

Die Beantwortung der übergeordneten Forschungsfragen setzt die Vergegenwärtigung des aktuellen empirischen Forschungsstandes zu verschiedenen gemeinsamen Prüfungsverfahren voraus, wie sie zur Standardisierung der Beurteilungspraxis in Schulen in unterschiedlichen Ländern eingesetzt werden. Ein Blick in die für diese Studie relevante, deutschsprachige Forschungsliteratur zeigt, dass, angesichts der Breite der möglichen Befunde zu gemeinsamen Prüfungsformen, drei Verfahren besonders interessant sind:

- a) *Orientierungs- oder Vergleichsarbeiten*: Dies sind Prüfungen, die im Verlauf des Schuljahres geschrieben werden und in der Regel zum Ziel haben, die Kompetenzen zu einem umfangreichen Themenbereich oder sogar zu einem gesamten Schuljahr zu erfassen. In Deutschland gibt es diesbezüglich die „Vergleichsarbeiten“ (VERA), in denen der Leistungsstand von Schüler/innen in den Fächern Deutsch und Mathematik in den Jahrgangsstufen 3 und 8 (hier zusätzlich Englisch und Französisch) erhoben wird. In einzelnen Kantonen in der Schweiz (z. B. Kantone im Bildungsraum Nordwestschweiz: Aargau, Basel-Landschaft, Basel-Stadt, Solothurn) werden Orientierungs- oder Vergleichsarbeiten als gemeinsame Prüfungen durchgeführt. Die Befunde dienen dabei in erster Linie der Diagnose der Kompetenzen der Schüler/innen. Teilweise werden sie aber auch zur Selektion herangezogen. Im Rahmen des Schweizer Bildungssystems können Orientierungs- und Vergleichsarbeiten sowohl kantonale als auch schulintern organisiert und durchgeführt sein.
- b) *Zentrale bzw. Gemeinsame Maturitäts- bzw. Abiturprüfungen*: Dabei handelt es sich um Prüfungen, die im Gegensatz zu Orientierungs- und Vergleichsarbeiten am Ende der allgemeinbildenden Schullaufbahn geschrieben werden. Die Art und Weise der Organisation und Durchführung dieser Abschlussprüfungen kann unterschiedlich gestaltet sein, wobei als eine Gemeinsamkeit gilt, dass sie

nicht – wie in vielen Schweizer Kantonen noch üblich – individuell von einer Lehrperson für die eigene Klasse entwickelt werden. Vielmehr lassen sich zwei Varianten unterscheiden, einerseits Abschlussprüfungen, die von schulexternen Fachkommissionen erstellt werden (‘zentrale Abschlussprüfungen’) und andererseits Abschlussprüfungen, die schulintern entwickelt werden (‘gemeinsame Abschlussprüfungen’). Zu Erstgenanntem zählen Abschlussprüfungen, die einheitlich für alle Gymnasien in einem Bundesland (z. B. Deutschland: Zentralabitur) oder sogar bundeslandübergreifend (z. B. Österreich: Zentralmatura) erstellt werden. Bei schulintern erstellten, gemeinsamen Abschlussprüfungen konzipiert ein Zusammenschluss von mehreren Lehrpersonen für mehrere Klassen innerhalb eines Gymnasiums die Prüfung (Schweiz: ‘Hausmatura’, ‘harmonisierte Maturitätsprüfung’). Der Grad der Partizipations- und Gestaltungsmöglichkeiten variiert erheblich zwischen den Prüfungsformen. So bieten schulextern organisierte und durchgeführte Abschlussprüfungen den Lehrpersonen – bis auf die Korrektur – kaum Gelegenheit zur Partizipation bzw. Mitgestaltung, während schulintern konzipierte gemeinsame Abschlussprüfungen sowohl bei der Entwicklung der Aufgaben als auch bei der Korrektur Partizipationsmöglichkeiten eröffnen. Gemeinsame und zentrale Maturitäts- bzw. Abschlussprüfungen dienen der Selektion von Schüler/innen und erhalten dadurch eine große Bedeutung für deren Laufbahn (‘high-stake’). Im Gegensatz zu entsprechenden Verfahren in anderen Ländern dienen die im deutschsprachigen Raum implementierten, zentralen Abitur- und Maturitätsprüfungen jedoch nicht der Beurteilung der Schulen oder Lehrpersonen und sind für diese somit ‘low-stake’.

- c) *In einzelnen Fachschaften oder Teilgruppen von Lehrpersonen im Gymnasium durchgeführte gemeinsamen Prüfungen:* Dies sind Prüfungen, die von den Fachschaften oder Teilgruppen von Lehrpersonen entwickelt werden. Sie können mündlich oder schriftlich sowie einmal oder mehrmals pro Jahr durchgeführt werden. Zudem kann der Stoffumgang variieren (Stoff eines ganzen Schuljahres oder eines einzelnen Themenbereichs). Diese Prüfungen weisen gewisse Parallelen zu Verfahren innerhalb eines schulinternen Evaluationsprozesses auf, da der gesamte Prozess von den Lehrpersonen durchgeführt wird, d. h. von der Entwicklung der Aufgaben über die konkrete Durchführung bis hin zur Auswertung. Besonders auffallend sind der fachschaftsinterne Entwicklungsprozess und das Finden einer für die einzelnen Akteure zufriedenstellenden Form des ‘Gemeinsamen Prüfens’. Die Partizipations- und Gestaltungsmöglichkeiten der Lehrpersonen sind hierbei groß.

Der nachfolgende Überblick über den empirischen Forschungsstand zu den oben genannten drei Typen zentraler bzw. gemeinsamer Prüfungen berücksichtigt insbesondere folgende Aspekte:

- Inwiefern führt die Durchführung gemeinsamer Prüfungen zu einer stärkeren Standardisierung in der Beurteilung der Leistungen der Schüler/innen?
- Inwiefern gehen mit der Durchführung gemeinsamer Prüfungen negative transintentionale Veränderungen im Unterricht, bei den Lehrpersonen und/oder den Schüler/innen einher (z. B. teaching-to-the-test, stärkere Unsicherheit bei Schüler/innen und Lehrpersonen)?
- Wie gehen Schulen und Lehrpersonen mit Datenrückmeldungen um und inwiefern führt die Durchführung gemeinsamer Prüfungen zur Weiterentwicklung des Unterrichts und/oder der Schule?

Bevor diese Themen bearbeitet werden, ergibt sich allerdings die Notwendigkeit der Klärung des Begriffes ‚gemeinsam‘.

2.1 Begriffliche Klärung: ‚gemeinsam‘ vs. ‚zentral‘

Wie im Bericht von Mero (2013) dargelegt, entspricht der aktuell genutzte Begriff des ‚Gemeinsamen Prüfens‘ nur teilweise dem, was in der Literatur in der Regel unter dem Begriff des ‚zentralen Prüfens‘ (oder ‚zentrale Prüfungen‘) subsumiert wird. Vielmehr ergeben sich bei genauer Betrachtung sowohl graduelle als auch inhaltliche Unterschiede.

Von *graduellen Unterschieden* kann beispielsweise hinsichtlich des Ausmaßes an zugrunde liegendem Konsens über die Durchführung der Prüfungen gesprochen werden. So besteht der Konsens beim ‚Gemeinsamen Prüfen‘ innerhalb einer größeren Lehrpersonengruppe bzw. Fachschaft, während er beim ‚zentralen Prüfen‘ über Zielkriterien schulübergreifender oder bundeslandweiter Lehrpläne definiert wird. Auch bezüglich der Vergleichbarkeit ergeben sich graduelle Unterschiede. Während beim ‚Gemeinsamen Prüfen‘ zwar eine gegenüber den traditionellen, d. h. nicht gemeinsam durchgeführten, Klassenprüfungen verbesserte Vergleichbarkeit zu erwarten ist, kann die Vergleichbarkeit als ein wesentliches Grundmerkmal des ‚zentralen Prüfens‘ angesehen werden (gleiche Leistung wird mit gleichen Noten bewertet).

Inhaltliche Unterschiede ergeben sich insbesondere aus der Relevanz der verschiedenen Bezugsnormen bei der Beurteilung der Leistungen der Schüler/innen. So werden beim ‚Gemeinsamen Prüfen‘ gemäß Mero (2013) alle drei Bezugsnormen (individuell, kriterial, sozial) genutzt, wobei im Vergleich zu traditionellen Klassenarbeiten die soziale Bezugsnorm in gemeinsamen Prüfungen einen erweiterten, da klassenübergreifenden Bezugsrahmen, aufweist. In zentralen Prüfungen, insbesondere wenn darunter Prüfungen oder Tests wie in den PISA- oder

TIMS-Studien¹ oder in den Vergleichsarbeiten in Deutschland (VERA) verstanden werden, kommt nur die kriteriale Bezugsnorm zum Einsatz, ohne eine Berücksichtigung der sozialen und individuellen Bezugsnorm.

