirkungen zentraler Abschlussprüfungen im Mehrebenensystem Schule. Eine Zwischenbilanz internationaler und nationaler Befunde und Forschungsdesiderata

E. Dominique Klein · Mirko Krüger · Svenja M. Kühn · Isabell van Ackeren

Zusammenfassung: Zentrale Abschlussprüfungen sind zu einem zentralen Steuerungsinstrument am Ende der Sekundarstufen I und II in Deutschland geworden. Im Beitrag wird im Hinblick auf intendierte und nicht intendierte Wirkungen der aktuelle Forschungsstand in Deutschland mit internationalen Vergleichsperspektiven systematisiert. Dabei werden Befunde auf bildungspolitischer/administrativer, schulischer und unterrichtlicher Ebene sowie mit Blick auf Lernerträge im Sinne einer Zwischenbilanz zu einem relativ jungen Forschungsbereich berichtet. Es zeigen sich länder-, fach- und kursspezifische Unterschiede, die keinen generellen Zentralprüfungseffekt nahelegen. In diesem Zusammenhang sollte in künftigen Forschungen die Heterogenität der Prüfungsverfahren stärker berücksichtigt werden.

Schlüsselwörter: Zentrale Prüfungen · Steuerung · Schulsystem · Unterricht · Internationaler Vergleich

Effects of statewide exit exams in a multi-level school system. An interim report of international and national findings and research desiderata

Abstract: Statewide exit exams have become a central governance instrument at the end of secondary schooling in Germany. In the article, the current state of research in Germany is systematized from an international comparative perspective in relation to intended and unintended effects. Findings regarding state exam policies and effects at the school and instructional level are presented as an interim report on a comparably young field of research. The findings show dif-

Online publiziert: 12.02.2014

© Springer Fachmedien Wiesbaden 2014

Dr. E. D. Klein (✉) · M. Krüger · Dr. S. M. Kühn · Prof. Dr. I. van Ackeren
Fakultät für Bildungswissenschaften, Universität Duisburg-Essen,
Universitätsstraße 11, 45117 Essen, Deutschland
E-Mail: dominique.klein@uni-due.de

M. Krüger
E-Mail: mirko.krueger@uni-due.de

Dr. S. M. Kühn
E-Mail: svenja.kuehn@uni-due.de

Prof. Dr. I. van Ackeren
E-Mail: isabell.van-ackeren@uni-due.de

ferences by states, subjects and courses which suggest that there is no general effect of statewide exams. In this context, future research should account for the heterogeneity of exam procedures to a greater degree.

Keywords: Governance · Instruction · International comparison · School system · Statewide exit exams

1 Einleitung

Die beinahe flächendeckende Einführung zentraler Prüfungen (ZP) zum Abschluss der Sekundarstufen I und II gehört zu den wichtigsten schulstrukturellen Veränderungen (fast) aller deutschen Länder der letzten Jahre.¹ Mit Ausnahme von Rheinland-Pfalz vergeben derzeit alle den Hauptschul-, den Mittleren Schulabschluss sowie das Abitur auf der Grundlage (teil-)zentraler Prüfungen (Klein et al. 2009; Kühn 2013). Diese gelten als zentrales Steuerungsinstrument mit dem Ziel, durch die formale und inhaltliche Standardisierung von Verfahren Qualität und Vergleichbarkeit zu sichern (ebd.), mehr Chancengleichheit bei der Vergabe von Zertifikaten zu erreichen (Berkemeyer et al. 2012) und die Signalwirkung und Akzeptanz der Zertifikate für die Anschlussysteme (z. B. Ausbildungsbetriebe) zu erhöhen (Klemm 2004; Piopiunik et al. 2013).

ZP sind *ein* Element der in der letzten Dekade deutlich ausgeweiteten bzw. neu implementierten standardisierten Vergleichsmaßnahmen im deutschen Schulsystem (vgl. im Überblick Altrichter und Maag Merki 2010). Dabei werden unter dem einheitlichen Etikett ‚zentrale Prüfung‘ verschiedene Systeme subsumiert, die bereits innerhalb Deutschlands unterschiedliche Verfahren (z. B. unterschiedliche Wahlmöglichkeiten bei den Aufgabenstellungen, schulinterne/-externe Korrektur zentral gestellter Aufgaben) und Funktionen (z. B. Zertifizierungs- und damit einhergehende Selektionsfunktion, Monitoring, Rechenschaftslegung, Schul-/Unterrichtsentwicklung²) umfassen (vgl. z. B. Maag Merki 2010). Nachfolgend werden *Prüfungen*, die am Ende eines institutionellen Bildungswegs/-abschnittes von *allen* Schülerinnen und Schülern in bestimmten Fächern abgelegt werden müssen, fokussiert. Die Gestaltung von und die Aufsicht über das Prüfungsverfahren obliegen einer *zentralen, schulexternen Instanz* (z. B. Schulministerien). Diese gibt *einheitliche Aufgabenstellungen* (bzw. einen Aufgabenpool mit Auswahlaufgaben) für die schriftlichen Prüfungen (international auch: mündliche oder fachpraktische Prüfungen) *in einer fest definierten Region* (z. B. ein Bundesland) vor, die sich auf die länderspezifischen Curricula für die prüfungsrelevante Schulstufe beziehen.

Ungeachtet der hohen Relevanz von ZP für Bildungsbiografien und den damit verbundenen beruflichen und gesellschaftlichen Teilhabechancen gab es in Deutschland lange nur eine geringe wissenschaftliche Auseinandersetzung mit dem Thema (vgl. Klein et al. 2009). ZP wurden implementiert, ohne dass empirisch fundierte Ergebnisse zu deren Wirksamkeit – im Hinblick auf intendierte und nicht-intendierte Wirkungen³ – vorlagen. Mit der Einführung von ZP in der Mehrheit der Bundesländer wurden verschiedene Forschungsprojekte initiiert, die auf das *Zentralabitur* konzentriert sind (z. B. Klein et al. 2009; Kühn 2010; van Ackeren et al. 2012; Maag Merki 2012b); Studien zu den Sek. I-Abschlüssen finden sich kaum.

Im Forschungsfeld „Zentrale Prüfungen“ werden mit unterschiedlichen theoretischen und methodischen Zugängen Fragestellungen der Educational Governance, der Schuleffektivität/Schulentwicklung, der Unterrichtsqualität sowie pädagogisch-psychologischer und bildungsökonomischer Forschung in den Blick genommen. Entlang der Mehrebenenstruktur des Bildungssystems (politische/administrative *Makrosteuerung*, schulische *Mesoebene*, unterrichtliche *Mikroebene*) werden im Folgenden Befunde nationaler und internationaler Studien systematisiert, um zu reflektieren, wo sich Ergebnisse ggf. zu einer stabileren Befundlage verdichten bzw. inwieweit sich differenzielle Wirkungen entfalten. Anschließend werden Befunde zu den Wirkungen von ZP auf Lernresultate präsentiert.

Die Literaturrecherche erfolgte über die Datenbanken ERIC und FIS Bildung mit folgenden Deskriptoren: Zentrale Prüfungen, Zentrale Abschlussprüfungen, Zentralabitur, Zentralmatura; Statewide exit exam/ination/s, High school exit exam/ination/s, Public (exit) exam/ination/s, National (exit) exam/ination/s, Central exit exam/ination/s sowie Eigennamen der Prüfungsverfahren, (Prüfungs-)Aufgaben, Abituraufgaben, task und assessment. Literatur wurde berücksichtigt, wenn sie sich auf zentrale Abschlussprüfungen nach der festgelegten Definition bezog und Wirkungen auf die Schul- und Unterrichtsebene sowie auf Schülerleistungen oder die mit den Prüfungen verknüpften Aufgaben und Funktionen/Intentionen *empirisch* untersucht wurden. Die in Tab. 1 dargestellten Ebenen sind nicht durchgängig stringent zu trennen; so können Setzungen auf der Systemebene Wirkungen auf nachfolgenden Ebenen entfalten (z. B. beeinflussen Korrekturvorgaben auf der Systemebene das Korrekturverhalten von Lehrkräften auf der Meso- bzw. Mikroebene).

2 Politisch-administrative Makroebene: Steuerungsintentionen und Steuerungsmechanismen

2.1 Steuerungsintentionen: vielgestaltig und ausdifferenziert

In der Literatur wird – u. a. auf der Grundlage von Dokumentenanalysen, z. B. zu bildungspolitischen Zielvorstellungen – berichtet, dass mit den ZP differenzierte Ziele verfolgt werden, welche im Kontext der traditionellen Funktion der *Zertifizierung* und darauf bezogenen *Herstellung von Vergleichbarkeit* auch eine erhöhte *staatliche Kontrolle über (neue) Inhalte*, über den *Erwerb als relevant erachteter Kompetenzen* und über die *Sicherung und Entwicklung schulischer Qualität* umfassen (vgl. Klein et al. 2009). So wird z. B. erwartet, dass Inhaltsbereiche aufgrund der auch für die Lehrkräfte unbekannt 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 (ebd.).

Im Rekurs auf Educational Governance-Konzepte und die damit verbundene Perspektive der wechselseitigen Bedingtheit von Handlungen innerhalb von Akteurskonstellationen (vgl. Brüsmeister 2007) kann man davon ausgehen, dass die auf die spezifischen Ziele abgestimmte Ausgestaltung der Prüfung sowie die Form der entsprechenden Imple-

Tab. 1: Struktur des Literatur- und Forschungsberichts

Analyseebene	Dimensionen	Indikatoren	Theoretische/disziplinäre Bezüge	Methodische Zugänge
Politisch-administrative Makroebene	<i>Steuerungsinentionen</i> (im Kontext landesspezifischer, historischer und kulturell geprägter Rahmenbedingungen)	Zertifizierung, Vergleichbarkeit, Sicherung bzw. Entwicklung von Standards (prozess- und wirkungsbezogen), Rechenschaftslegung (Schüler- bzw. Schulebene)	Bildungspolitische und -soziologische Analysen, Educational Governance-Konzepte zur Analyse von Akteurskonstellationen	Dokumentenanalysen, Ereigniszeitanalysen, Experteninterviews, Internationaler Vergleich
	Steuerungsmechanismen	<i>Prüfungsaufgaben</i> (u. a. Anspruchsniveau, Lehrplanvalidität, Innovationsgehalt, Testfairness) <i>Korrekturvorgaben</i> (Freiheitsgrade, Bezugsnormorientierung)	Fachdidaktische Analyse der Aufgabenqualität	Rating von Aufgabenmerkmalen, Qualitative Analysen von Korrekturverhalten
Schulische Mesoebene	Organisationale Strukturen im Umgang mit zentralen Prüfungen	Kooperation, Belastungserleben, thematische Fokussierung im Kontext kollegialer Selbstwirksamkeit, Bezugsnormorientierung, Ressourcenaallokation	Schuleffektivitäts- und Schulentwicklungsforschung, Educational Governance-Konzepte zur Analyse von Akteurskonstellationen	Fokusgruppeninterviews, quantitative Befragungen, qualitative Fallstudien (Einzelinterviews, Fokusgruppeninterviews, Beobachtungen), Internationaler Vergleich

Tab. 1: (Fortsetzung)

Analyseebene	Dimensionen	Indikatoren	Theoretische/disziplinäre Bezüge	Methodische Zugänge
Unterrichtliche Mikroebene	Lehrerperspektive	<i>Unterrichtsgestaltung</i> (Bezugsnormen, Prüfungsvorbereitung, Themenvarianz, individuelle Förderung) <i>Unterrichtsentwicklung</i> (Datennutzung, Verknüpfung interne/externe Evaluation) <i>Wahrnehmung und Akzeptanz der Prüfung</i> (Belastungserleben, Leistungsdruck, wahrgenommenes Leistungsniveau) <i>Motivation, Einstellungen, Wahrnehmungen</i> (Prüfungsangst, Interesse, Anstrengung, Miss-erfolgssänglichkeit, Selbstwirksamkeit, auch im Zusammenhang mit Leistung)	Pädagogische/psychologische Unterrichtsforschung, Educational Governance-Konzepte zur Analyse von Akteurskonstellationen	Quantitative Befragungen, auch im Längsschnitt, Unterrichtsbeobachtung, Internationaler Vergleich
Wirkungen und Erträge	Schülerperspektive	<i>Abschlüsse und Kompetenzen</i> (Leistungsstand, -entwicklung, Bildungsläufe, Benachteiligung) <i>Erträge von Abschlüssen</i> (Zugang zu Universität und Arbeitsmarkt, Studierenerfolg)	Bildungsökonomische Analysen	(Re-)Analyse von kompetenzbezogenen Daten (Re-)Analyse von statistischen Kennzahlen zu formalen Bildungserfolgen

mentations- und Kommunikationsstrukturen im jeweiligen System schulischer Qualitätsentwicklung maßgeblich beeinflussen dürften, welche Wirkungen die Prüfungen im schulischen Handlungskontext durch die Wahrnehmung und Deutung der schulischen Akteure entfalten.

