Z Erziehungswiss (2009) 12:691–714
DOI 10.1007/s11618-009-0099-6

ALLGEMEINER TEIL

ZfE

Vergleichbarkeit von Abiturleistungen

Leistungs- und Bewertungsunterschiede zwischen Hamburger und Baden-Württemberger Abiturienten und die Rolle zentraler Abiturprüfungen

Marko Neumann · Gabriel Nagy · Ulrich Trautwein · Oliver Lüdtke

Zusammenfassung: Der vorliegende Beitrag untersucht am Beispiel der Mathematik- und Englischleistungen vorhandene Leistungsunterschiede zwischen Baden-Württemberger (N=3526) und Hamburger (N=3734) Abiturienten daraufhin, ob und in welchem Ausmaß sich diese Leistungsunterschiede auch in den erteilten Noten im Abiturzeugnis widerspiegeln. Dabei wird auch der Frage nachgegangen, inwieweit Prüfungsnoten aus zentral vorgegebenen Abiturprüfungen eine bessere Vergleichbarkeit ermöglichen als die Fachnoten. Mehrebenenanalysen zur Vorhersage der Fachnoten ergaben für Mathematik deutliche Unterschiede in der Bewertungsstrenge zwischen den Bundesländern. Bei vergleichbaren Leistungen erhielten Abiturienten aus Hamburg in Mathematik bessere Fachnoten. Die Bewertungsunterschiede ließen sich zum Teil auf Referenzgruppeneffekte bei der Notenvergabe zurückführen. Für das Fach Englisch fanden sich keine Unterschiede in der Bewertungsstrenge zwischen den Bundesländern. Die Prüfungsnoten in Mathematik korrespondierten besser mit dem Leistungsniveau als die Fachnoten. Die Befunde werden vor dem Hintergrund der Verteilungsgerechtigkeit beim Zugang zu stark nachgefragten Studien- und Ausbildungsplätzen diskutiert.

Schlüsselwörter: Abitur · Hochschulzugang · Schulnoten · Zentrale Prüfungen · Referenzgruppeneffekte

Online publiziert: 23.01.2010

M. Neumann, M.A. (✉) · Dr. G. Nagy
Erziehungswissenschaft und Bildungssysteme, Max-Planck-Institut für Bildungsforschung,
Lentzeallee 94, 14195 Berlin, Deutschland
E-Mail: markoneumann@mpib-berlin.mpg.de

Dr. G. Nagy
E-Mail: nagy@mpib-berlin.mpg.de

Univ.-Prof. Dr. U. Trautwein · Univ.-Prof. Dr. O. Lüdtke
Fakultät 08, Institut für Erziehungswissenschaft, Universität Tübingen,
Europastr. 6, 72072 Tübingen, Deutschland
E-Mail: ulrich.trautwein@uni-tuebingen.de

Univ.-Prof. Dr. O. Lüdtke
E-Mail: oliver.luedtke@uni-tuebingen.de

Cross-state comparability of the abitur qualification in Germany. Differences in achievement and grading between abitur-level students from Baden-Württemberg and Hamburg and the role of central exams

Abstract: This article examines differences in the mathematics and English proficiency of academic-track students in Baden-Württemberg (N=3526) and Hamburg (N=3734), investigating whether and to what extent these differences are reflected in the Abitur grades the students are awarded. The article also examines the extent to which scores in centrally conducted examinations provide better comparability than do coursework grades. Multilevel analyses predicting coursework grades in mathematics revealed clear between-state differences, with students in Hamburg being awarded higher mathematics grades than comparably able students in Baden-Württemberg. These differences are partly attributable to frame of reference effects and their impact on teachers' grading practices. No corresponding between-state differences were found for English. Mathematics examination scores provided a much better measure of student achievement than mathematics coursework grades. The findings are discussed in terms of meritocratic access to sought-after university and training places.