Wesentlich bei diesen an der kriterialen Norm orientierten zentralen Prüfungen ist, dass sowohl die Entwicklung der Prüfungsaufgaben als auch die Bewertung der Leistungen der Schüler/innen nicht durch die Lehrpersonen der jeweiligen Klasse, sondern in der Regel durch schulexterne wissenschaftliche Institutionen erfolgt.

Werden die in Deutschland implementierten ‚Zentralen Abiturprüfungen‘ betrachtet, so bedeutet ‚zentral‘ wiederum etwas anders. Hier sind insbesondere der Entwicklungsprozess und die Durchführung der Abiturprüfungen ‚zentral‘ organisiert: Pro Bundesland (oder teilweise bundeslandübergreifend) entwickeln behördlich eingesetzte Fachkommissionen die Abiturprüfungen inkl. eines mehr oder weniger detaillierten Korrektur- und Beurteilungsrasters. Die Prüfungen werden pro Fach bzw. Fachkurs im ganzen Bundesland (oder dann bundeslandübergreifend) zum selben Zeitpunkt durchgeführt (Klein et al. 2009; Kühn 2012). In Abgrenzung zu den oben beschriebenen, im Rahmen von national oder international vergleichenden Untersuchungen durchgeführten, zentralen Prüfungen (z. B. PISA) findet aber in vielen Bundesländern die Korrektur und die Beurteilung der Leistungen der Schüler/innen durch die einzelnen Lehrpersonen, d. h. ‚dezentral‘ statt; dies allerdings unter Berücksichtigung der von der schulexternen Fachkommission festgelegten Korrektur- und Beurteilungsraster. Diese weisen Spielraum für die Beurteilung von Einzelschritten, Teilergebnissen oder Endergebnissen auf, sodass die Lehrperson die Beurteilungen in Abhängigkeit der Leistungen der Schüler/innen ihrer Klasse oder der Unterrichtsvorbereitung vornehmen oder Ergebnisse milder oder strenger bewerten kann. ‚Zentrale Abiturprüfungen‘ in Deutschland sind somit Prüfungen, die nicht von der einzelnen Lehrperson für die eigene Klasse, sondern schulübergreifend und schulextern in einem Bundesland von einer zentral eingesetzten Kommission erstellt und von allen Schüler/innen eines bestimmten Jahrgangs/Faches zum gleichen Zeitpunkt geschrieben werden. Sie entsprechen aber weder den zentralen Prüfungen, die ein höheres Ausmaß an Standardisierung aufweisen (wie beispielsweise die PISA- oder TIMSS-Tests), noch der Prüfungsform der Hausmatura, die in einzelnen Kantonen der Schweiz eingeführt worden ist.

Bei der Hausmatura handelt es sich um gemeinsame Prüfungen mit einem hohen Harmonisierungsgrad (Mero 2013, S. 9). Der Partizipationsgrad der Lehrpersonen im Bereich der Korrektur ist dabei vergleichbar mit demjenigen des Zen-

¹PISA = Programme for International Student Assessment; TIMSS = Trends in International Mathematics and Science Study.

tralabiturs. Bezüglich der Entwicklung und Abstimmung der Prüfungsaufgaben können sich die Lehrpersonen bei den harmonisierten Maturitätsprüfungen allerdings stärker einbringen als beim Zentralabitur. So werden beispielsweise im Kanton Aargau „Inhalte, Kompetenzen und Anforderungsniveau [...] schulhausintern festgelegt und zudem kantonal – mittels zentraler Rahmenvorgaben und im Dialog mit kantonal bestimmten Fachschaftsdelegierten – abgeglichen“ (Mero 2013, S. 9). Bei den zentralen Abiturprüfungen in Deutschland ist dies nicht möglich.

Zusammenfassend kann somit festgehalten werden, dass die Verwendung des Begriffes ‚gemeinsam‘, wie er für das ‚Gemeinsame Prüfen‘ in der Schweiz genutzt wird, spezifische Eigenheiten aufweist, die mit alternativen Begrifflichkeiten nicht vollständig abgedeckt werden können. Im Rahmen der hier präsentierten Studie wird deshalb – im Sinne einer Arbeitsdefinition – nur dann vom ‚Gemeinsamen Prüfen‘ bzw. von ‚gemeinsamen Prüfungen‘ gesprochen, wenn Prüfungen gemäß der Definition von Mero (2013) fokussiert werden. Der Begriff ‚zentrale Prüfungen‘ soll sonstige gemeinsam durchgeführte Prüfungsformen kennzeichnen, die die oben dargestellten, graduellen bzw. inhaltlichen Unterschiede zu den für die Schweiz spezifischen Formen des ‚Gemeinsamen Prüfens‘ aufweisen.

2.2 Stärkere Standardisierung der Beurteilungen durch ‚zentrale Prüfungen‘?

Inwiefern kann mit zentralen Prüfungen tatsächlich eine klassen- und schulübergreifende größere Vergleichbarkeit der Beurteilungspraxis erreicht werden? Diese Frage ist in verschiedenen Studien untersucht worden. Die empirische Befundlage, wie sie sich zum heutigen Zeitpunkt präsentiert, ist allerdings nicht eindeutig (Klein et al. 2014; Maag Merki 2016a, b). Zwar gibt es Studien, die in der Tendenz auf eine stärkere Standardisierung der Beurteilungspraxis hinweisen, allfällige Effekte werden jedoch lediglich in einzelnen Fächern oder Kursen sichtbar. Empirische Befunde zur Frage, inwiefern es gelingt, durch die Einführung zentraler Abiturprüfungen die bisherige – tendenziell vor allem an der sozialen Bezugsnorm orientierte – Beurteilung stärker an einer klassen- und schulübergreifenden Norm auszurichten, liegen nur vereinzelt vor (Holmeier 2013). Es dominieren Querschnittanalysen, in denen ‚stärker zentrale‘ mit ‚weniger zentralen‘ Beurteilungssystemen zu einem bestimmten Zeitpunkt miteinander verglichen werden und dies lediglich in wenigen Fächern. Was jedoch geschieht, wenn zentrale Prüfungen in Musik, Kunst oder sozialwissenschaftlichen Fächern eingeführt werden, ist bislang nicht analysiert worden.

Für die vorliegende Untersuchung besonders relevante Befunde bietet die TIMS-Studie für die Sekundarstufe II. Im Rahmen dieser Erhebung untersuchten Baumert und Watermann (2000) für den deutschsprachigen Raum den Effekt zentraler Abiturprüfungen auf die Benotungen im Abitur. Sie nutzten hierzu die regionalen Unterschiede zwischen den deutschen Bundesländern, da zum Zeitpunkt der Untersuchung nur einzelne Bundesländer in einzelnen Fächern die Abiturprüfungen zentral durchführten. Als Ergebnis zeigten sich kursniveau²- und fachspezifische Effekte. So wurde in Ländern mit zentralen Abiturprüfungen in den Grundkursen Mathematik strenger benotet, nicht aber in den Leistungskursen Physik und Mathematik.

Ebenfalls auf Basis der TIMSS-Daten für die Sekundarstufe II führten Neumann et al. (2011) Auswertungen zur Frage der Benotungsgerechtigkeit durch, allerdings nur im Bundesland Baden-Württemberg. In diesen Analysen stand die Frage im Zentrum, ob der Effekt der mittleren Klassenleistung auf die Benotung in zentralen Abschlussprüfungen geringer ist als auf die Halbjahresnoten im gymnasialen Unterricht. Dies geschah unter der Annahme, dass ein geringerer Effekt der mittleren Klassenleistung darauf hindeuten würde, dass die Beurteilungen der individuellen Leistungen der Schüler/innen weniger stark von der Klassenzusammensetzung abhängig und damit klassenübergreifend vergleichbarer sind. Die Analysen wurden für Mathematik und Englisch durchgeführt. Sie zeigten für Mathematik erwartungskonforme Befunde. Bezüglich der Klassennoten im Unterricht ergaben sich sowohl für die Grund- wie auch für die Leistungskurse signifikante Bezugsgruppeneffekte, d. h. gleiche Leistungen von Schüler/innen in unterschiedlichen Klassen wurden unterschiedlich beurteilt, je nachdem, wie leistungsstark eine Klasse war. Bei den Abiturprüfungen zeigte sich ein solcher Effekt in den Leistungskursen Mathematik allerdings nicht mehr und in den Grundkursen war er weniger stark. Die Autoren schlossen aus den Ergebnissen, dass die Beurteilungen im zentral geprüften Abitur nicht bzw. weniger von der durchschnittlichen Klassenleistung abhängig waren.