Besonders in den USA, wo zentrale Abschlussprüfungen in vielen Staaten eine jüngere Entwicklung darstellen⁴, werden die mit zentralen *High School Exit Exams* verknüpften Steuerungsintentionen vielfach thematisiert, wobei neben *Student Accountability* hier auch auf den Aspekt der *Rechenschaftslegung* von Schulen im Kontext des *No Child Left Behind*-Bildungsgesetzes rekurriert wird (z. B. Zabala et al. 2008); beispielsweise werden die Prüfungsergebnisse in einigen Bundesstaaten genutzt, um die Qualität der Schulen, etwa im *Adequate Yearly Progress*, abzubilden⁵. Mittels einer Ereigniszeitanalyse mit demographischen Daten zeigen zudem Warren und Kulick (2007), dass insbesondere US-Staaten mit hoher Arbeitslosenquote und Herkunftsvielfalt eher dazu neigen, ZP einzuführen. Eine von den Autoren diskutierte Erklärung ist, dass die Prüfungen *sozialen Benachteiligungen bei der Vergabe von Abschlüssen entgegenwirken* sollen (ebd., S. 226 f.). Inwiefern diese oder andere Erklärungsansätze (z. B. Sicherung von Mindeststandards und Demonstration von Strenge) die tatsächlichen Steuerungsintentionen der verschiedenen Bundesstaaten abbilden, lässt sich mit dem gewählten Forschungsansatz jedoch nicht klären.

In einer explorativen Vergleichsstudie beschreibt Klein (2013) die Funktionen von ZP am Ende der Sek. II in den exemplarisch gewählten Ländern Finnland, Irland und den Niederlanden mithilfe von Dokumentenanalysen und teilstandardisierten Experteninterviews mit Repräsentanten der Prüfungsbehörden. Die Funktionen überschneiden sich zwischen den Ländern allein hinsichtlich des Zertifizierungs- und Vergleichbarkeitsanspruchs. Akzentuierungen weiterer Funktionen, welche Prüfungen im Rahmen der Neuen Steuerung zugeschrieben werden – insbesondere Rechenschaft und Wettbewerb, aber auch Aspekte der Schulentwicklung –, unterscheiden sich dagegen erheblich, was auch in der Gestaltung der Prüfung und der Kommunikation gegenüber den Schulen zum Ausdruck kommt. Die Heterogenität der konkreten Nutzbarmachung von Prüfungsstrukturen und -resultaten in den Schulen erweist sich dabei auf der Grundlage der Daten einer in diesen Ländern ebenfalls durchgeführten Fragebogenstudie tatsächlich als stark beeinflusst durch landesspezifische, kulturelle und historisch verwurzelte Rahmenbedingungen in der Steuerung der Bildungssysteme und der Ausgestaltung der Qualitätsentwicklungssysteme (vgl. hierzu auch die Abschn. 3 und 4). Die Berücksichtigung der Systemebene und des Steuerungsanspruchs erweist sich demnach als ertragreich, um differenzielle Wirkungen der ZP im Kontext traditioneller und neuer Steuerungsansprüche zu verstehen.

2.2 Steuerungsmechanismen: Prüfungsaufgaben und Korrekturvorgaben im Fokus

Mit Blick auf die Realisierung der Steuerungsintentionen kommt den Aufgaben sowie deren Korrektur bzw. Korrekturvorgaben eine bedeutsame Rolle zu. Dabei stellt sich z. B. die Frage, inwiefern diese den zugrundeliegenden, bildungsadministrativ vorgegebenen Standards entsprechen (vgl. Kühn 2010 zu den Einheitlichen Prüfungsanforderungen EPA; Kühn und Drüke-Noe 2013, zu den Bildungsstandards für den Mittleren Abschluss). Darüber hinaus wird zum Aspekt der Fairness auch die Frage gestellt, ob

durch die Gestaltung von Prüfungsaufgaben bestimmte Schülergruppen bevorzugt bzw. benachteiligt werden – etwa mit Blick auf das Geschlecht⁶.

Aus steuerungstheoretischer Perspektive wird in der Literatur (z. B. Kühn 2010; Maier et al. 2011) zudem die Annahme formuliert, zentral gestellte Aufgaben seien für eine schnelle und flächendeckende Implementierung aufgabenbezogener Neuerungen auf Unterrichtsebene geeignet (z. B. durch kognitiv anspruchsvolle Aufgaben, innovative Inhalte oder Formate), da sich Lehrkräfte bei der Nutzung und Gestaltung von Aufgaben am ‚Modell‘ der Prüfungsaufgaben orientierten und letztere somit zur Weiterentwicklung der Unterrichts- und Aufgabekultur beitragen könnten.

Es liegen primär für die mathematisch-naturwissenschaftlichen Fächer in verschiedenen deutschen Ländern Beurteilungen von Aufgabenmerkmalen aus ZP am Ende der Sek. I/II vor. Einzelne vorliegende Studien weisen auf kognitiv weniger anspruchsvolle⁷, aber lehrplanvalide (z. B. Kühn 2010, 2011a, b; Kühn und Drüke-Noe 2013) und unter fachdidaktischen Gesichtspunkten in Ansätzen innovative Aufgabenformate hin (vgl. Schoppmeier et al. 2012; Otto und Kühn, im Erscheinen). Einschätzungen von Lehrkräften zum kognitiven Anspruchsniveau der Aufgaben verweisen auf fach- und länderspezifisch differenzielle Wahrnehmungen (LISUM, z. B. LISUM 2011; Appius und Holmeier 2012); dies bestätigen z. B. auch die Befunde eines standardisierten Aufgabenratings für das Fach Biologie (Florian et al. 2011).

Sowohl in Bezug auf die differenziell ausgestalteten Prüfungsverfahren in Deutschland (s. o.) als auch bezüglich der Gestaltung der (hier: naturwissenschaftlichen) Abiturprüfungsaufgaben hält Kühn (2010) fest, dass sich die Steuerungsintention der KMK und die Ausführung in den betrachteten Bundesländern offensichtlich durch den aktiven Gestaltungsanteil der beteiligten Akteure bei der Aufgabenentwicklung erheblich unterscheiden. Es sei anzunehmen, dass informelle Regelungen im Sinne einer bestimmten, tradierten Aufgabekultur und länderspezifischen Prüfungstradition steuernd wirkten. Abzuwarten bleibt die Qualität der Aufgaben bzw. die damit verbundene Steuerungswirkung im Kontext des geplanten länderübergreifenden Aufgabenpools, dessen Ziel es ist, die zentralen Prüfungen im Hinblick auf Aufgabenqualität, Aufgabenschwierigkeiten und Bewertungsstandards bundesweit vergleichbarer zu gestalten (KMK-Beschluss vom 20. und 21. Juni 2013 in Wittenberg). Damit wird sich der Standardisierungsgrad der Prüfungen erhöhen, obgleich die Auswahl bestimmter Aufgaben weiterhin den Ländern überlassen wird.

Aufgabenanalysen liegen aus internationaler Perspektive für mehrere Unterrichtsfächer in verschiedenen Staaten vor; sie zeichnen ein ähnliches Bild. Einerseits deuten z. B. Befunde aus England (Ofqual 2012), Finnland (Tikkanen 2010; Tikkanen und Aksela 2012), Japan (Britton und Raizen 1996) oder aus dem australischen New South Wales (Baumgart und Halse 1999) auf kognitiv anspruchsvolle Aufgaben⁷ im Rahmen von ZP hin. Andererseits wird auch eine kognitiv weniger anspruchsvolle Aufgabenpraxis in Verbindung mit ZP aus verschiedenen Bundesstaaten der USA (Achieve, Inc. 2004) sowie mehreren Ländern in Europa (Kühn 2012), Asien (Baumgart und Halse 1999) und Afrika (Valverde 2005) berichtet. Hier dürften wiederum kulturell unterschiedliche Prüfungstraditionen und damit verbundene differenzielle Steuerungsintentionen einen Einfluss haben. Letztere bewegen sich in einem Spannungsfeld von Vergleichbarkeitssicherung und Qualitätsentwicklung einerseits (mit entsprechender Signalwirkung gegenüber nach-

folgenden Bildungs- und Ausbildungssystemen und dem Arbeitsmarkt) sowie andererseits der Sorge dafür, dass die Prüfungsanforderungen die Anteile von Schüler/inne/n ohne bzw. mit gering qualifizierenden Schulabschlüssen nicht erhöhen.

Daneben spielen auch die Korrekturvorgaben eine zentrale Rolle für die Durchsetzung staatlicher Steuerungsintentionen – insbesondere in Deutschland, wo die Prüfungen schulintern korrigiert und bewertet werden, was die Bezeichnung „zentrale Abschlussprüfung“ verschleiert. Die Korrekturvorgaben sollen gewährleisten, dass die Bewertung vor allem an einem kriterialen Maßstab orientiert ist und dadurch Vergleichbarkeit innerhalb einer über Kurse und Schulen hinausgehenden Kohorte sowie möglichst auch zwischen Jahrgängen sicherstellt. Dabei ist die Befundlage zur Ausgestaltung von Korrektur- und Bewertungsrichtlinien eher dünn. Eine Studie von Zabka und Stark (2010) zur Gestaltung der im Zentralabitur Deutsch eingesetzten länderspezifischen Bewertungsvorgaben zeigt, dass sich diese in Darstellungsform, Umfang und Inhalt sowie hinsichtlich der Verbindlichkeit teils deutlich unterscheiden und z. T. nur aus unkommentierten Listen von Leistungsbeschreibungen mit disparaten Gültigkeitsansprüchen bestehen. Fraglich bleibt, ob es ggf. Unterschiede zwischen Fächergruppen mit eher klaren bzw. weniger klaren Fachinhalten und Fachgrenzen bzw. mehr oder weniger Konsens über Inhalte, ihre Abfolge (Sequenzialität) und Ziele gibt. In der Konsequenz könnten Aufgabenentwickler und Lehrkräfte unterschiedlicher Fächer verschieden ausgeprägte curriculare Autonomie für ihr Handeln beanspruchen (Stodolsky und Grossman 1995). Es bleibt zudem zu fragen, welche Freiheiten Prüfer/innen haben und inwiefern die Benotung tatsächlich vergleichbar ist. Neumann et al. (2011) zeigen für Hamburg und Baden-Württemberg, dass in die Bewertung im Zentralabitur Mathematik und Englisch schulbezogene soziale Bezugsmaßstäbe nur in geringem Maße mit einfließen; Holmeier (2013) berichtet diesbezüglich für Bremen und Hessen sowohl fach- als auch kursspezifische Unterschiede. Vereinzelt qualitative Analysen des Korrekturverhaltens weisen zusätzlich darauf hin, dass Korrektoren für die Notenfindung auch in stärker standardisierten Korrekturverfahren Vergleiche zwischen den Prüflingen, und somit einen sozialen Bezugsmaßstab, heranziehen (vgl. z. B. Crisp 2010, England).

In einigen Studien wird auch der Frage nachgegangen, inwiefern ZP dazu beitragen, *schulintern* durchgeführte formative oder summative Leistungsbewertungen stärker an einer kriterialen Norm zu orientieren. Die Befunde weisen auf keine grundsätzlich stärkere Orientierung an der kriterialen Bezugsnorm in dezentralen Prüfungsanteilen hin (vgl. z. B. van den Bergh et al. 2003, Niederlande; Neumann et al. 2009, Hamburg und Baden-Württemberg; Luyten und Dolkar 2010, Bhutan). Dies gilt auch für die vorangehende interne Leistungsbewertung (vgl. Holmeier 2012, 2013, Bremen und Hessen).