Keywords: Central exams · Frame of reference effects · School grades · University access · University-entrance diploma (Abitur)

1 Einleitung

Die großen internationalen und nationalen Schulleistungsstudien der letzten Jahre haben für Deutschland erhebliche Leistungsunterschiede zwischen Schülerinnen und Schülern aus unterschiedlichen Bundesländern, Schulformen innerhalb der Bundesländer sowie Schulen derselben Schulform im selben Bundesland aufgezeigt (Baumert et al. 2003; Bos et al. 2004a; Lehmann et al. 2002; Köller et al. 2004; Prenzel et al. 2005; Trautwein et al. 2007).

Leistungsunterschiede zwischen Schülerinnen und Schülern sind nicht von vornherein als problematisch einzustufen. Vor dem Hintergrund einer optimalen Förderung aller Schülerinnen und Schüler nach ihren individuellen Lernvoraussetzungen ist Leistungsvariation ein zu erwartendes und wünschenswertes Ergebnis institutionalisierter Bildungsprozesse (Baumert u. Watermann 2000). Kritisch werden diese Leistungsunterschiede jedoch zum einen dann, wenn im unteren Leistungsbereich Mindestanforderungen in bedeutsamem Ausmaß unterschritten werden. Zum anderen, und hier soll der Schwerpunkt des vorliegenden Beitrags liegen, sind Unterschiede im Kompetenzniveau von Schülerinnen und Schülern als problematisch einzustufen, wenn sich diese Kompetenzunterschiede nur unzureichend in Leistungsbewertungen und Abschlusszertifikaten niederschlagen, die den Zugang zu attraktiven Bildungswegen eröffnen (Baumert et al. 2003).

Im vorliegenden Beitrag untersuchen wir am Beispiel der Mathematik- und Englischleistungen vorhandene Leistungsunterschiede zwischen Baden-Württemberger und Hamburger Abiturienten daraufhin, ob und in welchem Ausmaß sich diese Leistungsunterschiede auch in den erteilten Noten im Abiturzeugnis widerspiegeln. Die Untersuchung liefert damit – wenn auch nur in Ausschnitten – erstmals Befunde zur Vergleichbarkeit von Abiturnoten, die von Schülerinnen und Schülern aus unterschiedlichen Bundesländern erzielt

wurden. Vor dem Hintergrund der aktuellen Diskussionen um ein bundesweites bzw. bundeslandübergreifendes Zentralabitur wird dabei auch der Frage nachgegangen, inwieweit Prüfungsnoten aus zentral vorgegebenen Abiturprüfungen eine bessere Vergleichbarkeit ermöglichen als Fachnoten, deren Abhängigkeit von lerngruppenspezifischen Referenzmaßstäben immer wieder bemängelt wird (vgl. Ingenkamp 1995; Wild u. Krapp 2001).

2 Leistungsbewertung und Verteilungsgerechtigkeit

Die besondere Relevanz der Vergleichbarkeit von Schulnoten und Abschlusszertifikaten ergibt sich vor allem aus der engen Verbindung von schulischer Abschlussqualifikation und Zugangsberechtigung für die weiterführenden und an bestimmte Aufnahmebedingungen geknüpften Ausbildungsstufen. Neben dem Übergang von der Grundschule in die weiterführenden Schulen (vgl. Bos et al. 2004b; Maaz et al. 2008) zeigt sich diese Kopplung besonders deutlich im Falle der Hochschulreife. Die im Abiturzeugnis enthaltenen Noten spielen eine entscheidende Rolle beim Zugang zu besonders begehrten Studienplätzen, auch wenn sich die Bedeutung der Abiturgesamtnote im Zuge der zunehmenden Autonomie der Hochschulen bei der Auswahl ihrer Studierenden gegenwärtig etwas relativiert und andere Aufnahmekriterien, wie z. B. studienfachbezogene Eingangstests und Auswahlgespräche, an Bedeutung gewinnen (vgl. Wissenschaftsrat 2004).