In Englisch zeigten sich für die Leistungskurse keine Bezugsgruppeneffekte, weder bei den Beurteilungen im Unterricht noch in Abiturbeurteilungen. In den Grundkursen hingegen stimmten die Ergebnisse mit jenen in den Mathematik-Leistungskursen überein: Die mittlere Klassenleistung beeinflusste dort die Halbjahresnoten, für die

²Das deutsche Kurssystem unterscheidet zwischen Grund- und Leistungskursen. Grundkurse dienen der Vermittlung grundlegender Inhalte und Kompetenzen, während Leistungskurse vertiefte Kenntnisse und Fähigkeiten innerhalb eines Faches vermitteln sollen. Leistungskurse umfassen zudem mehr Unterrichtsstunden pro Woche als Grundkurse.

Abiturnoten zeigte sich kein Effekt (Neumann et al. 2011, S. 214). Der Standardisierungseffekt ist somit nicht eindeutig, sondern analog zur Studie von Baumert und Watermann (2000) als abhängig vom Fach und vom Kursniveau zu bewerten.

Neumann et al. (2009) führten im Rahmen der TOSCA³ und der LAU-Studie⁴ ebenfalls eine Analyse zur Standardisierung der Benotung in der Sekundarstufe II durch. Dazu wurden Leistungsdaten in Mathematik und Englisch sowie die Fachnoten im letzten Jahr des Gymnasiums und im Abitur genutzt. Im Zentrum stand die Frage, inwiefern sich Leistungsunterschiede in den in Mathematik und Englisch erteilten Fachnoten des letzten Schuljahres in den Abiturnoten widerspiegeln. Dabei verglichen sie die Daten aus dem Bundesland Hamburg (dezentral geprüfte Abiturleistungen) mit jenen aus dem Bundesland Baden-Württemberg (zentral geprüfte Abiturleistungen). Im Fokus standen jene Schüler/innen, die die entsprechenden Fächer als Prüfungsfächer im Leistungskurs gewählt hatten. Die Ergebnisse wiesen ebenfalls auf einen fachspezifischen Effekt hin. Während die Befunde für Englisch zeigten, dass die Fach- und Abiturnoten die unterschiedlichen Leistungskompetenzen relativ gut abbilden konnten, ergaben sich für Mathematik bundeslandspezifische Effekte. Trotz gleicher Fachnoten in den beiden Bundesländern erhielten Hamburger Abiturient/innen im Mittel schlechtere Prüfungsnoten im Abitur als die Absolvent/innen in Baden-Württemberg. Diese Unterschiede resultierten jedoch nicht aus der milderer Bewertung der Schüler/innen in Baden-Württemberg, sondern aus ihren durchgehend besseren Leistungen gegenüber den Schüler/innen in Hamburg. Die deutlichen Leistungsunterschiede im Abitur spiegelten sich somit nicht in den Fachnoten im Gymnasium wider. Gemäß den Autoren lagen insofern zumindest für den Leistungskurs Mathematik deutliche Hinweise vor, als dass „Noten aus zentralen Abschlussprüfungen unter dem Aspekt der Verteilungsgerechtigkeit ein objektiveres Auswahlkriterium darstellen als referenzgruppenbezogene Fachnoten“ (ebd., S. 711). Allerdings wurde betont, dass die Befundlage keine Generalisierung erlaube, da einzig zwei Bundesländer sowie zwei Fächer untersucht wurden. Zudem war zu berücksichtigen, dass es sich bei dieser Untersuchung um eine Querschnittstudie handelte und somit nicht erfasst werden konnte, inwiefern sich mit der Einführung zentraler Prüfungen die Beurteilungsqualität verändert hat.

Dies war – in der Tendenz – in einer mehrjährigen, in den Bundesländern Bremen und Hessen durchgeführten Studie möglich, bei der Veränderungen in der schulischen Praxis aufgrund der 2007 durchgeführten Implementation zentraler Abiturprüfungen analysiert wurden (Maag Merki 2016a). Aufgrund ihres Lang-

³TOSCA = Transformation des Sekundarschulsystems und akademische Karrieren;

⁴LAU = Untersuchung Aspekte der Lernausgangslage und der Lernentwicklung

zeitcharakters konnten bei dieser Erhebung Auswirkungen des Wechsels von einem dezentralen zu einem zentralen Abiturprüfungssystem untersucht werden. Die Ergebnisse verwiesen teilweise auf einen Standardisierungseffekt bei der Beurteilungspraxis im Abitur und eine stärkere Standardisierung der Beurteilungen in den Halbjahresbeurteilungen im Gymnasium, dies allerdings in Abhängigkeit des Faches und des Kurstyps. Während der Effekt tendenziell für die Mathematik-Grundkurse in den schriftlichen Abiturprüfungen wie auch für die Halbjahresbeurteilungen in der gymnasialen Oberstufe zu beobachten war (Holmeier 2013; Maag Merki und Holmeier 2015; Maué 2013b), wurde bei den Mathematik-Leistungskursen nur die Halbjahresbeurteilung, nicht aber die Beurteilung des schriftlichen Abiturs vergleichbarer. Zudem gab es im Leistungskurs Mathematik Hinweise auf eine stärkere schul- bzw. kursübergreifende Standardisierung der Beurteilungspraxis im schriftlichen zentralen Abitur. Obwohl somit mit der Einführung zentraler Abiturprüfungen nicht der vor dem Abitur durchgeführte Unterricht im Zentrum stand, gab es Hinweise, dass in längerfristiger Perspektive die Einführung zentraler Abiturprüfungen auch mit Veränderungen der Beurteilungspraxis im Unterricht einherging. Dies könnte damit erklärt werden, dass Abschlussprüfungen einen Transfereffekt auf den davor gelagerten Unterricht ausüben können. Auch die Angleichung der Unterrichts- und Beurteilungspraxis an die Inhalte und Formate der Abschlussprüfungen kommt als mögliche Erklärung in Betracht, wie dies in der Literatur unter dem Begriff des „Washback“-Effekts gut dokumentiert ist (Cheng et al. 2004; Stecher 2002). Dieser Effekt beschreibt Anpassungen der Unterrichtspraxis an die Abschlussprüfungen, so beispielsweise eine stärkere Berücksichtigung der Prüfungsinhalte im Unterricht bei gleichzeitiger Reduktion von Inhalten, die nicht in der Prüfung geprüft werden oder eine stärkere Nutzung von Prüfungsformaten, wie sie in den Prüfungen vorzufinden sind (siehe Abschn. 2.3).

Innerhalb der Klassen haben die Lehrpersonen somit auch im dezentralen Abitur schon relativ fair beurteilt. Hingegen gelang es den Lehrpersonen unter dezentralen Prüfungsbedingungen weniger gut, das Leistungsniveau ihrer Schüler/innen schul- und kursübergreifend vergleichbar einzuschätzen (Maag Merki und Holmeier 2015). Aus der Perspektive der Chancengleichheit ist dies als problematisch zu beurteilen (Baumert et al. 2003). Mit der Einführung zentraler Abiturprüfungen scheint die vergleichende Einschätzung in der Tendenz eher zu gelingen. Allerdings zeigen weiterführende Analysen, dass die Beurteilungspraxis auch unter Bedingungen zentraler Abschlussprüfungen von nicht-leistungsbezogenen Kriterien, wie dem Geschlecht oder dem Migrationshintergrund beeinflusst wird, und dass die Einführung zentraler Abiturprüfungen zu keiner Veränderung der

Problemlage beitragen konnte (Holmeier 2013; Maué 2013a, 2013b). Wesentlicher Hintergrund ist hier die Tatsache, dass auch bei zentralen Abiturprüfungen die Leistungen der Schüler/innen nach wie vor von ihren Lehrpersonen beurteilt und benotet werden, allerdings auf der Basis eines kurs- und schulübergreifenden Beurteilungs- und Benotungsrasters. Diese dezentrale Beurteilungspraxis, die im Rahmen zentraler Abiturprüfungen häufig vorgenommen wird, weist zwar einzelne Vorteile für die Schulpraxis auf, ist aber anscheinend zu wenig in der Lage, einzelne Beurteilungsdifferenzen der Lehrpersonen aufzuheben (Beutel und Vollstädt 2000; Sacher 2009).

Ist nun aufgrund der Einführung und Durchführung des ‚Gemeinsamen Prüfens‘ von einem Standardisierungseffekt auf die Beurteilungspraxis der Lehrpersonen und von einer klassen- und schulübergreifend faireren Beurteilung auszugehen? Bisherige Befunde legen diesbezüglich keine generellen, sondern wenn, dann kurs- und fachspezifische Veränderungen nahe. Zudem kann nur dann von einer stärkeren Vergleichbarkeit der Beurteilungen ausgegangen werden, wenn zusätzlich weitere Aspekte berücksichtigt werden.