3 Schulische Mesoebene: Organisationale Strukturen im Umgang mit zentralen Prüfungen

Wie die mit den Prüfungen verknüpften Aufgaben und Anforderungen in Schulen wahrgenommen und umgesetzt werden, dürfte auch durch organisationale Verarbeitungsstrategien, Schulkultur und kollektive Normen und Werte einer Schule (z. B. die Einstellung gegenüber den Instrumenten Neuer Steuerung, datengestützte Schulentwicklungsstrate-

gien etc.) sowie ihre kontextuellen Rahmenbedingungen (z. B. Zusammensetzung der Schülerschaft, Vernetzung mit anderen Schulen) beeinflusst werden (vgl. z. B. Spillane 2012). Es ist insofern zu erwarten, dass sich Varianzen in den Wirkungen der Prüfungen zum Teil auch durch Unterschiede zwischen Schulen aufklären und sich unterschiedliche organisationale Verarbeitungsstrategien beobachten lassen, dass in umgekehrter Perspektive die zentralisierten Prüfungen aber auch dazu beitragen, Impulse für z. B. verstärkte Kooperationsbemühungen zu setzen.

In einer von Maag Merki (2012b) und Kolleginnen durchgeführten Studie zur Einführung zentraler Abiturprüfungen in Bremen und Hessen werden auf der Grundlage von Fokusgruppeninterviews (die allerdings nur mit den Bremer Lehrkräften durchgeführt wurden) keine Veränderungen der kollegialen Kooperation berichtet (Maag Merki 2008). Auf der Basis der weiteren quantitativen Befragungen in beiden Ländern zeichnet sich zudem ab, dass die *Intensität und Ausrichtung der Kooperation* über die drei Jahre des Längsschnitts hinweg relativ konstant geblieben ist. Die Intensität der Kooperation im Rahmen des Zentralabiturs korreliert kaum z. B. mit Unsicherheiten oder einem niedrigeren Belastungserleben (Appius 2012), zudem hat das Bedürfnis zu kooperieren – ebenso wie das grundsätzliche Interesse an der Auseinandersetzung mit dem Zentralabitur – insbesondere der hessischen Lehrkräfte im Laufe des Beobachtungszeitraums im Schnitt eher abgenommen (Oerke 2012b), so dass vermutet werden kann, dass die Kooperation vor allem mit Blick auf die Reduktion von Unsicherheiten in der Anfangsphase der Implementation bedeutsam gewesen ist.

In der Studie zeigt sich zudem, dass die Strategie von Lehrkräften, Unterrichtsinhalte stärker *auf prüfungsrelevante Themen zu fokussieren*, zwar nicht durch die Schulzugehörigkeit und schulischen Kooperationsstrukturen beeinflusst wird (vgl. Jäger et al. 2012), gleichwohl wird aber deutlich, dass Lehrkräfte, welche eine höhere kollektive Selbstwirksamkeit berichten und demnach überzeugt sind, dass das Kollegium z. B. mit schwierigeren Situationen umgehen kann, eher bereit sind, Themen zu variieren, als Lehrkräfte, die eine niedrigere kollektive Selbstwirksamkeit berichten (Jäger 2012). Die *Nutzung der kriterialen Bezugsnorm* im vorgelagerten Unterricht ist in der Studie im Zentralabitur zwar nicht grundsätzlich ausgeprägter als im dezentralen Abitur (vgl. Abschn. 2.2), ein höheres Maß der Kooperation hinsichtlich der Leistungsbewertung scheint aber die Nutzung der kriterialen Bezugsnorm zu begünstigen (Holmeier 2013). Zudem können die Schulen in Bremen hinsichtlich der im Unterricht angewandten Lernstrategien (u. a. verständnisorientiertes, unterstützendes Lernen) im Umgang mit Grund- und Leistungskursen mit und ohne ZP in verschiedenen Clustern gebündelt werden. So werden in einer Gruppe von Schulen in allen Grund- und Leistungskursen gleiche Lernstrategien angewandt, in einer weiteren Gruppe gibt es kursformspezifische Unterschiede, in einer dritten Gruppe Schulen korrelieren die eingesetzten Strategien vor allem mit der Prüfungsform (zentral/dezentral) (vgl. Maag Merki et al. 2008).

In den USA wird vor allem im Rahmen qualitativer Fallstudien untersucht, auf welche Weise Districts und Schulen zentrale *High School Exit Exams* nutzen, um im Rahmen von *school accountability* Schule weiterzuentwickeln. In einer Studie in zwei New Yorker Schulen erweist sich dabei der Umgang mit Änderungen in den Prüfungsanforderungen (das Bestehen der zentralen *Regents Exams* wurde von einer *möglichen* Art des Abschlusses zur grundlegenden *Voraussetzung* für einen Abschluss) als abhängig von bereits exist-

tierenden (oder eben nicht vorhandenen) Leitungsstrukturen, Formen der Kooperation, schulischen Ressourcen und Kompetenzen im Umgang mit den Prüfungen sowie Normen und Werten innerhalb der Schule (DeBray et al. 2001); die benannten Strukturen und Prozesse variieren zudem zwischen Fachbereichen innerhalb der Schule (DeBray 2005).

In einer fallbasierten Interviewstudie in fünf School Districts in New York zeigte sich, dass eine Verschärfung des Prüfungsreglements in allen beobachteten Districts eine Reallokation unterrichtlicher Ressourcen zuungunsten musisch-künstlerischer Fächer oder Freiarbeitszeiten zur Folge hatte (Sipple et al. 2004). Auf ähnliche Verschiebungen deutet eine ebenfalls qualitative Studie aus England hin (Perryman et al. 2011). Für beide Studien ist zu berücksichtigen, dass mit der Erfolgsquote in den Prüfungen auch Konsequenzen für die Schulen, etwa mit Blick auf die finanzielle oder personelle Ausstattung, verknüpft waren, welche den Druck auf die Schulen, Ressourcen umzuschichten, vergrößert haben dürften.

In der bereits erwähnten Vergleichsstudie zu Prüfungen in Finnland, Irland und den Niederlanden (vgl. Abschn. 2.1) geht Klein (2013) der Frage nach, auf welche Weise ZP sowie die durch sie zurückgemeldeten Leistungsdaten auf Schul- und Fachebene zur Schulentwicklung genutzt werden und stellt im Rahmen einer Fragebogenstudie „traditionellere“ Formen (Bewältigung der Prüfungsanforderungen, Leistungsbewertung, Prüfungsvorbereitung) und „neuere“ Formen (Schulentwicklung, interne Evaluation, Accountability) des Umgangs mit den Prüfungen gegenüber. Es zeichnet sich ab, dass die schulische Nutzung der Prüfungen auch von den vom Staat beschriebenen Intentionen beeinflusst zu sein scheint. So werden z. B. die in den Niederlanden im Rahmen eines umfassenden Qualitätsentwicklungssystems eingesetzten Abschlussprüfungen in den Schulen in umfassender, systematischer Weise für traditionelle wie neuere Anwendungsformen genutzt, während in den anderen beiden Ländern mit einer weniger auf Evaluation und Accountability ausgerichteten Steuerungsstruktur die Nutzung sehr gering ist (Finnland) bzw. stark auf traditionellere Nutzungsformen fokussiert (Irland). Ein grundsätzlicher Effekt von ZP auf die Ausgestaltung datengestützter Schulentwicklung kann hier nicht bestätigt werden.

4 Unterrichtliche Mikroebene I: Lehrerperspektive

Mit Blick auf die Unterrichtsgestaltung sind die Erwartungen an die positiven Wirkungen sowie Befürchtungen bezüglich potenzieller nicht-intendierter Effekte der ZP besonders hoch. Relevant sind dabei einerseits die tatsächliche Unterrichtsgestaltung in Systemen mit ZP, andererseits die Nutzung der *Befunde* der ZP zur Gestaltung und Weiterentwicklung des Unterrichts.

4.1 Unterrichtsgestaltung

In der deutschsprachigen Literatur wird die Unterrichtsebene besonders fokussiert. Die vorliegenden Befunde spiegeln dabei eine eher heterogene Wirkung von ZP auf die Oberflächen- und Tiefenstruktur des Unterrichts wider. So stellen van Ackeren et al. (2012) in einer Vergleichsstudie zwischen zwei Ländern mit traditionell zentralen (Baden-

Württemberg) bzw. dezentralen Abiturprüfungen (Rheinland-Pfalz; als Referenz wurden NRW-Lehrkräfte kurz nach dem Wechsel von dezentralen zu zentralen Prüfungen befragt) nur wenige Unterschiede fest. Dies betrifft z. B. eine systematisch geringere individuelle und soziale *Bezugsnormorientierung* bei der Benotung im zentralen gegenüber dem dezentralen Prüfungssystem, sowohl in Grund- als auch Leistungskursen. Zudem findet sich im zentralen Prüfungssystem eine hohe Intensität der *Prüfungsvorbereitung*, ohne dass sich allerdings Anzeichen für eine deutliche, länger andauernde thematische Engführung im Unterricht finden würden. In diesem Untersuchungskontext zeigen sich durchaus gewünschte Wirkungen von ZP bzw. eine nicht-intendierte Wirkung bleibt in diesem Kontext aus.

In der Längsschnittstudie von Maag Merki und Kolleginnen berichten Lehrkräfte in den (zunächst noch) dezentral geprüften Leistungskursen in Bremen im Schnitt eine breitere *Themenvariation* als Lehrkräfte in den bereits zentral geprüften Kursen (Maag Merki und Holmeier 2008). Nach der Einführung der ZP lässt sich für den Analysezeitraum von 2007 bis 2009 sowie eine Nacherhebung in 2011 feststellen, dass die Themenvarianz im Zentralabitur eher geringer ausfällt (Jäger et al. 2012; Oerke et al. 2013). Analog dazu weist auch eine retrospektive Befragung von Lehrkräften in NRW darauf hin, dass sich die Themenvarianz nach der Einführung des Zentralabiturs eher rückläufig entwickelt hat (Eickelmann et al. 2011).

In einer Fragebogenstudie von Kühn und Racherbäumer (2013) mit Deutsch- und Mathematiklehrkräften verschiedener Schulformen in NRW geben Gymnasial- und Gesamtschullehrkräfte in höherem Maße als Realschullehrkräfte an, hinsichtlich der *Möglichkeiten zur individuellen Förderung* durch die Prüfungen eingeschränkt zu sein (Racherbäumer und Kühn 2013). Die Befunde stärken die Annahme, dass das Zentralabitur den Unterricht stärker beeinflusst als ZP zum Mittleren Schulabschluss. Insbesondere im Deutsch-Leistungskurs wird verstärkt ein Widerspruch zwischen Standardisierung und individueller Förderung gesehen, was möglicherweise mit den Lektürevorgaben zusammenhängen könnte, welche die bisherige Themenvarianz offenbar erheblich einschränken (Kühn und Racherbäumer 2013). Gleichwohl berichtet die große Mehrheit retrospektiv und schulformübergreifend, ihren Unterricht nicht verändert zu haben (Kühn und Racherbäumer 2013; Racherbäumer und Kühn 2013). Offensichtlich sind die Einstellungen der Lehrpersonen zum Stellenwert individueller Förderung von vorherrschenden Handlungsroutinen überlagert.

Die *Tiefenstruktur* des Unterrichts im Kontext der ZP ist zudem in der Längsschnittstudie von Maag Merki (2012b) und Kolleginnen hinsichtlich der Lernstrategien aus Schülersperspektive analysiert worden. Hier legen die Angaben zwar insgesamt eine eher erhöhte *Nutzung von Elaborationsstrategien* im Unterricht nahe, was aber länderspezifisch sowie fachabhängig variiert (Maag Merki und Holmeier 2008; Maag Merki et al. 2010; Maag Merki 2011). Eine Differenzierung nach Grund-/Leistungskursen, wie sie zuvor Baumert und Watermann (2000) auf Basis von Reanalysen der TIMSS/III-Daten von 1999 für Deutschland konstatiert hatten⁸, wird von den Autorinnen nicht nachgewiesen.