2.1 Aussagekraft von Schulnoten und die Rolle von Referenzgruppeneffekten

Schulnoten übernehmen eine Vielzahl an Funktionen und entsprechend hoch sind die an sie gestellten Erwartungen (Jürgens 2005; Sacher 2005; Schrader u. Helmke 2001; Tent 2001; Wild u. Krapp 2001). Für die Abnehmerseite (z. B. weiterführende Schulen, Hochschulen und Arbeitgeber) besitzen Schulnoten in erster Linie eine Indikator- und eine Selektionsfunktion. Sie sollen Hinweise über Stärken und Schwächen (Kompetenzprofile) der Schülerinnen und Schüler liefern und eine verlässliche Verortung des Einzelnen im gesamten Leistungsspektrum ermöglichen, so zumindest die Idealvorstellung. In der Realität zeigt sich jedoch gerade in dieser Hinsicht eine eher eingeschränkte Aussagekraft von Schulnoten, die objektive, reliable und valide Schlüsse auf tatsächlich vorhandene Kompetenzunterschiede nur bedingt erlauben (Tent 2001; Wild u. Krapp 2001). In vielen Studien hat sich gezeigt, dass Lehrkräfte ganz unterschiedliche Kriterien und Maßstäbe bei der Vergabe von Schulnoten heranziehen und Noten entsprechend oftmals nur geringe Beurteilerübereinstimmungen und Wiederholungszuverlässigkeiten aufweisen (vgl. Ingenkamp 1995; Jürgens 2005; Schrader u. Helmke 2001; Tent 2001). Ein Grund dafür könnte sein, dass den Schulnoten neben ihrer Diagnosefunktion auch eine Anreiz- oder Motivationsfunktion zugeschrieben wird: Schulnoten sollen nicht ausschließlich auf Grundlage der gezeigten Leistung vergeben werden, sondern können auch unterrichtsstützende Aspekte wie Lernmotivation und Anstrengungsbereitschaft enthalten und neben einer rein sozialen oder kriterialen Vergleichsperspektive auch individuelle Entwicklungsfortschritte berücksichtigen (Rheinberg 2001). Insgesamt wird damit deutlich, dass Schulnoten eine Vielzahl wichtiger und zum Teil konfligierender Funktionen übernehmen. Dennoch wird man nicht bestreiten können, dass der Allokations- und Selek-

tionsfunktion insbesondere an Übergangspunkten in den individuellen Bildungsverläufen eine besondere Bedeutung zukommt.

Vor dem Hintergrund der Bedeutung von Schulnoten für Bildungsweg- und Berufslaufbahnentscheidungen ist ein Kritikpunkt von besonders zentraler Bedeutung: die mangelnde Vergleichbarkeit von Noten über verschiedene Klassen und Schulen hinweg (Wild u. Krapp 2001). Lehrkräfte vergeben in allen Klassen gute und schlechte Noten und bedienen sich dabei mehr oder weniger des gesamten Notenspektrums. Die erteilten Noten bilden zwar die Leistungsrangreihe der Schülerinnen und Schüler innerhalb der Referenzgruppe (z. B. der Klasse oder der Schule) recht gut ab (vgl. Schrader 2001), vergleicht man jedoch Schülerinnen und Schüler aus unterschiedlichen Referenzgruppen miteinander, finden sich bei gleichen Noten zum Teil substanzielle Unterschiede in den tatsächlich vorhandenen Kompetenzen. Die gleiche Leistung wird in Abhängigkeit der mittleren Leistungsstärke der Bezugsgruppe unterschiedlich benotet. Je höher das mittlere Leistungsniveau der Referenzgruppe, umso ungünstiger fällt – bei gleicher individueller Leistung – die Bewertung aus. In leistungsschwächeren Lerngruppen wird die gleiche Leistung besser bewertet als in leistungsstarken Lerngruppen (Ingenkamp 1969; Trautwein u. Baeriswyl 2007). Dies kann dazu führen, dass sich Klassen bzw. Schulen in ihren mittleren Fachnoten nicht unterscheiden, obschon die mit standardisierten Leistungstests gemessenen Kompetenzniveaus zwischen den Gruppen – z. B. Schülerinnen und Schülern aus unterschiedlichen Bundesländern – stark differieren.