So sind schulübergreifend vergleichbarere Beurteilungen in ‚gemeinsamen Prüfungen‘ nur möglich, wenn tatsächlich schulübergreifend gleiche Prüfungen durchgeführt und mit dem gleichen Benotungsraster beurteilt werden. Dies ist in der Tendenz jedoch nur für jene Verfahren zu erwarten, die nicht in der einzelnen Schule, sondern schulübergreifend durchgeführt werden, wie beispielsweise bei den Vergleichs- oder Orientierungsarbeiten. Bei schulintern durchgeführten, harmonisierten Maturitätsprüfungen sind allfällige Standardisierungen in der Beurteilung auch nur innerhalb der jeweiligen Schule zu erwarten. Eine zusätzliche Möglichkeit der schulübergreifenden Angleichung der Anforderungsniveaus von harmonisierten Maturitätsprüfungen bietet sich jedoch insofern, als dass die schulspezifischen Prüfungsaufgaben von einer schulübergreifenden Fachkommission begutachtet werden. Dadurch könnte und sollte es über mehrere Jahre hinweg innerhalb der Schulen – sowie langfristig gesehen auch schulübergreifend – gelingen, auf Basis gemeinsamer Vorstellungen darüber, welche (Teil-)Leistungen mit welchen Punkten/Noten beurteilt werden sollen, eine vergleichbare Beurteilung zu entwickeln.

Eine klassen- und schulübergreifende Beurteilung und Benotung bedingt jedoch nicht nur eine Festlegung gemeinsamer Prüfungsaufgaben, sondern ebenfalls die Festlegung eines gemeinsamen Korrektur- und Benotungsmodus. Dafür bedarf es zunächst eines Aushandlungsprozesses zwischen den Lehrpersonen, bei dem geklärt werden muss, welche Prüfungsinhalte, Prüfungsformate und Anforderungsniveaus sowie Kriterien für die Beurteilung und Benotung notwendig sind. Die Lehrpersonen sollten darüber hinaus in der Lage sein, die Voraussetzungen der

Schüler/innen in den Klassen und die entsprechende Vorbereitung im Unterricht angemessen zu berücksichtigen. Gemeinsame, für wenige Klassen durchgeführte Prüfungen, scheinen in dieser Hinsicht einen Vorteil zu haben, da sie die situationsspezifischen Voraussetzungen der Schule und der Klassen stärker berücksichtigen können.

Führen ‚gemeinsame Prüfungen‘ auch zu einer höheren Vergleichbarkeit der Leistungen und Beurteilungen, da sie klassen- oder schulübergreifend durchgeführt werden? Hierzu liegen kaum Ergebnisse vor. Die Frage ist allerdings kritisch zu diskutieren: möglicherweise werden zwar die gleichen Ergebnisse anhand der gleichen Kriterien mit den gleichen Noten beurteilt, aufgrund der unterschiedlichen Vorbereitung in den einzelnen Klassen und den unterschiedlichen Kompetenzen der Lehrperson sind aber nicht alle Schüler/innen gleich gut auf die Prüfungen vorbereitet worden. ‚Fairness‘ muss somit als Zielkategorie mit Vorsicht betrachtet werden.

2.3 Transintentionale Veränderungen durch ‚zentrale Prüfungen‘

Eine große Anzahl an Studien zeigt, dass im Zuge der Einführung zentraler Prüfungen mit transintentionalen Effekten zu rechnen ist, d. h. mit Veränderungen, die nicht in Einklang mit den ursprünglich intendierten Veränderungen stehen (für einen Überblick siehe Klein et al. 2014; Maag Merki 2016b). Dabei ist zu berücksichtigen, dass diese transintentionalen Veränderungen nicht nur negativ, sondern auch positiv sein können, beispielsweise wenn im Zusammenhang mit einem bestimmten Reformprojekt die Arbeitszufriedenheit der Lehrpersonen steigt und die entsprechende Einführung zusätzlich den Austausch zwischen den Lehrpersonen fördert. Meist werden aber unter ‚transintentionalen‘ Veränderungen vor allem jene verstanden, die für das Lernen der Schüler/innen (z. B. größere Leistungsangst), für das Arbeiten der Lehrpersonen (z. B. geringere Unterrichtsqualität) oder für die Gesellschaft (z. B. stärkere Segregationseffekte zwischen verschiedenen Bevölkerungsgruppen) als problematisch angesehen werden. Daher erscheint es sinnvoll, solche transintentionalen, d. h. indirekt intendierten Nebeneffekte, mit in die Analyse aufzunehmen.

Bisherige Befunde zu möglichen Veränderungen aufgrund der Einführung zentraler Prüfungen werden nachfolgend unter Rückgriff auf eine Zusammenstellung von Maag Merki (2016b) überblicksartig dargestellt (vgl. Tab. 2.1). Die Effekte zeigen sich insbesondere in Verfahren, deren Ergebnisse zentral für die Weiterbeschäftigung von Lehrpersonen und Schulleitungen, für die zukünftigen finanziellen Mittel der Schule oder für die Promotion von Schüler/innen (‚high-stakes‘-Verfahren) sind.

Tab. 2.1 Effekte der Einführung zentraler Prüfungen im Kontext von ‚high stakes‘-Assessment. (Maag Merki 2016b)

Positive Effekte	Ambivalente oder negative Effekte
Lehrpersonen	
• Zunahme der Motivation von LP • LP arbeiten effektiver • Nutzung der Testergebnisse für verbesserte Unterstützung des Lernens der SuS	• Steigerung des Leistungsdrucks • Reduktion von moralischem Verhalten • Wechsel in ein Bildungssystem ohne ‚high-stakes‘ Testsysteme • Deprofessionalisierung
Unterricht	
• Stärkere Verwendung anspruchsvoller Unterrichtsmethoden • Stärkere Unterstützung von SuS • Stärkere Fokussierung auf die Lernergebnisse der SuS • Integration von zentralen Inhalten in das eigene Curriculum • Höhere Konsistenz zwischen Lehrinhalten und Curriculum • Testsystem als Basis für Entscheidungen und Förderansätze	• <i>Reallocation</i>: Einengung des Curriculums in Abhängigkeit der Testinhalte und der Stärke der Konsequenzen (Vernachlässigung von nicht getesteten Fächern, Vernachlässigung von Themen innerhalb von Fächern, die nicht getestet werden) • <i>Aligning</i>: Anpassung der Lehrmethoden und Prüfungsmethoden an Testformate • <i>Negative Coaching</i>: umfangreiche Testvorbereitungstrainings, die eine inhaltliche Auseinandersetzung mit Unterrichtsinhalten verhindern • <i>Cheating</i>: z. B. Testfragen im Voraus zugänglich machen, Hinweise/Unterstützung während der Testdurchführung geben • <i>Leistungsabhängige Förderung</i>: Besondere Förderung für SuS mit Leistungen nahe des minimal erforderlichen Leistungsniveaus (‚bubble kids‘), weniger Förderung von SuS deutlich unter oder über Leistungslimit

(Fortsetzung)

Tab. 2.1 (Fortsetzung)

Positive Effekte	Ambivalente oder negative Effekte
Schule	
• Stärkere Professionalisierungsmaßnahmen für LP • Zunahme von Kooperation zwischen LP • Zunahme der Bedeutung von Qualität für Schulentwicklung • Durchführung weiterer Evaluationen für Analyse des Entwicklungsfortschrittes • <i>Reallocation of resources</i>: Stärkere Förderung von SuS mit Defiziten; zusätzliche ‚after-school‘-Programme; mehr/besseres Lehrpersonal	• Neuzuteilung von LP zu den geprüften Klassenjahrgängen mit dem Ziel der Verbesserung der Prüfungsleistungen der Klassen • Verteilung von Belohnungen für SuS, damit sich diese im Test anstrengen • Ausschluss von leistungsschwachen SuS; Zurückbehaltung von SuS mit schlechten Leistungen in Klassenstufen, die nicht getestet werden; Zulassen von Absenzen an Testtagen • Häufigere Diagnose von Sonderbeschulung • Vergrößerung des Anteils von SuS, die die Schule ohne Abschluss verlassen • Nicht-Versetzen in die nächste Klasse • Reduktion oder Streichung von Fächern, die nicht getestet werden • Unterstützung für LP durch SL, Unterricht auf Tests zu fokussieren
Staat/Gesellschaft	
• Definition von zentralen Inhalten/Curriculum • Aufbau von technischen Unterstützungssystemen mit dem Ziel einer höheren Konsistenz zwischen schulischen Prozessen und Testsystem • Bereitstellung von Unterstützungsmaßnahmen und -materialien für Schulen und Distrikts (Bezirke/Gemeinden) im Umgang mit Testergebnissen und für Förderung von SuS • Vergrößerung des Lehrumfangs für leistungsschwache SuS	• Entscheidung über Zulassung von Lehrmitteln auf Ebene des Staates und nicht auf Ebene Distrikts (Bezirk/Gemeinde) oder Schule • Manipulation von Ergebnissen • Willkürliche Bestimmung des ‚cut score‘ • Validitätsprobleme bei Testkonstruktion, Fehler in den Tests • Keine oder negative Effekte auf Gerechtigkeit des Schulsystems • Veränderung des Bildes über die Profession der Lehrperson in der Gesellschaft

SuS = Schüler/innen, LP = Lehrpersonen, SL = Schulleitung

Unterschieden wird dabei zwischen positiven Effekten, so beispielsweise der stärkeren Unterstützung der Schüler/innen, und negativen oder zumindest ambivalenten Veränderungen, wie zum Beispiel einer Einengung des Curriculums auf Testinhalte

und -formate der Abschlussprüfungen. Entsprechende Effekte können auf allen Ebenen des Bildungssystems, von der individuellen bis zur Makroebene (Gesamtsystem), beobachtet werden.