In der anglophonen Forschung finden sich insgesamt nur wenige Studien zur Unterrichtsgestaltung im Kontext von ZP. In einer älteren englischen Studie stellen Scarth und Hammersley (1988) mithilfe von quantitativ ausgewerteten Unterrichtsbeobachtungen heraus, dass Lehrkräfte ihren Unterricht in für den Abschluss der Sek. I relevanten

Kursen lehrerzentrierter gestalten als in nicht abschlussrelevanten Kursen – unabhängig davon, ob die Kurse mit einer ZP oder auf anderem Wege abschließen. In einer quantitativen Fragebogenstudie in den USA werden zwei Prüfungssysteme mit unterschiedlichen Konsequenzen für Schulen und Schüler/innen verglichen (Vogler 2006, 2008; Vogler und Carnes 2009). Die Lehrkräfte in abschlussrelevanten Kursen geben gegenüber nicht abschlussrelevanten Kursen zwar an, durch die Konsequenzen für die Schüler/innen in ihrer Unterrichtsgestaltung beeinflusst zu werden, ein stärker lehrerzentrierter Unterricht wird in der Studie für abschlussrelevante Kurse aber faktisch nicht nachgewiesen. Hingegen arbeiten Krüger et al. (2013) in einer Interviewstudie heraus, dass erfahrene und für ihre Unterrichtsleistungen ausgezeichnete Lehrkräfte in Western Australia den Unterricht in abschlussrelevanten Biologiekursen in den letzten zwei Schuljahren angesichts der ZP lehrerzentrierter gestalten als in den Schuljahren davor.

Bezieht man in einer breiteren Perspektive schullaufbahnbegleitende, standardbasierte Leistungstests in die Betrachtung ein, so wird eine normierende Wirkung von Leistungstests auf den vorangehenden Unterricht als „washback“- bzw. „backwash“-Effekt beschrieben (vgl. z. B. Cheng und Curtis 2012), welcher vor allem in Sprachfächern untersucht und nachgewiesen wurde. Demnach werden über die Form der Prüfung in hohem Maße die Lernstrategien der Lernenden in der Phase der Prüfungsvorbereitung bestimmt – insbesondere, wenn diese auf die bloße Wiedergabe von Wissen und weniger auf verständnisorientiertes Lernen zielen. Darüber hinaus lassen sich auch „washback“-Effekte auf Lehrerkognitionen bzw. das Unterrichtshandeln von Lehrkräften, etwa mit Blick auf Inhalte und genutzte Materialien, nachweisen (vgl. z. B. die Metasynthese von Spratt 2005).

4.2 Unterrichtsentwicklung

Aus organisationstheoretischer Sicht ist davon auszugehen, dass ZP vor allem dann eine Verbesserung der Lernleistungen begünstigen, wenn die mit den Prüfungsergebnissen zurückgemeldeten Informationen von den Lehrkräften genutzt werden, um – *retroaktiv* für nachfolgende Schülerkohorten (vgl. Somerset 1996) – Unterricht zu verbessern. Dieser Aspekt ist bislang in der Forschung im deutschsprachigen Raum nicht und international wenig beleuchtet worden.

In England erschienen in den 1990er Jahren zwei Studien zur Nutzung von ZP zur datengestützten Schulentwicklung. Demnach fällt die *Datennutzung* im Kontext von ZP eher gering aus – auch dann, wenn die Schulen an entsprechenden Entwicklungsprojekten teilgenommen haben. Dabei zeigt sich in einer Fragebogenstudie, dass Fachzugehörigkeit, die Erfahrung im Projekt sowie die Teilnahme an Trainings die (insgesamt geringe) Datennutzung positiv beeinflussen (vgl. Tymms 1995), was u. a. auf die Bedeutung außerschulischer Unterstützungssysteme hinweist. In einer qualitativen Fallstudie zeigt sich zudem, dass die Nutzung zwischen Schulen und innerhalb der Schulen zwischen den einzelnen Fachbereichen erheblich variiert, was auch mit unterschiedlichen Überzeugungen auf Seiten der Schul- und Fachleitung zusammenhängt (vgl. Saunders und Rudd 1999; Saunders 2000).

In der bereits genannten finnisch/irisch/niederländischen Vergleichsstudie von Klein (2013) deutet sich zudem für die berichtete Datennutzung durch einzelne Lehrkräfte, ana-

log zu den organisationalen Strukturen (s. o.), eine deutliche Verankerung im rahmenden Prüfungs- und Steuerungssystem an. Auch hier ist in den Niederlanden die Nutzung stärker auf die datengestützte Unterrichtsentwicklung ausgerichtet. Hier dürften wiederum die mit den Prüfungen verknüpfte Rechenschaftspflicht der Schulen sowie die systematische Verknüpfung externer und interner Evaluationsverfahren zum Tragen kommen.

4.3 Wahrnehmung und Akzeptanz der Prüfung durch Lehrkräfte

Die Wahrnehmung des Prüfungsverfahrens (z. B. Unsicherheit und Belastungserleben) durch die Lehrkräfte dürfte in entscheidendem Maße die Unterrichtsvorbereitung/-durchführung beeinflussen. Insbesondere mit Blick auf neuere, auf Rechenschaftslegung und Schulentwicklung ausgelegte Steuerungsintentionen ist zudem bedeutsam, ob und inwieweit das Prüfungsverfahren von den Lehrkräften grundsätzlich als sinnvoll und nutzbringend wahrgenommen und akzeptiert wird.

Verschiedene Studien widmen sich im Zuge der Einführung von ZP in den deutschen Ländern der Einschätzung des neuen Verfahrens durch schulische Akteure. So deutet eine Befragung von Gymnasial- und Gesamtschullehrkräften in NRW an, dass die Lehrkräfte drei Jahre nach der Einführung des Zentralabiturs zwar die höhere Vergleichbarkeit der Prüfungen anerkennen, dabei aber vor allem hinsichtlich des in der Wahrnehmung eher niedrigen Leistungsniveaus und der eher geringen Aufgabenkomplexität Skepsis zeigen (Lorenz et al. 2011). Mit Blick auf die *Korrektur- und Bewertungsrichtlinien* erscheint die Wahrnehmung seitens der Lehrkräfte ambivalent und differiert jeweils in Abhängigkeit von Fach und Bundesland. So zeigt sich, dass die in der Längsschnittstudie von Maag Merki (2012b) befragten Lehrkräfte die jeweils länderspezifischen Korrektur- und Bewertungsrichtlinien eher positiv beurteilen (vgl. Appius und Holmeier 2012). Gleichwohl wird gerade im sprachlich-geisteswissenschaftlichen Bereich Kritik an sehr engen oder fehlerhaften Bewertungsvorgaben formuliert (Odendahl 2008; Sturm 2008).

Das durch Lehrkräfte wahrgenommene *Belastungserleben* und ein möglicher Leistungsdruck sind in Deutschland bislang nur mit Blick auf neu implementierte Prüfungsverfahren beleuchtet worden: In der Studie von Maag Merki (2012b) und Kolleginnen mit anschließender Ergänzungsstudie zeigt sich, dass die Lehrkräfte sich durch das Zentralabitur im zeitlichen Verlauf von 2007 bis 2011 zunehmend entlastet fühlen. Unsicherheiten gegenüber der Prüfung waren von Beginn an eher gering und haben sich mit zunehmender Erfahrung reduziert; zudem ist auch der – ohnehin eher gering ausgeprägte – gefühlte *Leistungsdruck* im Kontext des Zentralabiturs gesunken (vgl. Maué et al. 2012; Oerke 2012a).

Mit Blick auf die Wahrnehmung des Prüfungsverfahrens zeigt sich in der Vergleichsstudie von Klein (2013), dass diese insgesamt in allen drei beobachteten Staaten sehr positiv ausfällt, sich gleichwohl wiederum Bezüge zwischen der Gestaltung schulischer Steuerung und der Wahrnehmung der Prüfung herstellen lassen. So stimmen niederländische Lehrkräfte innerhalb des stärker auf Accountability ausgerichteten Systems eher zu, dass die Ergebnisse ein guter Indikator für die Qualität ihrer Schule sind. Geht es nicht um das Prüfungssystem als Ganzes, sondern um Änderungen (z. B. der Prüfungsstandards), legt die Studie von Krüger et al. (2013) mit australischen Lehrkräften nahe, dass die Änderungen vor allem dann als positiv wahrgenommen werden, wenn sie den

Schüler/inne/n zugutekommen (z. B. curriculare Innovationen mit vermehrtem Lebensweltbezug) und eher negativ, wenn die Rigorosität des Prüfungsverfahrens erhöht und dadurch ein Rückgang der Lernfreude und Kursbelegung befürchtet wird.

In den USA werden Einstellungen gegenüber ZP bislang vor allem im Rahmen qualitativer Studien erhoben, deren Fokus insbesondere auf Accountability und in diesem Kontext auf Motivation und Belastungserleben schulischer Akteure liegt. Hier ergeben sich einerseits Hinweise darauf, dass Lehrkräfte – im Gegensatz zu Schulleitung und District Administration – nicht in erster Linie durch die Konsequenzen für die Schulen dazu motiviert werden, die Prüfungsstandards in ihrem Unterricht umzusetzen. Vielmehr ist es vor allem der Wunsch, den bestmöglichen Abschluss für die Schüler/innen zu erreichen (Sipple et al. 2004, New York). Für die Frage, ob zentrale Standards angenommen und umgesetzt werden, stehen demnach nicht nur der Prüfungserfolg, sondern vor allem auch die Lernentwicklung der Schüler/innen (vgl. ebd.) sowie das professionelle Selbstverständnis der Lehrkräfte im Vordergrund (Goertz und Massell 2005; Massell et al. 2005, USA). Andererseits kann sich gerade in Systemen, die Konsequenzen für Schüler/innen und Schulen miteinander verbinden, ein Konflikt zwischen den Anforderungen ergeben. Dies ist z. B. dann der Fall, wenn Lehrkräfte ihren Schüler/inne/n die Information vorenthalten, dass sie ein schlechtes Prüfungsergebnis auf anderem Wege kompensieren können, damit diese trotzdem versuchen, in der auch für die Schule im Rahmen der Rechenschaftslegung relevanten Prüfung das bestmögliche Ergebnis zu erreichen (vgl. Zhang 2009, Arizona).

Ebenfalls ambivalente Einstellungen gegenüber den ZP deuten sich im Kontext einer qualitativen Fallstudie in England und Flandern an, in der die englischen Lehrkräfte zwar einerseits die Relevanz der Prüfungen als Selektions- und Evaluationsinstrument anerkennen, andererseits aber die Validität der Aussagen über die Schülerleistungen hinterfragen (Stevens und van Houtte 2011). Dabei zeigt sich zumindest in England, dass Wahrnehmungen von Kontrollverlust, Frustration und fehlende Möglichkeiten der kreativen Entfaltung in der Sek. I nicht nur in Schulen mit einem hohen Anteil an Schüler/inne/n aus eher bildungsfernen Milieus berichtet werden, sondern auch von Lehrkräften aus solchen Schulen, deren Schüler/innen in den Prüfungen durchschnittlich abschnitten (Perryman et al. 2011).

5 Unterrichtliche Mikroebene II: Schülerperspektive

In der Literatur wird im Hinblick auf das motivationale und emotionale Erleben der Schüler/innen auf der einen Seite angenommen, dass sie angesichts der großen Relevanz der ZP grundsätzlich eine höhere Anstrengungsbereitschaft und Leistungsmotivation zeigen, was sich letztlich in höheren Prüfungsleistungen niederschlagen könne (Bishop 1999). Demgegenüber wird neben einer vermuteten höheren Belastung (z. B. größere Prüfungsangst) durch die nicht mehr von der Fachlehrkraft gestellten Prüfungen (vgl. Baumert und Watermann 2000) angenommen, dass sich die Betonung prüfungsrelevanter Themen und die damit verbundene thematische Engführung sowie der Wegfall interessanter Themen negativ auf die Motivation auswirken könnten.

Hierzulande ist das motivationale und emotionale Erleben von Schüler/inne/n in Verbindung mit ZP bislang nur wenig erforscht. Reanalysen der TIMSS/III-Daten zeigen einerseits keinen Zusammenhang zwischen zentralen gegenüber dezentralen Abiturprüfungen und einer erhöhten *fachspezifischen Prüfungsangst* (vgl. Baumert und Watermann 2000). Andererseits werden – ebenfalls auf der Basis von Reanalysen der TIMSS/III-Daten – hinsichtlich eines höheren Schwierigkeitsgrades und mehr Langeweile auch negative Effekte auf das motivational-emotionale Erleben der Lernenden im Zentralabitur für das Fach Mathematik beschrieben (Jürges et al. 2009; Jürges und Schneider 2010). Im Zusammenhang mit der Einführung des Zentralabiturs in Bremen und Hessen variieren teilweise die Wirkungen auf *fachliches Interesse, Anstrengungsbereitschaft (Persistenz), Misserfolgsängstlichkeit, Unsicherheit und schulische Selbstwirksamkeitsüberzeugungen* der Schüler/innen in Abhängigkeit vom Fach, Kursniveau und Land (Maag Merki 2012c; Oerke 2012a).