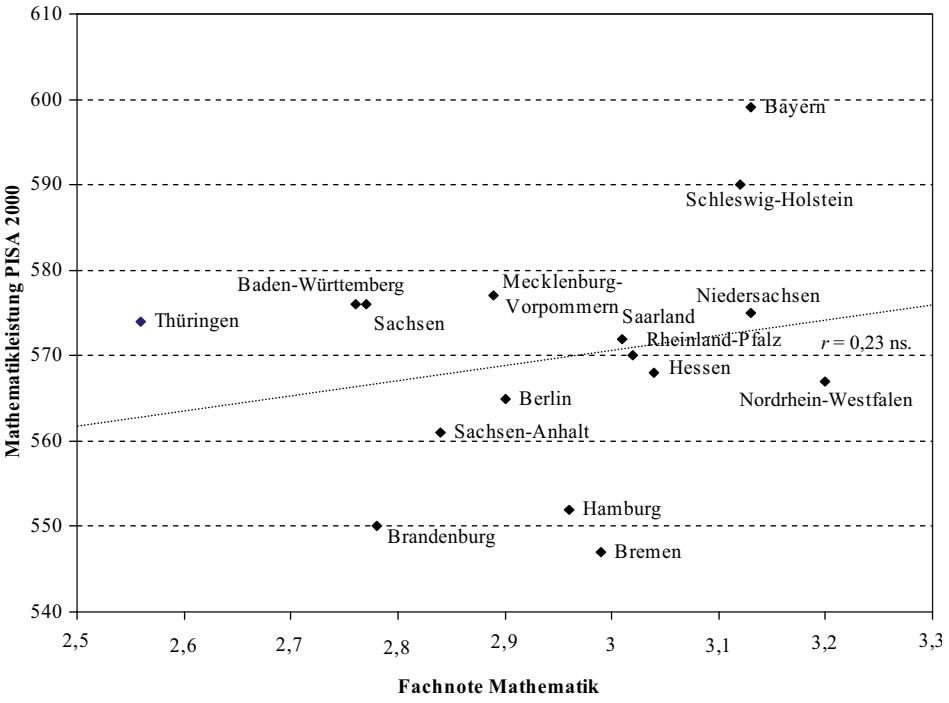
2.2 Regionale Leistungs- und Bewertungsunterschiede in der Grundschule und der Sekundarstufe I

Bei der Frage nach der Leistungsgerechtigkeit bei der Notenvergabe ist zunächst eine begriffliche Präzisierung vorzunehmen. So ist sorgfältig zwischen Leistungsunterschieden, Notenunterschieden und Bewertungsunterschieden zu differenzieren. Mit *Leistungsunterschieden* sind im Folgenden tatsächlich vorhandene und mit standardisierten Leistungstests erhobene Unterschiede in den Kompetenzen von Schülerinnen und Schülern bzw. zwischen Schülergruppen gemeint. *Notenunterschiede* sagen hingegen nur etwas über Unterschiede in den individuellen oder mittleren vergebenen Schulnoten aus, ohne dass dabei das tatsächliche Leistungsniveau (Leistungsunterschiede) mit einbezogen wird (z. B.: „In Bundesland A beträgt der Notendurchschnitt in Mathematik 2,6, in Bundesland B 3,1."). Von *Bewertungsunterschieden* bzw. unterschiedlicher *Bewertungsstrenge* sprechen wir nur dann, wenn vorhandene Notenunterschiede nicht mit entsprechenden Leistungsunterschieden einhergehen bzw. identische Leistungen unterschiedlich benotet werden (z. B.: „Bei identischen Testleistungen werden die Schülerinnen und Schüler in Bundesland B [mittlere Note 3,1] strenger bewertet als in Bundesland A [mittlere Note 2,6]."). Denkbar wäre in diesem Beispiel aber auch, dass in Bundesland A strenger benotet wird, nämlich dann, wenn der Leistungsvorsprung zugunsten Bundesland A größer ausfällt als der entsprechende Notenunterschied.