Aufgrund der Abhängigkeit dieser Effekte von den Rahmenbedingungen der zentralen Prüfungen und ihren Durchführungsmodalitäten ist nun zu untersuchen, inwiefern auch in weniger zentral organisierten Systemen wie Deutschland oder der Schweiz entsprechende Veränderungen sichtbar werden. Die vorliegenden Befunde zeigen, dass auch dort mit ambivalenten oder negativen Veränderungen zu rechnen ist, obwohl diese im Vergleich zu anderen OECD-Staaten nur einen geringen Standardisierungsgrad aufweisen (Klein et al. 2009).

Unterschiede zwischen den Abiturprüfungssystemen sind zunächst *in der Unterrichtspraxis* zu identifizieren. So lässt sich gemäß van Ackeren et al. (2012) in Systemen mit zentralen Abiturprüfungen eine intensivere Vorbereitung der abiturrelevanten Unterrichtsinhalten im Vergleich zu ebenfalls lehrplanmäßig eingeforderten, aber nicht im Abitur geprüften Inhalten beobachten. Zudem stellten die Autor/innen in einem Bundesland mit zentralen Abiturprüfungen im Vergleich zu einem Bundesland ohne zentrale Abiturprüfungen ein höheres Unterrichtstempo und ein häufigeres Vergeben von Hausaufgaben fest. Auch Jürges und Schneider (2010) konstatierten auf Grundlage ihrer Daten, dass die Menge an Hausaufgaben bei zentralen Abschlussprüfungen größer sei als in dezentralen Prüfungssystemen und dass diese zudem im Unterricht häufiger kontrolliert und diskutiert werden.

Die Analysen in Bremen und Hessen bestätigen, dass aufgrund der Einführung zentraler Abiturprüfungen die Abiturschwerpunktthemen im Unterricht deutlich favorisiert werden, wodurch die Interessen der Schüler/innen, weitere curricular relevante Themen oder fachbezogene Alltagsaktualitäten weniger Berücksichtigung finden als in Prüfungssystemen mit dezentralem Abitur (Oerke und Maag Merki 2013; Oerke et al. 2013). Diese Analysen zeigten auch, dass sich der Effekt im Jahresvergleich nicht verändert hat, sodass nicht davon ausgegangen werden kann, dass die zunehmende Erfahrung der Lehrpersonen mit dem Zentralabitur zu einer stärkeren Berücksichtigung von nicht abiturrelevanten Inhalten oder von Schüler/inneninteressen führt.

Grundsätzlich muss ein solcher teaching-to-the-test-Effekt nicht unbedingt problematisch sein. Vor allem dann nicht, wenn a) die entsprechenden Prüfungsaufgaben curricular valide und anspruchsvoll sind und die Anpassungsprozesse der Lehrpersonen somit keine Reduktion der Unterrichtsinhalte oder des kognitiven Bearbeitungsniveaus zur Folge haben und wenn b) die Lehrpersonen trotz guter Aufgaben keine zu rigide Eingrenzung der Unterrichtsinhalte auf die Schwerpunktthemen der Abiturprüfungen vornehmen oder die Interessen der Schüler/innen im Unterricht weniger berücksichtigen, als dies möglich wäre.

Allerdings zeigen die Analysen im Zusammenhang mit der Einführung zentraler Abiturprüfungen in Bremen, dass die Beschränkung auf abiturrelevante Themen zur Vorbereitung auf das Zentralabitur durchaus einen Risikofaktor für die Schüler/innen auf motivationaler Ebene darstellt. Dort hatte sich ein negativer Effekt auf das Interesse am Fach gezeigt (Oerke et al. 2013). Dahingegen schlägt sich eine solche Beschränkung der Themen nicht in besseren Abiturleistungen nieder (Oerke und Maag Merki 2013). Eine zu starke Fokussierung auf die Abitur-schwerpunktthemen lohnt sich auch für die Lehrpersonen nicht, da dies für sie mit Unsicherheiten verbunden ist: Sie kennen die konkreten Abituraufgaben im Zentralabitur nicht und können trotz der vorliegenden Hinweise nicht sicher sein, ob ihre Auswahl der Unterrichtsinhalte die richtige ist. Der in verschiedenen Studien beobachtete teaching-to-the-test-Effekt ist insofern zumindest als ambivalent zu interpretieren.

Es sind jedoch auch positive Veränderungen im Zusammenhang mit der Einführung zentraler Abiturprüfungen zu beobachten. So gibt es für einzelne Fachkurse Hinweise darauf, dass sich die Unterstützungsqualität im Unterricht aus Sicht der Schüler/innen kurzfristig (Holmeier und Maag Merki 2012) und längerfristig (Maag Merki und Oerke 2016) verbessert hat. In längerfristiger Perspektive war dies in den Englisch-Leistungskursen in Bremen besonders deutlich sichtbar. Hier erlebten Schüler/innen, die in diesem Fachkurs zentrale Abiturprüfungen ablegten im Vergleich zu jenen, die dezentral geprüft wurden, in einem stärkeren Ausmaß, fachbezogene Rückmeldungen zu ihrem Leistungsstand, Anerkennung oder Lob für ihre Leistungen oder dass ihre Lehrpersonen ihnen etwas zutrauen.

Werden mögliche *Effekte auf die Schüler/innen* in den Fokus genommen, so ist ebenfalls von fach- und kursspezifischen Befunden auszugehen (Klein et al. 2014; Maag Merki 2016b). Im Gegensatz zu theoretischen Ansätzen der Bildungsökonomie (Bishop 1999; Jürges und Schneider 2010) kann nicht vermutet werden, dass sich der Einsatz und die Durchhaltefähigkeit der Schüler/innen während der Vorbereitung auf das Abitur positiv verändert. Vielmehr zeigten einzelne Studien, dass mit einem Wechsel des Prüfungssystems mindestens kurzfristig Belastungen wie Unsicherheit oder Angst vor Misserfolg steigen (Jürges und Schneider 2010; Jürges et al. 2012; Maag Merki 2011) oder aber keine Veränderungen zu beobachten sind (Baumert und Watermann 2000). Weitere Analysen gaben zudem vor allem für die Englisch-Leistungskurse Hinweise, dass motivationale Dimensionen des Lernens der Schüler/innen durch die Unterrichtspraxis der Lehrpersonen, insbesondere durch die von den Schüler/innen erlebte Kompetenzunterstützung, beeinflusst werden (Maag Merki und Oerke 2016). So schien der Wechsel des Systems Verschlechterungen hinsichtlich der Persistenz und der

schulischen Selbstwirksamkeit zu begünstigen. Gleichzeitig erlebten die Schüler/innen aber mit zentralen Abiturprüfungen eine stärkere Kompetenzunterstützung durch ihre Lehrpersonen als noch im dezentralen System. In der Fünfjahresperspektive wurde der anfänglich negative Effekt somit neutralisiert, d. h. bei diesen beiden Dimensionen gab es keine Unterschiede mehr zwischen dem alten und dem neuen Prüfungssystem. Weiterhin zeigte sich, dass zumindest in einzelnen Fächern das Fachinteresse stieg, wobei eine stärkere Unterstützung im Unterricht durch die Lehrpersonen das Interesse der Schüler/innen zusätzlich beeinflusste (Maag Merki und Oerke 2016). Hinsichtlich des Attributionsmusters der Schüler/innen zur Erklärung der eigenen Abiturnoten wiesen die Ergebnisse darauf hin, dass sich die Bedeutung der Lehrperson für die Einschätzung des eigenen Erfolgs leicht vergrößerte, ohne dass aber angenommen werden kann, dass die Einführung des Zentralabiturs bei den Schüler/innen zu einem Gefühl des Kontrollverlusts führte (Oerke 2012; Oerke et al. 2011).