International liegt der Fokus im Bereich des motivational-emotionalen Erlebens von Schüler/inne/n in ZP vor allem auf der Erforschung von durch die Prüfungen induziertem Stress und Prüfungsangst. Studien aus Australien (Smith und Sinclair 2000), England (Denscombe 2000; Putwain 2009) und Marokko (Bensmansour 1999) weisen darauf hin, dass die Teilnahme an ZP als stressinduzierendes Ereignis wahrgenommen wird. Darüber hinaus werden in England Zusammenhänge zwischen Prüfungsangst und -leistungen (Putwain 2008a; Putwain et al. 2010; Chamberlain et al. 2011) sowie damit verbundenen Konsequenzen für die Schüler/innen (vgl. Putwain 2008a) in Form schlechter Prüfungsergebnisse und der daraus ggf. resultierenden eingeschränkten beruflichen Zukunftsperspektiven identifiziert. Weitere Studien zeigen, dass das Geschlecht (vgl. Putwain 2011) und der sozioökonomische Status (vgl. Putwain 2008b) einen moderierenden Effekt auf den Zusammenhang zwischen Prüfungsangst und Prüfungsleistungen haben. Gemäß diesen Befunden weisen allgemein Mädchen sowie Prüflinge mit niedrigem sozioökonomischen Status einen höheren Grad an Prüfungsangst auf und besitzen dadurch eine größere Wahrscheinlichkeit, in den Prüfungen schlechter abzuschneiden. Mehrere der genannten Untersuchungen machen zudem deutlich, dass die Wahrnehmung von Stress und Prüfungsangst zwischen den Prüflingen deutlich divergiert und z. B. in Abhängigkeit vom Lehrerenthusiasmus oder der Unterrichtsmethodik variiert (vgl. z. B. Bensmansour 1999; Putwain 2009, 2011).

6 Effekte zentraler Abschlussprüfungen auf Schülerleistungen

Schließlich stellt sich die Frage, inwiefern ZP Wirkungen auf Lernergebnisse entfalten können. Dabei interessiert einerseits, inwiefern sie insgesamt dazu beitragen, den *Leistungsstand* zu erhöhen; andererseits, inwieweit sich *Bildungsbenachteiligungen* ausgleichen lassen oder weiter verfestigen⁹. Hinsichtlich dieser Fragestellungen werden meist Lernleistungen bzw. Leistungszuwächse in Systemen mit und ohne ZP unter Berücksichtigung weiterer Prädiktoren auf individueller, schulischer und institutioneller Ebene miteinander verglichen (vgl. Holme et al. 2010).

Einige international sowie zwischen den deutschen Ländern vergleichende Studien (z. B. Bishop 1997, 1998; Wößmann 2007, 2008) kommen zu dem Ergebnis, dass Schü-

ler/innen in Staaten bzw. Ländern mit ZP am Ende der Sek. II deutlich bessere Leistungen erbringen als in Staaten ohne ZP: „Über vier internationale Schüler Leistungsvergleiche [...] hinweg, zeigen umfassende Regressionsanalysen der Schülerindividuen, 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). Die Postulierung eines Kausalzusammenhangs zwischen ZP und Leistungen erweist sich allerdings als voraussetzungsreich (Schümer und Weiss 2008). Zudem ist die Dichotomisierung der Prüfungsverfahren (zentral/dezentral) teilweise schlecht validiert (Cosentino de Cohen 2010) und bildet die Heterogenität sowohl von ZP als auch dezentralen Abschlussverfahren nicht hinreichend ab (Klein et al. 2009). Es kann auch hinterfragt werden, ob sich aus den Leistungen 13-Jähriger (TIMSS/II) und 15-Jähriger (PISA) eine Fernwirkung auf die Abschlussprüfungen am Ende der Sek. II ablesen lässt. Zudem werden Zusammenhänge zwischen den Kompetenzen der Schüler/innen und anderen Faktoren nicht auf der Individualebene, sondern mithilfe auf Länder- bzw. Staatsebene aggregierter Daten berechnet. Andere Studien, welche diese Aspekte (zum Teil) berücksichtigen, kommen daher zu zurückhaltenderen Einschätzungen.

Eine Reanalyse der PISA-E 2003-Daten, in der nur die Leistungen von Schüler/innen berücksichtigt werden, welche die abiturrelevante Schulform Gymnasium besuchten, kann eine generelle Überlegenheit von Ländern mit Zentralabitur gegenüber Ländern mit dezentralem Abitur nicht bestätigen (Block et al. 2011). Büchel et al. (2003) berechnen mit TIMSS/II-Daten den Effekt der Prüfungsorganisation am Ende der Sek. I auf Mathematikleistungen in der 7. Klasse und beziehen als Kontrollgröße auch Leistungen im Nicht-Prüfungsfach Naturwissenschaften ein. Die Autor/innen berichten einen Vorteil für Länder mit ZP von 0,12 Standardabweichungen. Baumert und Watermann (2000) nutzen TIMSS/III-Daten, um die Mathematik- und Physikleistungen in Bundesländern mit und ohne Zentralabitur zu vergleichen und machen dabei keinen generellen Effekt des Zentralabiturs, sondern kurs- und fachspezifische Wirkungen aus (zugunsten von Mathematik-Grundkursen). Zu vergleichbaren Ergebnissen kommt auch eine Analyse der durchschnittlichen Leistungen sowie der Leistungsstreuung von Schüler/innen in Englisch und Mathematik in den ersten drei Jahren nach Einführung des Zentralabiturs in Bremen und Hessen (Maag Merki 2012a), wobei sich substantielle Veränderungen im Leistungsniveau nur im Mathematik-Grundkurs in Bremen zeigen; hier reduzierte sich auch systematisch die Gruppe der Schüler/innen auf dem niedrigsten Niveau, ohne dass die obere Leistungsgruppe kleiner wurde. Da für diesen Kurs allerdings lediglich Daten nach der Einführung des Zentralabiturs vorliegen, lässt sich hieraus nicht zwingend ein Effekt der Prüfungsform ablesen. In den Kursen, für die Informationen für beide Konstellationen (mit/ohne ZP) vorliegen, lassen sich keine substantiellen Veränderungen in den Leistungen der Schüler/innen feststellen.

International finden sich zudem Studien, welche die Dichotomisierung von Prüfungsverfahren umgehen und dadurch zu differenzierteren Ergebnissen kommen. In einer Reanalyse der durchschnittlichen Leistung sowie der Leistungsstreuung in TIMSS/III 1995 bezieht Cosentino de Cohen (2010) im internationalen Vergleich auch die Konsequenzen für Schüler/innen – etwa inwiefern auch Resultate interner Leistungsbewertung in die Gesamtqualifikation mit hineinspielen – in die Kategorisierung der Prüfungsverfahren

mit ein. Dabei deuten sich allein aus der Interaktion zwischen externer Abschlussprüfung und hohen Konsequenzen für die Schüler/innen positive Effekte auf die Leistungsstärke vor allem für Schüler/innen im unteren Leistungsspektrum sowie eine insgesamt eher geringere Leistungsstreuung an. In einem Vergleich kanadischer Provinzen mit und ohne ZP wird dagegen kein systematischer Zusammenhang zwischen Prüfungsorganisation und Leistung berichtet (vgl. ebd.). Auf Basis von Längsschnittdaten setzt sich Shuster (2012) mit Leistungszuwächsen in Mathematik in 49 US-Bundesstaaten mit unterschiedlichen Abschlussprüfungen (*minimum competency*, *standards-based* oder *end-of-course exams*¹⁰) auseinander. *End-of-course exams* als der Prüfungstyp, der den deutschen Prüfungsverfahren am nächsten kommt, haben demnach zwar einen kleinen positiven Effekt auf die Leistungen der Schüler/innen im oberen Leistungsquartil, hängen in den unteren beiden Leistungsquartilen allerdings – anders als bei Cosentino de Cohen (2010) – eher mit einem geringeren Leistungszuwachs zusammen.

Neben dem hier skizzierten *Student Achievement* werden insbesondere in der US-amerikanischen Forschung auch die Wirkungen zentraler Abschlussprüfungen auf *Student Attainment* – etwa mit Blick auf Wiederholer- bzw. Schulabbrecherquoten (z. B. Griffin und Heidorn 1996; Greene und Winters 2004; Carnoy 2005; Dee und Jacob 2006) – sowie auf *Student Postsecondary Outcomes* – etwa hinsichtlich des Zugangs zu Universitäten (z. B. D’Agostino und Bonner 2009) oder zum Arbeitsmarkt (z. B. Martorell 2004) – untersucht. In beiden Kategorien zeigen sich kleinere Effekte vor allem „strengerer“ Prüfungsverfahren auf die Erfolge von leistungsschwächeren Schüler/innen, welche zudem durch die soziale Herkunft der Schüler/innen konfundiert sind (vgl. dazu im Überblick Holme et al. 2010).

7 Fazit

Im Beitrag wird der bisherige Forschungsstand zu zentralen Abschlussprüfungen, einem aus der Perspektive vieler deutscher Länder vergleichsweise neuen Thema, auf mehreren Ebenen des Schulsystems in prozess- und outputorientierter Hinsicht mit internationalen Befunden zusammengeführt.

Auf der *Makroebene der Bildungspolitik und -administration* zeigen sich vielgestaltige und ausdifferenzierte Steuerungsintentionen, die sich von der grundlegenden Funktion der Zertifizierung und Sicherung von Vergleichbarkeit hin teils zu Rechenschaftslegung und zu Erwartungen an Schulentwicklungsprozesse ausgeweitet haben. Landesspezifische Muster der bildungspolitischen und -administrativen Wirkungserwartung scheinen dabei durch kulturell und historisch verwurzelte Rahmenbedingungen in der Steuerung der Bildungssysteme beeinflusst zu sein; insofern erweist sich eine entsprechende systematische Analyse der Makroebene unter Berücksichtigung kultureller Hintergrundmerkmale als ertragreich für ein vertiefteres Verständnis von Steuerungsintentionen und -wirkungen. In Bezug auf die Aufgabengestaltung als zentralem Kern der Prüfungen finden sich keine einheitlichen Befunde; vielmehr variieren diese ebenfalls nach (Bundes-)Land, Unterrichtsfach und Kursniveau. Diese Ausdifferenzierung wird sich in Deutschland möglicherweise durch die Einführung eines gemeinsamen Aufgabenpools im Sinne besserer Vergleichbarkeit verändern und ggf. auch Wirkungen auf Unterrichts-

ebene entfalten können (etwa im Rahmen innovativer Aufgabenformate, z. B. mit Blick auf eine stärkere Kontextorientierung). Zugleich geht der stärkere Zentralisierungs- und Standardisierungsanspruch mit einer Wahloption bezüglich der Aufgabennutzung auf dezentraler Länderebene einher; auch die schulinterne Erstkorrektur bleibt erhalten, so dass es weiterhin zentrale und dezentrale Momente im Prüfungsverfahren „zentraler Abschlussprüfungen“ geben wird. Das Wechselspiel von Zentralität und Dezentralität im Prüfungsverfahren schafft letztlich einen Ausgleich zwischen der Anforderung der Vergleichbarkeitssicherung und Qualitätsentwicklung (mit einer Signalwirkung gegenüber nachfolgenden Bildungs- und Ausbildungssystemen und dem Arbeitsmarkt) sowie dem Anknüpfen an die hierzulande traditionell starke pädagogische Professionalität von Lehrerinnen und Lehrern.

Als Forschungsdesiderat erweist sich die Analyse von Bewertungs- und Korrekturvorgaben, insbesondere für die deutsche Situation, wo die Prüfungen schulintern ausgewertet werden. Hier scheint es zudem in der Wahrnehmung der Vorgaben fachspezifisch unterschiedlich stark empfundene Einschränkungen des eigenen professionellen Handelns der Lehrer/innen zu geben, die offensichtlich mit der Fachstruktur und -kultur einhergehen (z. B. stärker in Deutsch als in Mathematik).