Für den Grundschulbereich haben Bos et al. (2004a) auf der Datengrundlage der Internationalen Grundschul-Lese-Untersuchung (IGLU) und deren deutscher Erweiterung (IGLU-E) die erzielten Fachnoten und erreichten Testleistungen in Lesen und Mathematik für mehrere Bundesländer gegenübergestellt. Ihre Analysen erbrachten klare Hin-

weise auf Bewertungsunterschiede zwischen den betrachteten Bundesländern. Im Bereich Lesen wiesen Schülerinnen und Schüler in Baden-Württemberg und Bayern im Mittel auf allen Notenstufen bessere Testleistungen auf, als dies in Nordrhein-Westfalen und Hessen der Fall war. Ähnliche Befunde zeigten sich auch für das Fach Mathematik. Die Analysen ergaben darüber hinaus auch deutliche Hinweise auf Bewertungsunterschiede innerhalb der Bundesländer. In Anbetracht der Tatsache, dass die Deutsch- und Mathematiknoten in den meisten Bundesländern eine wesentliche Rolle für die Schullaufbahnempfehlung spielen, sind die Befunde als besonders relevant einzustufen (Bos et al. 2004b).

Für die Sekundarstufe I haben Baumert et al. (2003) Leistungs- und Bewertungsunterschiede auf der Basis der deutschen PISA-2000-Erweiterungsstichprobe untersucht. Auch ihre Analysen erbrachten deutliche Hinweise auf regionale Bewertungsunterschiede. So bewegten sich die über alle Schulformen gemittelten Mathematiknoten der Neuntklässler zwischen den Bundesländern in einer Bandbreite von 2,74 und 3,16 – allerdings ohne dass diese Notenunterschiede in Zusammenhang mit den in den Bundesländern erzielten Testleistungen standen. Ein weitgehend vergleichbares Bild zeigt sich, wenn ausschließlich die Schülerinnen und Schüler am Gymnasium betrachtet werden, wie in Abb. 1 illustriert. Über alle Bundesländer ergibt sich ein leicht positiver, aber statistisch nicht signifikanter Zusammenhang von $r = 0{,}23$ zwischen mittlerer Mathematikleistung und mittlerer Fach-



*Eigene Berechnung auf Basis der PISA-E-2000 Daten

Abb. 1: Mittlere Fachnote* und Fachleistung in Mathematik von Neuntklässlern am Gymnasium (PISA-E 2000) nach Bundesland (Quelle: Neubrand u. Klieme 2002, S. 123)

note, das heißt, je höher die Leistung, desto höher und damit schlechter fallen die Noten aus. Bei angenommener Bewertungsgerechtigkeit würde man hier jedoch einen substanziellen negativen Zusammenhang erwarten: Je besser die mittleren Leistungen, umso besser sollte auch das mittlere Notenniveau ausfallen.

2.3 Regionale und institutionelle Bewertungsunterschiede in der gymnasialen Oberstufe

Für das Abitur existierten bislang kaum Untersuchungen, die verlässliche Schlüsse über länderübergreifende Leistungs- und Bewertungsunterschiede am Ende der gymnasialen Oberstufe erlauben. Eine Ausnahme bilden die von Baumert u. Watermann (2000) im Rahmen der *Third International Mathematics and Science Study* (TIMSS/III, vgl. Baumert et al. 2000) durchgeführten Analysen zur regionalen Variabilität von Bewertungsmaßstäben in der Oberstufe. Die Autoren teilten die Bundesländer in vier größere Gebietseinheiten mit jeweils ähnlichen Schulbesuchsquoten in der gymnasialen Oberstufe ein, wobei die neuen Länder eine separate Kategorie bildeten. Aussagen auf Ebene einzelner Bundesländer waren aufgrund zu geringer Fallzahlen nicht möglich. Während sich für die Grundkurse in Mathematik und Physik keine Unterschiede der Bewertungsstrenge in Abhängigkeit von der Gebietszugehörigkeit nachweisen ließen, fanden sich für die Leistungskurse deutliche Hinweise auf unterschiedliche Bewertungsmaßstäbe in den betrachteten Gebietseinheiten. So wurde in Mathematik in der Gruppe der alten Länder mit geringer Oberstufenquote vergleichsweise streng benotet. Im Fach Physik wurde in den neuen Ländern über alle Notenstufen hinweg deutlich milder benotet, was die Autoren zu dem Fazit führte, „dass man wohl nicht mehr von äquivalenten Bewertungsgrundsätzen zwischen Ost und West sprechen kann“ (Baumert u. Watermann 2000, S. 340).