Auch bei *den Lehrpersonen* – die bei der Einführung zentraler Prüfungen nicht direkt im Fokus standen – zeigten sich Veränderungen, die als transintentional bezeichnet werden können (Maag Merki 2012). So ging die Einführung dieser Verfahren mit einer Steigerung der Unsicherheit und des Leistungsdrucks bei den Lehrpersonen einher. Allerdings war nach der Einführung eine Abnahme dieser Gefühle zu beobachten. Auch stieg mit zunehmender Erfahrung mit dem Zentralabitur das Gefühl, durch das Zentralabitur entlastet zu werden. Wesentlich war dabei die Kooperation zwischen den Lehrpersonen, da diese deren Unsicherheit reduzieren konnte. Der Effekt kam insbesondere dann zum Tragen, wenn die Lehrpersonen nur über geringe Erfahrungen mit dem Zentralabitur verfügten. Allerdings kann aufgrund der vorliegenden Befunde kaum davon ausgegangen werden, dass mit der Einführung von zentralen Abiturprüfungen die Kooperation zwischen den Lehrpersonen steigt. Wer vor der Einführung bereits mit anderen Lehrpersonen kooperiert hatte, tat dies auch bei zentralen Abiturprüfungen, wer vorgängig nicht kooperierte, tat dies auch bei zentralen Prüfungen nicht. Bereits fehlende Kooperation wurde durch die entsprechende Einführung somit nicht zusätzlich angeregt.

Für die *Durchführung von Formen des ‚Gemeinsamen Prüfens‘* ist es somit bedeutsam zu untersuchen, inwiefern sich die – hier für zentrale Prüfungen kurz zusammengefassten – transintentionalen Veränderungen ebenfalls zeigen werden. Gerade für die stärker ‚bottom-up‘ eingeführten, gemeinsamen Prüfungen, die von einzelnen Lehrpersonen oder Fachschaften entwickelt worden sind, auf Freiwilligkeit basieren und dadurch eine hohe Eigeninitiative aufweisen, könnten die negativen oder ambivalenten Befunde weniger ausgeprägt sein oder sogar verschwinden. In den stärker ‚top-down‘ eingeführten und verordneten Verfahren – bei denen für

die Lehrpersonen weniger Spielraum, dafür jedoch eine Distanz zwischen den durchgeführten Verfahren bezüglich Inhalt, Aufgabenformat oder Beurteilungskriterien besteht – sind hingegen in größerem Ausmaß ambivalente und negative transintentionale Veränderungen zu erwarten.

2.4 Umgang mit Datenrückmeldungen aus ‚zentralen Prüfungen‘

Mit der Durchführung gemeinsamer Prüfungen verbindet sich die Erwartung, dass die Beurteilungen langfristig vergleichbarer und die Leistungen der Schüler/innen fairer beurteilt werden. Auch Veränderungen in der Unterrichtspraxis und bei den Kompetenzen der Lehrpersonen oder im Bildungssystem werden antizipiert (Mero 2013). Basis dieser Erwartungen ist ein kybernetisches Feedbacksystem, in dem ein Zusammenhang zwischen der Durchführung gemeinsamer Prüfungen auf der einen und Unterrichts-, Schul- und Systementwicklung auf der anderen Seite postuliert wird.

Ein Blick in die Literatur zeigt, dass verschiedene Theoriemodelle einen Zusammenhang zwischen der Durchführung von zentralen oder gemeinsamen Prüfungen und Schul- bzw. Unterrichtsentwicklungsprozessen beschreiben können (Maag Merki 2016b). Aus einer bildungswissenschaftlichen Perspektive heraus wird angenommen, dass über die Rückmeldung der Ergebnisse aus den zentralen Lernstandserhebungen, den zentralen Abschlussprüfungen oder den gemeinsamen Prüfungen sowie einem Vergleich von Ist- und Soll-Ergebnissen bestimmte Handlungsschritte in der jeweiligen Schule realisiert werden. Diese sollen – wiederum kurz- oder langfristig – zu einer besseren Anpassung der schulischen Prozesse und der Ist-Ergebnisse an die erwarteten Soll-Ergebnisse führen (vgl. Abb. 2.1). Dabei wird Bezug genommen auf ein Verständnis von Schulentwicklung, bei dem diese

einem selbstreflexiven Prozess der Verarbeitung von internen und externen Zielvorgaben und Ergebnissen von internen und externen Monitoringverfahren mit dem Ziel der Gestaltung einer optimalen Lernumgebung für die Bildungsprozesse der Schüler/innen [entspricht] (Maag Merki 2008, S. 25).

Die Schulen bzw. Lehrpersonen und Schulleitungen werden als aktive Akteure verstanden, die in der Lage sind, oder darin unterstützt werden müssen, über rezeptive und reflexive Prozesse der Auseinandersetzung mit einer möglichen Ist-Soll-Differenz geeignete Schulentwicklungsmaßnahmen abzuleiten und diese in funktionaler Weise in der schulischen Praxis umzusetzen.

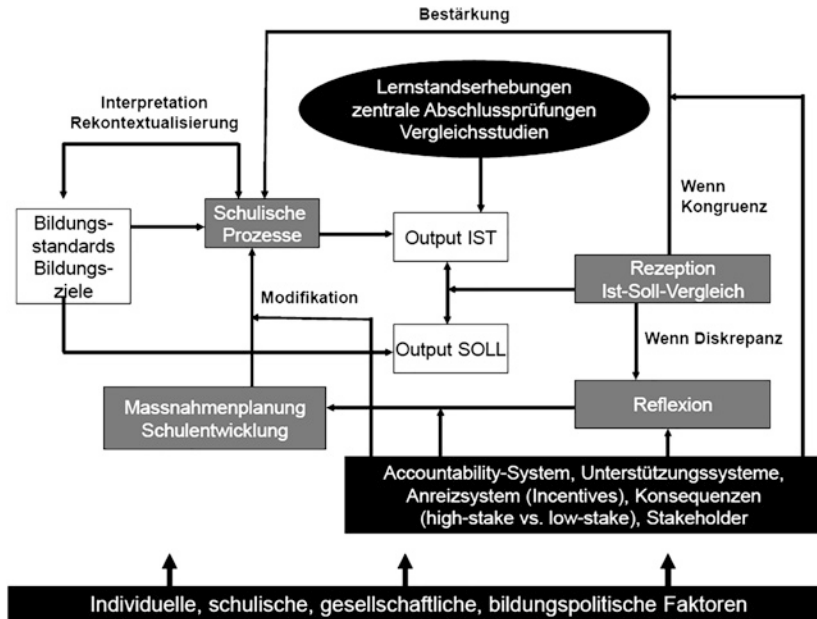


Abb. 2.1 Theoretisches Wirkungsmodell. (Maag Merki 2016b ©)

Beeinflusst wird dieser Prozess allerdings durch Rekontextualisierungsprozesse der Akteure (Fend 2008) und somit durch die Interpretationsleistungen, Kompetenzen und Einstellungen der Lehrpersonen in den Schulen. Zudem kann davon ausgegangen werden, dass gesellschafts- und bildungspolitische Kontextbedingungen sowie das Setzen von Anreiz- und Unterstützungselementen den Schulentwicklungsprozess und damit die Wirkungen dieser Verfahren deutlich beeinflussen. Damit ist zum einen zu erwarten, dass der Umgang mit zentralen oder gemeinsamen Prüfungen von Schule zu Schule unterschiedlich ist. Zum anderen kann aber auch angenommen werden, dass diese Auseinandersetzung aufgrund der unterschiedlichen Kontextbedingungen variiert und beispielsweise davon abhängt, wie viel Druck die Schulbehörden auf die Schulen und Lehrpersonen ausüben oder welche Unterstützung ihnen für eine nachhaltige Weiterentwicklung geboten wird.

Konkret zeichnet sich dieser Feedbackprozess durch verschiedene Klärungsschritte aus und ist spiralförmig zu verstehen. Nach dem ersten Durchlauf wird er auf einer höheren Wissensebene und gegebenenfalls unter Fokussierung eines

neuen Zieles fortgeführt. Folgende Fragen sind von großer Bedeutung, damit die Durchführung gemeinsamer oder zentraler Prüfungen zu positiven Entwicklungen im Unterricht, in der Schule oder sogar im Bildungssystem führen kann:

1. Welche Ziele sollen erreicht werden?
2. Wie können diese Ziele in der eigenen Schulpraxis erreicht werden?
3. Welche Ergebnisse erreichen die eigenen Schüler/innen in den Prüfungen?
4. Wo ergeben sich bedeutsame Differenzen zwischen den Leistungen der eigenen Schüler/innen und den Vorgaben?
5. Worin liegen die Ursachen dieser Differenzen?
6. Welche Maßnahmen sind geeignet, um diese Differenzen zu minimieren?
7. Wie können diese Maßnahmen in der eigenen schulischen Praxis umgesetzt werden?
8. Führen diese Maßnahmen tatsächlich zu besseren schulischen Leistungen?