Auf der *schulischen Mesoebene* steht bislang die Kooperation von Lehrkräften im Kontext veränderter Prüfungsverfahren im Vordergrund. Lediglich in der Anfangsphase der Implementation scheint es eine Reduktion von Unsicherheiten durch Zusammenarbeit zu geben. International weisen die Ergebnisse darauf hin, dass die Bewältigung neuer Prüfungsanforderungen in Schulen bereits vorhandener Leitungs- und Kooperationsstrukturen bzw. insbesondere bei fehlenden Ressourcen und Kompetenzen entsprechender Unterstützungsstrukturen durch Externe bedarf. Entsprechende Hinweise finden sich auch in der Literatur zur evidenzbasierten Schulentwicklung (vgl. z. B. van Ackeren et al. 2013).

Die *unterrichtliche Mikrobene* erweist sich als Kernbereich der für Deutschland vorliegenden Forschung. Zu den vermuteten Wirkungen der ZP auf eine eingeschränkte Themenvarianz im Unterricht zeigen sich nach der Einführung des Zentralabiturs eher geringe Wirkungen. Insgesamt lässt die Befundlage keinen generellen ZP-Effekt auf die Oberflächen- und Tiefenstruktur des Unterrichts vermuten; vielmehr scheint die Unterrichtsgestaltung wiederum fach-, kurs- und länderspezifisch zu variieren. Hierbei ist neben kulturellen Unterschieden (zwischen Fächern wie zwischen Ländern) auch zu vermuten, dass die heterogen gestalteten Prüfungsverfahren und -aufgaben sowie die damit verbundenen unterschiedlich weitreichenden Konsequenzen die Lehrkräfte auf unterschiedliche Weise dazu veranlassen, ihren Unterricht zu verändern oder eben nicht.

Bei der Wahrnehmung und Akzeptanz der Prüfung durch Lehrkräfte gibt es Hinweise auf Implementationseffekte. Routine und Sicherheit im Umgang mit neuen Prüfungsverfahren kehren zunehmend ein. International zeigen sich recht ambivalente Einstellungsmuster mit Blick auf die Sicherung von Qualität und Vergleichbarkeit oder auch die eingeschätzte Validität der Prüfungen. Gleichwohl bleibt zu bedenken, dass die Befunde aus unterschiedlichen Staaten aufgrund unterschiedlicher Forschungszugänge und Prüfungsstrukturen nicht unmittelbar vergleichbar sind.

Im Hinblick auf die Schülerwahrnehmung lassen sich in erster Linie nicht-intendierte Wirkungen auf das motivational-emotionale Erleben beobachten (u. a. Prüfungsangst,

fachliches Interesse, Motivation). Die Befundlage in Deutschland ist jedoch dünn, inkonsistent und methodisch optimierbar, die internationale Befundlage basiert ebenfalls auf wenigen Studien aus unterschiedlichen Bildungskontexten.

Mit Blick auf das erreichte Leistungsniveau der Schüler/innen scheinen fach- und kurspezifische Ausprägungen bedeutsam zu sein. Die vorliegenden Befunde für Deutschland weisen – bei aller vorsichtigen Interpretation – möglicherweise in Richtung eines positiven Effekts (höhere Leistungen, Verringerung der Gruppe auf der untersten Kompetenzstufe) für das Fach Mathematik auf Grundkursniveau. Der Sachverhalt, dass die möglichen Steuerungseffekte in Abhängigkeit vom Abiturfach deutlich differieren, kann als Indiz dafür verstanden werden, dass keine generellen Zentralprüfungseffekte existieren. Die Diskussion möglicher fach- und kursniveauspezifischer Effekte des Zentralabiturs steht allerdings erst am Anfang und bedarf der Vertiefung. Aus einer governance-theoretischen Perspektive erscheinen in diesem Zusammenhang fachdidaktische Analysen allein nicht ausreichend. Die potenziellen fachdifferenziellen Effekte des Zentralabiturs deuten möglicherweise auch auf ein fächerspezifisch differierendes Professionsverständnis der Lehrkräfte und/oder besondere fachkulturelle Unterschiede hin. Die internationalen Befunde verweisen zudem darauf, dass vermutlich auch Prüfungsformate sowie damit verknüpfte Konsequenzen für die Schüler/innen eine Rolle dafür spielen dürften, welche Wirkungen ZP auf die tatsächlichen Fachleistungen der Schüler/inne/n entfalten können.

Die Bilanzierung zeigt insgesamt, dass ein direkter Vergleich der Studienergebnisse durch die sehr unterschiedlichen Prüfungsstrukturen sowie durch die differenziellen Forschungszugänge erschwert wird. Was sich unter dem Etikett „Zentrale Abschlussprüfung“ zunächst als einheitliches Steuerungsinstrument darstellt, erweist sich letztlich als zu differenzierendes Konstrukt; entsprechende quantitative wie qualitative Analysen haben somit Fallstudiencharakter. Dies ist durchaus eine zentrale Erkenntnis vor allem der nationalen Studien der vergangenen Jahre. In diesem Zusammenhang wird aber auch deutlich, dass die vorliegenden empirischen Befunde zu ZP sowohl national als auch international auf unterschiedliche Forschungsdesigns, Stichproben, Erhebungsinstrumente zurückgeführt werden können. Die Zuverlässigkeit der empirischen Aussagen kann damit grundsätzlich erheblich schwanken. Beispielsweise sind einige Studien retrospektive Lehrer- und Schülerbefragungen, so dass Schlussfolgerungen auf mögliche Veränderungen im Unterricht vorsichtig zu diskutieren sind. Schließlich fällt auf, dass Befragungen und Interviews als methodische Zugänge überwiegen; beobachtende Studien z. B. zu Fragen des Prüfungscoachings bzw. *teaching to the test* finden sich nicht.

Unter dem Strich bleiben Zweifel an der generellen Steuerungswirksamkeit des Zentralabiturs – zumal in den vielfältigen, aber wenig standardisierten Organisationsformen, in denen die Abschlussprüfungen derzeit in den einzelnen Bundesländern implementiert sind. Im internationalen Vergleich lässt sich zudem konstatieren, dass sich zwar zum Teil Hinweise auf Veränderungen in schulischen Organisationsprozessen und in den Lehrerkognitionen finden, diese sich aber aufgrund der differenziellen Accountability-Systeme, in die die Prüfungen eingebettet sind, auf Basis der bisherigen Befunde nicht eindeutig auf den Kern der zentralen Abschlussprüfungen zurückbinden lassen und darüber hinaus einer quantitativen Überprüfung bedürfen. Systematische Effekte im Sinne einer *intendierten* Verbesserung von Unterrichts- und Arbeitsprozessen sowie einer Sicherung von Leistungsstandards lassen sich aber auch hier nicht stabil abbilden. Die empirischen

Befunde deuten daher insgesamt eher auf eine – gemessen an den Ansprüchen – eingeschränkte Steuerungswirksamkeit zentraler Abschlussprüfungen hin.

Anmerkungen

- 1 Während ein Teil der Länder auf eine längere ZP-Tradition zurückblickt (z. B. Bayern, Baden-Württemberg), ist diese Form der Prüfungsorganisation in der Mehrheit der Länder erst im Kontext der bildungspolitischen Debatten über die Entwicklung schulischer Qualität in den letzten Jahren implementiert worden. International finden sich hingegen zahlreiche Länder mit längerer Tradition (vgl. Klein et al. 2009).
- 2 vgl. z. B. Klein 2013 oder den „Leitfaden zur Nutzung der Ergebnisse zentraler Prüfungen“ des Ministeriums für Bildung und Kultur Schleswig-Holstein; http://www.schleswig-holstein.de/Bildung/DE/Service/Broschueren/Bildung/NutzungErgebnisseZentralePruefungen__blob=publicationFile.pdf [03.12.2013].
- 3 intendiert: im Sinne der Zielvorgaben (Vergleichbarkeit, Qualitätsentwicklung etc.); nicht intendiert: z. B. verstärkte Nicht-Versetzungen und Abschulungen, thematische und methodische Engführung von Unterricht.
- 4 Nur eine geringe Zahl von US-Bundesstaaten (z. B. New York) hatte bereits vor 2000 sogenannte *High School Exit Exams* implementiert (z. B. Zabala et al. 2008, S. 5 ff.); diese waren mehrheitlich als Minimum Competency Tests konzipiert und entsprechen damit nicht der Definition von „zentralen Abschlussprüfungen“, die diesem Beitrag zugrundeliegt; erst etwa während des vergangenen Jahrzehntes haben verschiedene US-Bundesstaaten begonnen, auf andere, stärker auf tatsächliche Kursinhalte abzielende Prüfungsformate umzustellen (vgl. dazu auch Endnote 10).
- 5 So beispielsweise in Kalifornien, siehe <http://www.cde.ca.gov/ta/ac/ay/documents/aypinfoguide13.pdf>.
- 6 Lorenz et al. (2013) überprüfen z. B. die Fairness von Abituraufgaben im Fach Englisch (NRW) mit Blick auf das Geschlecht der Prüflinge, können eine differenzielle Itemfunktion, welche darauf hinweisen würde, dass die Aufgaben systematisch Mädchen oder Jungen benachteiligen, aber nur bei wenigen Einzelitems feststellen.
- 7 Für das erfolgreiche Bearbeiten und Lösen einer Aufgabe sind kognitive Prozesse unterschiedlichen Komplexitätsgrades anzuwenden – von einfachen Routinetätigkeiten (z. B. Wiedergabe von Fakten) bis hin zur Bearbeitung komplexer Fragestellungen (z. B. um Probleme zu formulieren). Der Forschungsstand macht fachübergreifend deutlich, dass in ZP weitgehend Aufgaben fehlen, die reflektierende, verallgemeinernde oder bewertende (d. h. kognitiv anspruchsvolle) Anforderungen zum Gegenstand haben.
- 8 Hier fielen Elaborationsstrategien vor allem in zentral geprüften Leistungskursen höher aus.
- 9 In dieser Perspektive liegt der Fokus in der Regel auf der Benachteiligung bestimmter sozialer Gruppen (z. B. *ethnic minorities*) z. B. mit Blick auf Nichtbestehensquoten. Die entsprechenden Studien kommen dabei zu inkonsistenten Ergebnissen (vgl. hierzu auch Holme et al. 2010). Für Deutschland liegt bislang eine vergleichbare Studie vor, die auf Basis von Daten aus NRW vermuten lässt, dass durch das neu eingeführte Zentralabitur insbesondere für Schüler/innen in Gesamtschulen geringere Chancen bestehen, das Abitur zu erreichen, wobei sich hier zusätzlich Effekte des Migrationsstatus zeigen (vgl. Schräpler und Weishaupt 2013).

- 10 *Minimum competency exams* testen grundlegende Kenntnisse unterhalb des High School-Niveaus. Während *standards-based exams* (teilweise auch *comprehensive exams*) prüfen, ob Schüler/innen in verschiedenen Kernfächern die für den High School-Abschluss bestimmten State Standards erreicht haben, testen *end-of-course exams* explizit, ob die Schüler/innen die Inhalte spezifischer High School-Kurse beherrschen (vgl. hierzu auch Zabala et al. 2008).