Verschiedene Studien haben sich mit der Vergleichbarkeit von Abiturleistungen innerhalb eines Bundeslandes zwischen unterschiedlichen Oberstufenformen befasst. Köller et al. (1999) sind der Vergleichbarkeit von Bewertungsmaßstäben zwischen Gymnasien und Oberstufen an Gesamtschulen auf der Datengrundlage einer Schülerstichprobe in Nordrhein-Westfalen nachgegangen. Ihre Analysen für das Fach Mathematik erbrachten einen deutlichen Leistungsvorsprung zugunsten der Schülerinnen und Schüler am allgemeinbildenden Gymnasium, der jedoch nicht mit Unterschieden in den Fachnoten korrespondierte. Die Fachnoten fielen in den beiden Schulformen mehr oder weniger identisch aus.

Watermann et al. (2004) haben – ebenfalls für Mathematik – die Vergleichbarkeit von Bewertungsmaßstäben an den allgemeinbildenden Gymnasien und den verschiedenen Richtungen der beruflichen Gymnasien in Baden-Württemberg untersucht. Die vorgefundenen substanziellen Leistungsunterschiede zwischen allgemeinbildenden und technischen Gymnasien auf der einen und den übrigen beruflichen Gymnasialformen (wirtschaftswissenschaftlich, ernährungswissenschaftlich, agrarwissenschaftlich und sozialpädagogisch) auf der anderen Seite gingen kaum mit entsprechenden Differenzierungen bei den erteilten Mathematiknoten einher. Die aufgeführten Befunde machen damit insgesamt deutlich, dass auch in der gymnasialen Oberstufe mit Bewertungsunterschieden zu rechnen ist.

2.4 Bessere Vergleichbarkeit durch zentrale Abschlussprüfungen?

Angesichts der eingeschränkten Vergleichbarkeit von Schulnoten und nominell gleichwertigen Abschlusszertifikaten stellt sich die Frage, welche Möglichkeiten bestehen, die zum Teil beträchtlichen Diskrepanzen zwischen Leistungsniveau und Leistungsbeurteilung zu reduzieren. Zur Diskussion stehen hierbei gegenwärtig vor allem die Definition einheitlicher Leistungsanforderungen, wie sie in Form von bundesweit gültigen Bildungsstandards von der Kultusministerkonferenz (KMK) im Oktober 2007 nun auch für das Abitur auf den Weg gebracht wurden, sowie zentrale Abschlussprüfungen. Letztere sind zwar in Form des Zentralabiturs mittlerweile in nahezu allen Bundesländern implementiert, allerdings handelt es sich bislang nur um einheitliche Prüfungen innerhalb der Bundesländer (vgl. im Überblick Klein et al. 2009).

Bezüglich der objektivierenden Wirkung einheitlicher Leistungsanforderungen und zentraler Abschlussprüfungen ist jedoch auch ein gewisses Maß an Skepsis geboten, denn bereits seit den 1970er-Jahren existieren „Einheitliche Prüfungsanforderungen für das Abitur“ (EPA), über deren Implementierung und Wirksamkeit in den einzelnen Bundesländern allerdings kaum aussagekräftige Befunde vorliegen. Dasselbe gilt für die vermutete stärkere Objektivität von Abiturnoten aus zentral vorgegebenen Abschlussprüfungen. Die Erstkorrektur der schriftlichen Prüfungsarbeiten erfolgt i. d. R. durch einen Lehrer an der Schule des betreffenden Schülers, sodass Referenzgruppeneffekte bei der Benotung auch hier nicht vollkommen ausgeschlossen werden können. Zwar sind verschiedene Untersuchungen der Frage nachgegangen, inwieweit zentrale Prüfungen das Leistungsniveau anheben können (vgl. Baumert u. Watermann 2000; Büchel et al. 2003; Wößmann 2005), zur Frage der höheren Leistungsgerechtigkeit zentraler Abiturprüfungen liegen bislang jedoch kaum aussagekräftige Befunde vor (vgl. jedoch Maag Merki 2008).