Anhand dieser Fragen wird bereits deutlich, dass der Prozess hohe Anforderungen an die Schulen und Lehrpersonen stellt und keinesfalls davon ausgegangen werden kann, dass eine Realisierung einfach so, d. h. ohne Hindernisse möglich ist. Vielmehr ist er von einem komplexen System aus Bedingungen und Voraussetzungen abhängig. Besonders hervorzuheben sind die Fragen, die auf eine mögliche Differenz zwischen Ist und Soll abzielen, da die entsprechenden Antworten nicht einfach sind, werden die komplexen Bedingungen schulischen Lernens berücksichtigt (Helmke und Schrader 2006).

Neben der ersten Herausforderung, gemeinsame Prüfungen zu konzipieren, die curricular zulässig sind und es ermöglichen, die Leistungen der Schüler/innen valide und fair zu erfassen, besteht die zweite Herausforderung darin, einen Modus zu finden, wie Lehrpersonen aus den gemeinsamen Prüfungen Erkenntnisse für die eigene Unterrichtspraxis und die bessere Förderung der Schüler/innen gewinnen und nutzbar machen können. Es ist anzunehmen, dass darin die eigentlichen Schwierigkeiten in Bezug auf die Implementation des ‚Gemeinsamen Prüfens‘ liegen. Schließlich gibt es selten einfache Erklärungen, warum in einer Klasse oder einer Schule bestimmte Ziele nicht erfüllt werden und/oder mit welchen Strategien eine Leistungssteigerung erreichbar ist.

Weiterhin ist zu beachten, dass das angesprochene Wirkungsmodell (Abb. 2.1) zwar theoretisch plausibel sein mag, dass sich in der Praxis jedoch verschiedene Stolpersteine zeigen. Wie Altrichter et al. (2016) in ihrer Übersicht über den aktuellen Stand der Forschung zu den Effekten von Ergebnismrückmeldungen bei Vergleichsarbeiten auf die Schul- und Unterrichtsentwicklung resümieren, ist eher davon auszugehen, dass dieses Wirkungsmodell sehr viel differenzierter skizziert werden müsste, um die Realität gut abzubilden.

So zeigen die bisherigen Forschungsbefunde, dass die Daten und Ergebnisse aus Vergleichsarbeiten zu wenig für interne Analysen genutzt werden und sie keine nachhaltigen Impulse für die Unterrichtsentwicklung ermöglichen. So hat

das Idealmodell outputorientierter Steuerung, der Regelkreis von Zielbestimmungen, Überprüfungen, Rückmeldungen, Entwicklungsmaßnahmen, Unterstützungssystemen und Verbesserungen [...] seinen kritischsten Punkt, aber auch gleichzeitig seine größte Schwachstelle dort, wo Ziele und Rückmeldungen in Entwicklungsarbeit und verändertes Unterrichtshandeln übersetzt werden sollen (Altrichter et al. 2016, S. 252).

Gemäß den Autor/innen könnten dafür verschiedene Gründe herangezogen werden (ebd., S. 253 ff.), so beispielsweise:

- die nur mangelhafte Qualität der Ergebnisrückmeldungen (z. B. Umfang, Lesbarkeit, Verständlichkeit, Aufbereitung der Daten, Kontextinformationen, Zeit zwischen Datengewinnung und Datenrückmeldung),
- die geringe Akzeptanz und Kompetenz der Lehrpersonen, die Daten und Ergebnisse valide interpretieren zu können; ihre Grundüberzeugungen bzw. eine mangelnde Passung zwischen den bereitgestellten Informationen durch Datenrückmeldungen und den Arbeits- und Denkstilen der Lehrpersonen,
- die Organisation der Einzelschule und die Kultur im Kollegium oder in Fachschaften (keine Kultur der Kooperation und der fachbezogenen Zusammenarbeit im Schulhaus oder keine Kultur der gemeinsamen Weiterentwicklung der eigenen Schule oder Fachschaft); dabei ist davon auszugehen, dass die Fähigkeit einer Schule, Schulentwicklung zu betreiben, Voraussetzung dafür ist, dass eine lernwirksame Verarbeitung von Daten und Ergebnissen aus klassenübergreifenden Tests realisiert werden kann,
- eine fehlende Implementationskompetenz in der Schule und mangelnde interne und externe Unterstützungsleistungen (z. B. durch Schulleitungen, Schulbehörden, Lehrmittel, Weiterbildungskurse),
- fehlende fordernde und fördernde Systemstrukturen, so beispielsweise genügend Zeitressourcen, um sich intensiv mit den Ergebnissen von vergleichenden Leistungstests auseinanderzusetzen, fehlende schulische Verfahren der Qualitätssicherung und -entwicklung, in die das Datenfeedback eingebunden ist.
- Hinweise im Feedback, die zu Fehlinterpretationen führen oder die nicht für das eigene Lernen als Lehrperson oder die Weiterentwicklung der Aufgaben genutzt werden können.

- Unübliches Verständnis von Feedback: Lehrpersonen erhalten Rückmeldungen über andere Personen (Schüler/innenleistungen) und sollen diese für die eigene Arbeit und Unterrichtsentwicklung nutzen.

Inwiefern die Einführung von ‚Gemeinsamem Prüfen‘ in den Schulen längerfristigen Zielen in der Unterrichts-, Schul- und Systementwicklung dienen kann, ist somit von einem komplexen Bedingungsgefüge abhängig. ‚Bottom-up‘ eingeführte, gemeinsame Prüfungen könnten eine größere Effektivität für Unterrichts- und Schulentwicklung erreichen, da sie selber in den Fachschaften und Schulen entwickelt worden sind und somit eine hohe Passung mit den konkreten situationalen Bedingungen aufweisen. Zudem sind die Lehrpersonen eng in den Entscheidungs- und Entwicklungsprozess involviert, können dadurch stärker motiviert werden und die Verfahren in höherem Maße akzeptieren. Im Gegensatz dazu besteht allerdings die Gefahr, dass ‚bottom-up‘ Verfahren ‚auf einem Auge blind sind‘ und letztlich im Systemvergleich nicht zu einer verbesserten Vergleichbarkeit der Beurteilung der Leistungen der Schüler/innen führen.

‚Top-down‘ eingeführte, klassen- und schulübergreifende Verfahren wiederum sind vergleichbarer, scheinen aber eine größere Distanz zu den konkreten Unterrichtsaufgaben und -herausforderungen in den einzelnen Klassen aufzuweisen. Die Motivation und Akzeptanz der Lehrpersonen könnte geringer sein und die Ergebnisse nur bedingt für den aktuellen Unterricht sowie für die Verbesserung des Unterrichts genutzt werden, zumal die Rückmeldung an die Lehrpersonen und Schulen meist auch mit einer gewissen zeitlichen Verzögerung erfolgt. Damit ist es besonders interessant zu analysieren, wie diese unterschiedlichen Verfahren hinsichtlich ihres Potenzials für die Unterrichts-, Schul- und Systementwicklung eingeschätzt werden. Dabei könnte entscheidend sein, wie die Stärken und Schwächen der stärker ‚top-down‘ und ‚bottom-up‘ durchgeführten Verfahren ausgestaltet sind und ob in spezifischen Verfahren eine optimale Balance zwischen ‚top-down‘ und ‚bottom-up‘ durchgeführten Verfahren erreicht werden kann.

Aus theoretischer Sicht (Creemers und Kyriakides 2012; Fend 2008) und gestützt durch bisherige Befunde bei der Einführung von standardisierten Prüfungsformen oder darüber hinausgehenden Unterrichtsreformen (Bryk et al. 2010; Maag Merki 2016a; Stringfield et al. 2008) ist zudem anzunehmen, dass die Realisierung des ‚Gemeinsamen Prüfens‘ zwar von formalen Vorgaben und Rahmenbedingungen beeinflusst wird, dass aber die konkrete Umsetzung Ergebnis eines komplexen ‚Übersetzungsprozesses‘ der schulinternen Akteure ist. Zudem hängt der Übersetzungsprozess von den Kompetenzen, Einstellungen und Vorerfahrungen der Akteure, vor allem der Lehrpersonen und der Schulleitung ab,

ebenso wie vom fachlichen Kontext, insofern als die Problemlagen und Möglichkeiten bei der Umsetzung des ‚Gemeinsamen Prüfens‘ fachspezifisch variieren. Davon ausgehend, dass die Implementation gemeinsamer Prüfungsformen als ein Schul- und Unterrichtsentwicklungsunterfangen verstanden wird (Mero 2013), sind des Weiteren kollektive Erfahrungen in der Schule wie auch schul- und fachteamspezifische Schul- und Unterrichtsentwicklungskompetenzen mögliche Einflussfaktoren auf die Realisierung des ‚Gemeinsamen Prüfens‘. Diese Faktoren stehen in einem engen Beziehungsgeflecht zueinander. Um Hinweise auf mögliche dominante Einflussfaktoren, insbesondere aus dem Kontext der Schule, des Verfahrens und des Faches zu erhalten, realisiert die vorliegende Studie unterschiedliche vergleichende Analysen. Im nachfolgenden Kapitel werden die Analysebereiche sowie das theoretische Rahmenmodell vorgestellt.