Literatur

- Achieve, Inc. (2004). *Do graduation tests measure up? A closer look at state high school exit exams*. Washington, D. C.: Achieve, Inc. <http://www.achieve.org/files/TestGraduation-Final-Report.pdf>. Zugegriffen: 26. Aug. 2013.
- van Ackeren, I., Heinrich, M., & Thiel, F. (2013). *Evidenzbasierte Steuerung im Bildungssystem? Befunde aus dem BMBF-SteBis-Verbund* (Die Deutsche Schule, Beiheft 12). Münster: Waxmann.
- van Ackeren, I., Block, R., Klein, E. D., & Kühn, S. M. (2012). The impact of statewide exit exams: A descriptive case study of three German states with differing low stakes exam regimes. *Education Policy Analysis Archives*, 20(8), 1–28. <http://epaa.asu.edu/ojs/article/view/1011>. Zugegriffen: 26. Aug. 2013.
- Altrichter, H., & Maag Merki, K. (2010). *Handbuch Neue Steuerung im Schulsystem*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Appius, S. (2012). Kooperation zwischen Lehrpersonen im Zusammenhang mit dem Abitur. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Prozesse und Wirkungen der Einführung zentraler Abiturprüfungen in zwei Bundesländern* (S. 95–117). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Appius, S., & Holmeier, M. (2012). Beurteilung der Abituraufgaben und Korrekturhinweise. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Prozesse und Wirkungen der Einführung zentraler Abiturprüfungen in zwei Bundesländern* (S. 353–381). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Baumert, J., & Watermann, R. (2000). Standardisierung durch die Abiturprüfung: Zentralabitur oder dezentrale Prüfungsorganisation? In J. Baumert, W. Bos & R. Lehmann (Hrsg.), *TIMSS/III. Dritte internationale Mathematik- und Naturwissenschaftsstudie-mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn* (Bd. 2, S. 341–351). Opladen: Leske+Budrich.
- Baumgart, N., & Halse, C. (1999). Approaches to learning across cultures: The role of assessment. *Assessment in Education*, 6(3), 321–339.
- Bensmansour, N. (1999). Motivational Orientations, Self-efficacy, Anxiety, and Strategy Use in Learning High School Mathematics in Morocco. *Mediterranean Journal of Educational Studies*, 4(1), 1–15.
- van den Bergh, H., Rohde, E., & Zwarts, M. (2003). Is het ene examen het andere? Over de stabiliteit van schoolonderzoek en centraal examen. *Pedagogische Studiën*, 80(3), 176–191.
- Berkemeyer, N., Bos, W., & Manitius, V. (2012). *Chancenspiegel. Zur Chancengerechtigkeit und Leistungsfähigkeit der deutschen Schulsysteme*. Gütersloh: Bertelsmann.
- Bishop, J. H. (1997). The effect of national standards and curriculum-based exams on achievement. *The American Economic Review*, 87(2), 260–264.
- Bishop, J. H. (1998). The effect of curriculum-based external exit exams on student achievement. *Journal of Economic Education*, 29(2), 172–182.
- Bishop, J. H. (1999). Are national exit examinations important for economic efficiency? *Swedish Economic Policy Review*, 6(2), 349–401.

- Block, R., Klein, E. D., van Ackeren, I., & Kühn, S. M. (2011). Leistungseffekte des Zentralabiturs? Eine kritische Auseinandersetzung mit bildungsökonomischen Interpretationen zu den Effekten der Prüfungsorganisation auf der Basis von PISA E 2003-Daten. *Bildungsforschung*, 8(1), 215–238.
- Britton, E. D., & Raizen, S. A. (1996). *Examining the examinations. An international comparison of science and mathematics examinations for college-bound students*. Boston: Kluwer Academic Publishers.
- Brüsemeister, T. (2007). Steuerungsakteure und ihre Handlungslogiken im Mehrebenensystem der Schule. In J. Kussau & T. Brüsemeister (Hrsg.), *Governance, Schule und Politik. Zwischen Antagonismus und Kooperation* (S. 63–95). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Büchel, F., Jürges, H., & Schneider, K. (2003). Die Auswirkungen zentraler Abschlussprüfungen auf die Schulleistung: Quasi-experimentelle Befunde aus der deutschen TIMSS-Stichprobe. *Vierteljahreshefte zur Wirtschaftsforschung*, 72(2), 238–251.
- Carnoy, M. (2005). Have state accountability and high-stakes tests influenced student progression rates in high school? *Educational Measurement: Issues and Practice*, 24(4), 19–31.
- Chamberlain, S., Daly, A. L., & Spalding, V. (2011). The fear factor: Students' experiences of test anxiety when taking A-level examinations. *Pastoral Care in Education*, 29(3), 193–205.
- Cheng, L., & Curtis, A. (2012). Test impact and washback: Implications for teaching and learning. In C. Coombe, B. O'Sullivan, P. Davidson & S. Stoyhoff (Hrsg.), *Cambridge guide to second language assessment* (S. 89–95). Cambridge: Cambridge University Press.
- Cosentino de Cohen, C. (2010). *Examination regimes and student achievement*. Dissertation, Princeton University. Princeton: Princeton University.
- Crisp, V. (2010). Judging the grade: Exploring the judgement processes involved in examination grading decisions. *Evaluation & Research in Education*, 23(1), 19–35.
- D'Agostino, J. V., & Bonner, S. M. (2009). High school exit exam scores and university performance. *Educational Assessment*, 14(1), 24–37.
- DeBray, E. (2005). A comprehensive high school and a shift in New York state policy: A study of early implementation. *The High School Journal*, 89(1), 18–45.
- DeBray, E., Parson, G., & Woodworth, K. (2001). Patterns of response in four high schools under state accountability policies in Vermont and New York. *Yearbook of the National Society for the Study of Education*, 100(2), 170–192.
- Dee, T. S., & Jacob, B. A. (2006). *Do high school exit exams influence educational attainment or labor market performance?* (NBER Working Paper 12199). Cambridge. National Bureau of Economic Research.
- Denscombe, M. (2000). Social conditions for stress: Young people's experience of doing GCSEs. *British Educational Research Journal*, 26(3), 359–374.
- Eickelmann, B., Kahnert, J., Lorenz, R., & Bos, W. (2011). Das Zentralabitur in Nordrhein-Westfalen aus der Lehrerperspektive: Veränderungen für den Unterricht. *Schulverwaltung NRW*, 22(11), 31–32.
- Florian, C., Schmiemann, P., & Sandmann, A. (2011). Schriftliche Prüfungsaufgaben im Zentralabitur Biologie: Ergebnisse deskriptiver Aufgabenanalysen. In D. Krüger, A. Uppmeier zu Belzen, P. Schmiemann & A. Sandmann (Hrsg.), *Erkenntnisweg Biologiedidaktik 10* (S. 69–86). Kassel: Universitätsdruckerei Kassel.
- Goertz, M. E., & Massell, D. (2005). Holding high hopes: How high schools respond to state accountability policies (CPRE Policy Briefs). http://www.cpre.org/sites/default/files/policy-brief/882_rb42.pdf. Zugegriffen: 22. Jan. 2014.
- Greene, J. P., & Winters, M. A. (2004). *Pushed out or pulled up? Exit exams and dropout rates in public high schools* (Education Working Paper 5). http://www.manhattan-institute.org/html/ewp_05.htm. Zugegriffen: 22. Jan. 2014.
- Griffin, B. W., & Heidorn, M. H. (1996). An examination of the relationship between minimum competency test performance and dropping out of high school. *Educational Evaluation and Policy Analysis*, 18(3), 243–252.

- Holme, J. J., Richards, M. P., Jimerson, J. B., & Cohen, R. W. (2010). Assessing the effects of high school exit examinations. *Review of Educational Research, 80*(4), 476–526.
- Holmeier, M. (2012). Bezugsnormorientierung im Unterricht im Kontext zentraler Abschlussprüfungen. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Prozesse und Wirkungen der Einführung zentraler Abiturprüfungen in zwei Bundesländern* (S. 237–261). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Holmeier, M. (2013). *Leistungsbeurteilung im Zentralabitur*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Jäger, D. J. (2012). Herausforderung Zentralabitur: Unterrichtsinhalte variieren und an Prüfungsthemen anpassen. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Prozesse und Wirkungen der Einführung zentraler Abiturprüfungen in zwei Bundesländern* (S. 179–205). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Jäger, D. J., Maag Merki, K., Oerke, B., & Holmeier, M. (2012). Statewide low-stakes tests and a teaching to the test effect? An analysis of teacher survey data from two German states. *Assessment in Education: Principles, Policy & Practice, 19*(4), 451–467.
- Jürges, H., & Schneider, K. (2010). Central exit examinations increase performance... But take the fun out of mathematics. *Journal of Population Economics, 23*(2), 497–517.
- Jürges, H., Schneider, K., Senkbeil, M., & Carstensen, C. H. (2009). *Assessment drives learning: The effect of central exit exams on curriculum knowledge and mathematical literacy (CESifo working paper, 2666)*. München: Ifo Institute for Economic Research.
- Klein, E. D. (2013). *Statewide exit exams, governance, and school development: An international comparison*. Münster: Waxmann.
- Klein, E. D., Kühn, S. M., van Ackeren, I., & 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.
- Klemm, K. (2004). Strukturfragen und kein Ende. In H. G. Holtappels, H. Pfeiffer, H.-G. Rolff, R. Schulz-Zander & K. Klemm (Hrsg.), *Jahrbuch der Schulentwicklung. Daten, Beispiele und Perspektiven* (Bd. 13, S. 83–96). Weinheim: Juventa.
- Krüger, M., Won, M., & Treagust, D. F. (2013). Teachers' perceptions on the changes in the curriculum and exit examinations for biology and human biology. *Australian Journal of Teacher Education, 38*(3), 41–58.
- Kühn, S. M. (2010). *Steuerung und Innovation durch Abschlussprüfungen?* Wiesbaden: VS Verlag für Sozialwissenschaften.
- Kühn, S. M. (2011a). Und was kommt an? Analysen zur Implementation von Innovationen in der gymnasialen Oberstufe und im Abitur. *Die Deutsche Schule, 103*(1), 24–37.
- Kühn, S. M. (2011b). Weiterentwicklung der Aufgabenkultur im naturwissenschaftlichen Unterricht in der gymnasialen Oberstufe und im Abitur. *Zeitschrift für Didaktik der Naturwissenschaften, 17*, 35–55.
- Kühn, S. M. (2012). Exploring the use of statewide exit exams to spread innovation – The example of context in science tasks from an international comparative perspective. *Studies in Educational Evaluation, 37*(4), 189–195.
- Kühn, S. M. (2013). Vergleichbarkeit beim Mittleren Schulabschluss? Ein Überblick über die Vielfalt schulstrukturell möglicher Bildungswege und Prüfungsverfahren in den deutschen Ländern. *Die Deutsche Schule, 105*(1), 87–101.
- Kühn, S. M., & Drüke-Noe, C. (2013). Qualität und Vergleichbarkeit durch Bildungsstandards und zentrale Prüfungen? – Ein bundesweiter Vergleich von Prüfungsanforderungen im Fach Mathematik zum Erwerb des Mittleren Schulabschlusses. *Zeitschrift für Pädagogik, 59*(6), 912–932.
- Kühn, S. M., & Racherbäumer, K. (2013). Standardisierung und/oder Individualisierung? Empirische Befunde zur Umsetzung von Maßnahmen zur individuellen Förderung im Kontext zentraler Abschlussprüfungen. *Unterrichtswissenschaft, 41*(2), 172–189.