3 Fragestellung

Im Zentrum der vorliegenden Untersuchung steht die länderübergreifende Betrachtung von Bewertungsmaßstäben beim Abitur. Auf Basis der aus einem Leistungsvergleich zwischen Baden-Württemberger und Hamburger Abiturienten vorliegenden Leistungsdaten und Notenangaben für die Fächer Mathematik und Englisch gehen wir der Frage nach, in welchem Maß sich vorhandene Leistungsunterschiede am Ende der gymnasialen Oberstufe auch in den erteilten Fach- und Prüfungsnoten widerspiegeln. Mit Fach- und Prüfungsnoten sind dabei in der vorliegenden Untersuchung die jeweils erreichten Punktzahlen nach den Vorgaben des Credit-Systems der gymnasialen Oberstufe gemeint. Die zu erreichenden Punktzahlen bewegen sich zwischen 0 und 15 Punkten und sind auch in die herkömmliche Notenskala (1 = sehr gut bis 6 = ungenügend) überführbar.[1] Höhere Punktwerte implizieren dabei bessere Noten. Wenn im Folgenden von „besseren Noten“ die Rede ist, so sind damit also stets höhere Punktzahlen gemeint.

Die bereits publizierten Auswertungen des Leistungsvergleichs zwischen Baden-Württemberg und Hamburg (vgl. Trautwein et al. 2007) haben für Mathematik einen deutlichen Leistungsvorsprung von etwa einer Standardabweichung ($d = 0{,}98$) zugunsten der Baden-Württemberger Gymnasiasten ergeben (Nagy et al. 2007). Hingegen fielen die

mittleren Testleistungen in Englisch in Hamburg und Baden-Württemberg ähnlich aus (Jonkmann et al. 2007). Damit wird bereits unmittelbar einsichtig, dass eine länderübergreifende Betrachtung von Leistungs- und Bewertungsunterschieden in jedem Falle fachspezifisch zu erfolgen hat und Befunde in einer fachlichen Domäne nicht ohne weiteres auf andere Domänen übertragbar sind.

Fachnoten, Bewertungsstrenge und die Bedeutung von Referenzgruppeneffekten. Welche Befunde sind nun vor dem Hintergrund der Leistungsunterschiede zwischen Hamburger und Baden-Württemberger Abiturienten in Bezug auf das mittlere Notenniveau (d. h. die erreichten Punkte auf der 15er-Metrik) und die Bewertungsstrenge zu erwarten? Ausgehend von den in Abschnitt 2.2 dargestellten Befunden aus dem PISA-Ländervergleich für die Gymnasien erwarteten wir für Mathematik in Baden-Württemberg tendenziell ein höheres mittleres Punkteniveau bei den Fachnoten als in Hamburg. Aufgrund der großen Leistungsunterschiede und der vorliegenden Befunde zur Wirksamkeit von Referenzgruppeneffekten bei der Notenvergabe (vgl. Ingenkamp 1995; Trautwein u. Baeriswyl 2007) gehen wir jedoch in Hinblick auf die Bewertungsstrenge davon aus, dass sich die vorhandene Leistungsdifferenz in Mathematik nicht bzw. nur teilweise in den Fachnoten niederschlägt. Entsprechend sollten sich für Hamburger Abiturienten bei Kontrolle der individuellen Mathematikleistung bessere Fachnoten (d. h. höhere Notenpunktzahlen) ergeben als für Baden-Württemberger Gymnasiasten vergleichbaren Leistungsniveaus.

Um zu überprüfen, ob und inwieweit mögliche Bewertungsunterschiede zwischen den Ländern auf Referenzgruppeneffekte bei der Leistungsbewertung zurückzuführen sind, werden wir in den Analysen neben der individuellen Testleistung auch das mittlere Leistungsniveau der Lerngruppe als weiteren Bezugspunkt für die Notenvergabe berücksichtigen. Sofern Referenzgruppeneffekte am Wirken sind, sollte sich dies in negativen Effekten des mittleren Leistungsniveaus auf die Fachnote und in einer Reduktion der nach Kontrolle der individuellen Leistung verbleibenden Notenunterschiede zwischen den Bundesländern zeigen. Im Vergleich zu Mathematik erwarteten wir für Englisch, dass mögliche Diskrepanzen zwischen Testleistung und Fachnoten aufgrund des vergleichbaren mittleren Leistungsniveaus in beiden Ländern – sofern überhaupt nachweisbar – geringer ausfallen sollten.