Literatur

- Altrichter, H., Moosbrugger, R., & Zuber, J. (2016). Schul- und Unterrichtsentwicklung durch Datenrückmeldung. In H. Altrichter & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem*. 2., erweiterte Aufl. (S. 235–277). Wiesbaden: Springer VS.
- Baumert, J., Trautwein, U., & Artelt, C. (2003). Schulumwelten – institutionelle Bedingungen des Lehrens und Lernens. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann & M. Weiss (Hrsg.), *PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 259–330). Opladen: Leske+Budrich.
- Baumert, J., & Watermann, R. (2000). Institutionelle und regionale Variabilität und die Sicherung gemeinsamer Standards in der gymnasialen Oberstufe. 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. Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (S. 317–372). Opladen: Leske+Budrich.
- Beutel, S.-I., & Vollstädt, W. (2000). *Leistung ermitteln und bewerten*. Hamburg: Bergmann + Helbig Verlag.
- Bishop, J. H. (1999). Are national exit examinations important for educational efficiency. *Swedish Economic Policy Review*, 6, 349–398.
- Bryk, A. S., Bender Sebring, P., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organizing schools for improvement. Lessons from Chicago*. Chicago, London: The University of Chicago Press.
- Cheng, L., Watanabe, Y., & Curtis, A. (2004). *Washback in Language Testing. Research Contexts and Methods*. Mahwah, N.J.: Lawrence Erlbaum.
- Creemers, B. P. M., & Kyriakides, L. (2012). *Improving quality in education. Dynamic approaches to school improvement*. New York: Routledge.
- Fend, H. (2008). *Schule gestalten. Systemsteuerung, Schulentwicklung und Unterrichtsqualität*. Wiesbaden: VS Verlag für Sozialwissenschaften.

- Helmke, A., & Schrader, F.-W. (2006). Determinanten der Schulleistung. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie. 3. überarb. und erweiterte Ausgabe* (S. 83–94). Weinheim: Beltz.
- Holmeier, M. (2013). *Leistungsbeurteilung im Zentralabitur*. Wiesbaden: Springer VS.
- Holmeier, M., & Maag Merki, K. (2012). Unterstützung im Unterricht im Kontext der Einführung zentraler Abiturprüfungen. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 155–178). Wiesbaden: Springer VS.
- Jürges, H., & Schneider, K. (2010). Central exit examinations increase performance, but take the fun out of mathematics. *Journal of Population Economics* 23, 497–517.
- Jürges, H., Schneider, K., Senkbeil, M., & Carstensen, C., H. (2012). Assessment drives learning: The effect of central exit exams on curricular knowledge and mathematical literacy. *Economics of Education Review*, 31, 56–65.
- Klein, E. D., Krüger, M., Kühn, S. M., & Van Ackeren, I. (2014). Wirkungen zentraler Abschlussprüfungen im Mehrebenensystem Schule. Eine Zwischenbilanz internationaler und nationaler Befunde. *Zeitschrift für Erziehungswissenschaft*, 17, 7–33.
- 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.
- Kühn, S. M. (2012). Zentrale Abiturprüfungen im nationalen und internationalen Vergleich mit besonderer Perspektive auf Bremen und Hessen. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 27–44). Wiesbaden: Springer VS.
- Maag Merki, K. (2008). Die Architektur einer Theorie der Schulentwicklung. Strukturanalyse und Interdependenzen. *Journal für schulentwicklung*, 12(2), 22–30.
- Maag Merki, K. (2011). Effects of the implementation of state-wide exit exams on students' self-regulated learning. *Studies in Educational Evaluation*, 37, 196–205.
- Maag Merki, K. (2012). Die Einführung zentraler Abiturprüfungen – eine (Zwischen-) Bilanz nach drei Jahren. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 385–407). Wiesbaden: Springer VS.
- Maag Merki, K. (2016a). Die Einführung zentraler Abiturprüfungen in Bremen und Hessen und ihre Effekte auf Unterricht und Lernen. Zentrale Befunde nach fünf Jahren. In J. Kramer, M. Neumann & U. Trautwein (Hrsg.), *Abitur und Matura im Wandel. Historische Entwicklungslinien, aktuelle Reformen und ihre Effekte. Edition ZfE, Volume 2* (S. 129–159). Wiesbaden: Springer VS.
- Maag Merki, K. (2016b). 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, 2., erweiterte Ausgabe* (S. 151–182). Wiesbaden: Springer VS.
- Maag Merki, K., & Holmeier, M. (2015). Comparability of semester and exit exams grades: Long-term effect of the implementation of state-wide exit exams. *School Effectiveness and School Improvement*, 26(1), 57–74.
- Maag Merki, K., & Oerke, B. (2016). Long-term effects of the implementation of state-wide exit exams: a multilevel regression analysis of mediation effects of teaching practices on students' motivational orientations. *Educational Assessment, Evaluation and Accountability*. doi: 10.1007/s11092-016-9244-y.

- Maué, E. (2013a). *Comparability of grades. Does this vision become reality by means of state-wide exit exams?* Paper presented at the AERA Annual Meeting 2013, San Francisco, CA.
- Maué, E. (2013b). Vergleichbarkeit von Abiturnoten – eine Fiktion? Längerfristige Effekte der Implementation zentraler Abiturprüfungen in Bremen. In J. Asdonk, S. U. Kuhnen & P. Bornkessel (Hrsg.), *Übergang Schule – Hochschule. Analysen – Konzeptionen – Gestaltungsperspektiven* (S. 114–128). Münster: Waxmann.
- Mero, R. (2013). *EDK-Projekt «Gymnasiale Maturität langfristige Sicherung des prüfungsfreien Hochschulzugangs».* Bericht zum Teilprojekt 2 «Unterstützungsangebote zum Gemeinsamen Prüfen». Bern: wbz cps.
- Neumann, M., Nagy, G., Trautwein, U., & Lüdtke, O. (2009). Vergleichbarkeit von Abiturleistungen. Leistungs- und Bewertungsunterschiede zwischen Hamburger und Baden-Württemberger Abiturienten und die Rolle zentraler Abiturprüfungen. *Zeitschrift für Erziehungswissenschaft*, 12(4), 691–714.
- Neumann, M., Trautwein, U., & Nagy, G. (2011). Do central examinations lead to greater grading comparability? A study of frame-of-references effects on the University entrance qualification in Germany. *Studies in Educational Evaluation*, 37, 206–217.
- Oerke, B. (2012). Emotionaler Umgang von Lehrkräften und Schüler/-innen mit dem Zentralabitur: Unsicherheit, Leistungsdruck und Leistungsattribution. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 119–153). Wiesbaden: Springer VS.
- Oerke, B., & Maag Merki, K. (2013). Teaching-to-the-test infolge des Zentralabiturs – gut für die Abiturleistung, schlecht für das Fachinteresse? *Schulverwaltung* (2), 48–51.
- Oerke, B., Maag Merki, K., Holmeier, M., & Jäger, D. J. (2011). Changes in student attributions due to the implementation of central exit exams. *Educational Assessment, Evaluation and Accountability*, 23(3), 223–241.
- Oerke, B., Maag Merki, K., Maué, E., & Jäger, D. J. (2013). Zentralabitur und Themenvarianz: Lohnt sich Teaching-to-the-Test? In D. Bosse, F. Eberle & S.-T. Barbara (Hrsg.), *Standardisierung in der gymnasialen Oberstufe* (S. 27–50). Wiesbaden: Springer VS.
- Sacher, W. (2009). *Leistungen entwickeln, überprüfen und beurteilen: Bewährte und neue Wege für die Primar- und Sekundarstufe, 5. überarbeitete und erweiterte Aufl.* Bad Heilbrunn: Klinkhardt.
- Stecher, B. M. (2002). Consequences of large-scale, high-stakes testing on school and classroom practice. In L. S. Hamilton, B. M. Stecher & S. P. Klein (Hrsg.), *Making sense of test-based accountability in education* (S. 79–100). Santa Monica, CA: Rand Corporation.
- Stringfield, S., Reynolds, D., & Schaffer, E. C. (2008). Improving secondary students' academic achievement through a focus on reform reliability: 4- and 9-year findings from the High Reliability Schools project. *School Effectiveness and School Improvement*, 19(4), 409–428.
- 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 regims. *Education Policy Analysis Archives*, 20(8), 1–32.

Gemeinsames Prüfen

Eine Fallanalyse in Gymnasien der Schweiz

Holmeier, M.; Maag Merki, K.; Hirt, C.

2017, X, 273 S. 5 Abb., Softcover

ISBN: 978-3-658-17059-2