- Landesinstitut für Schule und Medien Berlin-Brandenburg. (2011). Evaluationsbericht. Schriftliche Prüfungsaufgaben im Zentralabitur 2010 des Landes Brandenburg. http://bildungsserver.berlin-brandenburg.de/fileadmin/bbb/unterricht/pruefungen/abitur_bb/Evaluation_ZABI_2010.pdf. Zugegriffen: 26.08.2013.
- Lorenz, R., Kahnert, J., Eickelmann, B., & Bos, W. (2011). Mehr Gerechtigkeit durch Zentralabitur? Analysen einer Lehrerbefragung in NRW. *Schul-Management*, 42(6), 24–27.
- Lorenz, R., Eickelmann, B., & Dohe, C. (2013). Geschlechtsspezifische Unterschiede im Fach Englisch in NRW – zur Frage der Fairness von zentralen Abituraufgaben. *Jahrbuch der Schulentwicklung*, 17, 236–263.
- Luyten, H., & Dolkar, D. (2010). School-based assessments in high-stakes examinations in Bhutan: A question of trust? Exploring inconsistencies between external exam scores, school-based assessments, detailed teacher ratings, and student self-ratings. *Educational Research and Evaluation*, 16(5), 421–435.
- Maag Merki, K. (2008). Die Einführung des Zentralabiturs in Bremen – Eine Fallanalyse. *Die Deutsche Schule*, 100(3), 357–368.
- Maag Merki, K. (2010). Theoretische und empirische Analysen der Effektivität von Bildungsstandards, standardbezogenen Lernstandserhebungen und zentralen Abschlussprüfungen. In H. Altrichter & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (S. 145–170). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Maag Merki, K. (2011). The introduction of state-wide exit examinations: Empirical effects on math and English teaching in German academically oriented secondary schools. In M. A. Pereyra, H.-G. Kotthoff & R. Cowen (Hrsg.), *Pisa Under Examination. Changing Knowledge, Changing Tests, and Changing Schools* (S. 125–141). Rotterdam: Sense.
- Maag Merki, K. (2012a). Die Leistungen der Gymnasiastinnen und Gymnasiasten in Mathematik und Englisch. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Prozesse und Wirkungen der Einführung zentraler Abiturprüfungen in zwei Bundesländern* (S. 263–292). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Maag Merki, K. (2012b). *Zentralabitur. Die längsschnittliche Analyse der Prozesse und Wirkungen der Einführung zentraler Abiturprüfungen in zwei Bundesländern*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Maag Merki, K. (2012c). Zentrale Prüfungen – empirische Evidenzen der Effekte der Einführung zentraler Abiturprüfungen auf Motivation und Emotion der Schüler/innen. In A. Wacker, U. Maier & J. Wissinger (Hrsg.), *Schul- und Unterrichtsreform durch ergebnisorientierte Steuerung. Empirische Befunde und forschungsmethodische Implikationen* (S. 247–276). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Maag Merki, K., & Holmeier, M. (2008). Die Implementation zentraler Abiturprüfungen: Erste Ergebnisse zu den Effekten der Einführung auf das schulische Handeln der Lehrpersonen. In E.-M. Lankes (Hrsg.), *Pädagogische Professionalität als Gegenstand empirischer Forschung* (S. 233–244). Münster: Waxmann.
- Maag Merki, K., Klieme, E., & Holmeier, M. (2008). Unterrichtsgestaltung unter den Bedingungen zentraler Abiturprüfungen: Differenzielle Analysen auf Schulebene mittels Latent Class Analysis. *Zeitschrift für Pädagogik*, 54(6), 791–808.
- Maag Merki, K., Holmeier, M., Jäger, D. J., & Oerke, B. (2010). Die Effekte der Einführung zentraler Abiturprüfungen auf die Unterrichtsgestaltung in Leistungskursen in der gymnasialen Oberstufe. *Unterrichtswissenschaft*, 38(2), 173–192.
- Maier, U., Bohl, T., Kleinknecht, M., & Metz, K. (2011). Einflüsse von Merkmalen des Testsystems und Schulkontextfaktoren auf die Akzeptanz und Rezeption von zentralen Testrückmeldungen durch Lehrkräfte. *Journal for Educational Research Online*, 3(2), 62–93.
- Martorell, F. (2004). *Do high school graduation exams matter? A regression discontinuity approach*. Working Paper, University of California, Berkeley. http://www.utdallas.edu/research/tsp-erc/pdf/wp_martorell_2004_high_school_graduation_exams.pdf. Zugegriffen: 22. Jan. 2014.

- Massell, D., Goertz, M. E., Christensen, G., & Goldwasser, M. (2005). The press from above, the pull from below: High school responses to external accountability. In B. Gross & M. E. Goertz (Hrsg.), *Holding high hopes. How high schools respond to state accountability policies* (S. 17–41). CPRE Research Report Series.
- Maué, E., Maag Merki, K., & Oerke, B. (2012). Emotionales Erleben des Zentralabiturs von Lehrpersonen in Bremen. Längerfristige Effekte der Implementation zentraler Prüfungen. In S. Hornberg & M. Parreira do Amaral (Hrsg.), *Deregulierung im Bildungswesen* (S. 109–130). Münster: Waxmann.
- Neumann, M., Trautwein, U., & Nagy, G. (2011). Do central examinations lead to greater grading comparability? A study of frame-of-reference effects on the university entrance qualification in Germany. *Studies in Educational Evaluation*, 37(4), 206–217.
- Neumann, M., Nagy, G., Trautwein, U., & Lüdtke, O. (2009). Vergleichbarkeit von Abiturleistungen. *Zeitschrift für Erziehungswissenschaft*, 12(4), 691–714.
- Odendahl, J. (2008). Zentral, nicht zerebral: Wider die Benotung zentral gestellter Klausuren anhand eines Bewertungsrasters. *Pädagogische Rundschau*, 62(4), 451–457.
- Oerke, B. (2012a). Emotionaler Umgang von Lehrkräften und Schüler/-innen mit dem Zentralabitur: Unsicherheit, Leistungsdruck und Leistungsattributionen. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Prozesse und Wirkungen der Einführung zentraler Abiturprüfungen in zwei Bundesländern* (S. 119–154). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Oerke, B. (2012b). Auseinandersetzung der Lehrpersonen mit der Einführung des Zentralabiturs: Stages of Concern. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Prozesse und Wirkungen der Einführung zentraler Abiturprüfungen in zwei Bundesländern* (S. 207–236). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Oerke, B., Maag Merki, K., Maué, E., & Jäger, D. (2013). Zentralabitur und Themenvarianz im Unterricht: Lohnt sich Teaching-to-the-test? In D. Bosse, F. Eberle & B. Schneider-Taylor (Hrsg.), *Standardisierung in der gymnasialen Oberstufe* (S. 27–49). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Ofqual. (2012). *International comparisons in senior secondary assessment. Full report*. Coventry: Office of Qualifications and Examinations Regulation.
- Otto, S., & Kühn, S. M. (im Erscheinen). Zwischen Tradition und Innovation. Eine Analyse zentraler Prüfungsaufgaben im Fach Deutsch im nationalen Vergleich. *Schulpädagogik-heute* 9(5).
- Perryman, J., Ball, S., Maguire, M., & Braun, A. (2011). Life in the pressure cooker – School league tables and English and mathematics teachers' responses to accountability in a results-driven era. *British Journal of Educational Studies*, 59(2), 179–195.
- Piopiunik, M., Schwerdt, G., & Woessmann, L. (2013). Central school exit exams and labor-market outcomes. *European Journal of Political Economy*, 31(9), 92–108.
- Putwain, D. W. (2008a). Do examinations stakes moderate the test anxiety–examination performance relationship? *Educational Psychology*, 28(2), 109–118.
- Putwain, D. W. (2008b). Test anxiety and GCSE performance: The effect of gender and socio-economic background. *Educational Psychology in Practice*, 24(4), 319–334.
- Putwain, D. W. (2009). Assessment and examination stress in key stage 4. *British Educational Research Journal*, 35(3), 391–411.
- Putwain, D. W. (2011). How is examination stress experienced by secondary students preparing for their general certificate of secondary education examinations and how can it be explained? *International Journal of Qualitative Studies in Education*, 24(6), 717–731.
- Putwain, D. W., Connors, L., & Symes, W. (2010). Do cognitive distortions mediate the test anxiety–examination performance relationship? *Educational Psychology*, 30(1), 11–26.
- Racherbäumer, K., & Kühn, S. M. (2013). Zentrale Prüfungen und individuelle Förderung. *Zeitschrift für Bildungsforschung*, 3(1), 27–45.

- Saunders, L. (2000). Understanding schools' use of 'Value Added' data: The psychology and sociology of numbers. *Research Papers in Education*, 15(3), 241–258.
- Saunders, L., & Rudd, P. (1999). *Schools' use of 'Value Added' Data: A science in the service of an art?* Slough: NFER.
- Scarth, J., & Hammersley, M. (1988). Examinations and teaching: An exploratory study. *British Educational Research Journal*, 14(3), 231–249.
- Schoppmeier, F., Borowski, A., & Fischer, H. E. (2012). Mathematische Bereiche in Leistungskursklausuren. *Physik und Didaktik in Schule und Hochschule*, 11(1), 28–40.
- Schräpler, J.-P., & Weishaupt, H. (2013). Auswirkung des Zentralabiturs auf den Abiturserfolg an Gymnasien und Gesamtschulen in Nordrhein-Westfalen. In N. McElvany & H. G. Holtappels (Hrsg.), *Empirische Bildungsforschung. Theorien, Methoden, Befunde und Perspektiven* (S. 249–266). Münster: Waxmann.
- Schümer, G., & Weiss, M. (2008). *Bildungsökonomie und Qualität der Schulbildung: Kommentar zur bildungsökonomischen Auswertung von Daten aus internationalen Schulleistungsstudien*. Frankfurt a. M.: GEW.
- Shuster, K. (2012). Re-examining exit exams: New findings from the educational longitudinal study of 2002. *Education Policy Analysis Archives*, 20(3), 1–32. <http://epaa.asu.edu/ojs/article/view/797>. Zugegriffen: 26. Aug. 2013.
- Sipple, J. W., Killeen, K., & Monk, D. H. (2004). Adoption and adaptation: School district responses to state imposed learning and graduation requirements. *Educational Evaluation and Policy Analysis*, 26(2), 143–168.
- Smith, L., & Sinclair, K. (2000). Stress and learning in the higher school certificate. *Change: Transformations in Education*, 3(2), 67–79.
- Somers, A. (1996). Examinations and educational quality. In A. Little & A. Wolf (Hrsg.), *Assessment in transition. learning, monitoring, and selection in international perspective* (S. 263–284). Oxford: Pergamon.
- Spillane, J. P. (2012). Data in practice: Conceptualizing the data-based decision-making phenomena. *American Journal of Education*, 118(2), 113–141.
- Spratt, M. (2005). Washback and the classroom: The implications for teaching and learning of studies of washback from exams. *Language Teaching Research*, 9(1), 5–29.
- Stevens, P. A. J., & van Houtte, M. (2011). Adapting to the system or the student? Exploring teacher adaptations to disadvantaged students in an English and a Belgian secondary school. *Educational Evaluation and Policy Analysis*, 33(1), 59–75.
- Stodolsky, S. S., & Grossman, P. L. (1995). The impact of subject matter on curricular activity: An analysis of five academic subjects. *American Educational Research Journal*, 32(2), 227–249.
- Sturm, R. (2008). Bemerkungen zum Zentralabitur in Niedersachsen im Fach Geschichte. *Geschichte in Wissenschaft und Unterricht*, 59(4), 218–230.
- Tikkanen, G. M. (2010). *Kemian ylioppilaskokeen tehtävät summatiivisen arvioinnin välineenä* [Gefälligkeitsübersetzung in Englisch: Chemistry matriculation examination as summative assessment tool]. Dissertation. Helsinki: Universität Helsinki.
- Tikkanen, G. M., & Aksela, M. (2012). Analysis of Finnish chemistry matriculation examination questions according to cognitive complexity. *NorDiNa*, 8(3), 258–268.
- Tymms, P. (1995). Influencing educational practice through performance indicators. *School Effectiveness and School Improvement*, 6(2), 123–145.
- Valverde, G. A. (2005). Curriculum policy seen through high-stakes examinations: Mathematics and biology in a selection of school-leaving examinations from the Middle East and North Africa. *Peabody Journal of Education*, 80(1), 29–55.
- Vogler, K. E. (2006). Impact of an exit examination on English teachers' instructional practices. *Essays in Education*, 16 (Spring). <http://www.usca.edu/essays/vol162006/vogler.pdf>. Zugegriffen: 26. Aug. 2013.

- Vogler, K. E. (2008). Comparing the Impact of Accountability Examinations on Mississippi and Tennessee Social Studies Teachers' Instructional Practices. *Educ. Assessment, 13*(1), 1–32.
- Vogler, K. E., & Carnes, G. N. (April 2009). *Comparing the Impact of a High School Exit Examination on Science Teachers' Instructional Practice. Paper presented at the annual meeting of the American Educational Research Association.* San Diego, California.
- Warren, J. R., & Kulick, R. B. (2007). Modeling States' Enactment of High School Exit Examination Policies. *Social Forces, 86*(1), 215–229.
- Wößmann, L. (2007). *Fundamental determinants of school efficiency and equity: German states as a microcosm for OECD countries* (CESifo Working Papers Nr. 1981).
- Wößmann, L. (2008). Zentrale Abschlussprüfungen und Schülerleistungen: Individualanalysen anhand von vier internationalen Tests. *Zeitschrift für Pädagogik, 54*(6), 810–827.
- Zabala, D., Minnici, A., McMurrer, J., & Briggs, L. (2008). *State high school exit exams: A move toward end-of-course exams.* Washington, D. C.: Center on Education Policy (CEP).
- Zabka, T., & Stark, T. (2010). Aufgabenstellungen und Erwartungshorizonte als Steuerungsinstrumente: Zum Umgang mit Problemen der Literaturinterpretation im Zentralabitur. *Der Deutschunterricht, 62*(1), 19–29.
- Zhang, Y. (2009). *Conflicts between state policy and school practice: Learning from Arizona's experience with high school exam policies.* Washington, D. C.: Center on Education Policy (CEP). <http://www.cep-dc.org/displayDocument.cfm?DocumentID=193>. Zugegriffen: 26. Aug. 2013.