Zentrale Abiturprüfungen. In einem weiteren Schritt untersuchen wir, ob und in welchem Maß schriftliche Prüfungsnoten aus zentral vorgegebenen Abiturprüfungen eine bessere Vergleichbarkeit zwischen den Bundesländern ermöglichen als Fachnoten. Zwar existiert in Deutschland bislang kein bundeslandübergreifendes Zentralabitur, sodass eine Überprüfung dieser Fragestellung nicht auf der Basis von für Hamburger und Baden-Württemberger Abiturienten identischen Prüfungsaufgaben erfolgen kann; jedoch haben mittlerweile bis auf Rheinland-Pfalz alle Länder zentrale schriftliche Abiturprüfungen auf Bundeslandebene eingeführt, auf die zurückgegriffen werden kann. In Baden-Württemberg hat das Zentralabitur eine lange Tradition und wurde bereits nach dem Zweiten Weltkrieg eingeführt. In Hamburg wurde es erst im Schuljahr 2004/05 implementiert.

Warum sollten landesspezifisch vorgegebene zentrale Abiturprüfungen eine bessere länderübergreifende Vergleichbarkeit von Bewertungsmaßstäben zur Folge haben? Aus unserer Sicht sprechen zwei Punkte dafür. Zum einen lässt sich argumentieren, dass die Prüfungsaufgaben in beiden Bundesländern an den bundesweit gültigen EPA orientiert sind und somit ein vergleichbares Schwierigkeitsniveau aufweisen sollten. Zum ande-

ren gilt es zu berücksichtigen, dass die Prüfungsaufgaben in beiden Bundesländern nicht nur zentral vorgegeben, sondern auch zentrale und vollständig anonymisierte Zweit- und wenn nötig Drittkorrekturen vorgenommen werden. Die Spielräume für klassen- bzw. schulbezogene Referenzgruppeneffekte auf die Benotung sollten entsprechend geringer sein als für die Fachnoten. Dies könnte ebenfalls eine bessere Passung von Testleistung und Prüfungsnote zwischen den Bundesländern bewirken.

4 Methode

4.1 Datengrundlage

Die Daten der vorliegenden Untersuchung entstammen der im Jahr 2002 in Baden-Württemberg durchgeführten Studie „Transformation des Sekundarschulsystems und akademische Karrieren“ (TOSCA; vgl. Köller et al., 2004) und der im Jahr 2005 in Hamburg durchgeführten Studie „Aspekte der Lernausgangslage und Lernentwicklung – Jahrgangsstufe 13“ (LAU-13; vgl. Lehmann et al. 2006; Trautwein et al., 2007). In beiden Studien wurden Gymnasiasten am Ende der 13. Jahrgangsstufe untersucht, die in weiten Teilen identische Erhebungsinstrumente bearbeiteten. Bei der LAU-13-Studie handelt es sich um eine Vollerhebung ($N=5.507$ aus 108 Schulen) an allen gymnasialen Oberstufen Hamburgs. Für Baden-Württemberg wurde eine repräsentative Stichprobe von $N=5.775$ Schülerinnen und Schülern (aus 149 Schulen) aus den allgemeinbildenden Gymnasien und den verschiedenen Richtungen der in Baden-Württemberg vorhandenen beruflichen Gymnasien gezogen, die durch Gewichtung an die tatsächliche Populationsverteilung aus dem Schuljahr 2001/02 angepasst wurde (vgl. Lüdtke et al. 2007a). In der vorliegenden Untersuchung beschränken wir uns auf die allgemeinbildenden Gymnasien, die in Hamburg 67,8% und in Baden-Württemberg 68,9% der Abiturienten binden. Dieses Vorgehen hat den Vorteil, Länder- und Schulformunterschiede bei der Leistungsbewertung zu entmischen. Tabelle 1 gibt Auskunft über die Verteilung der Schülerinnen und Schüler auf die Grund- und Leistungskurse in Mathematik und Englisch. Wie ersichtlich wird, fiel der Anteil der Mathematik-Leistungskursschüler in Baden-Württemberg fast doppelt so hoch aus wie in Hamburg. Vertiefende Analysen im Rahmen der TOSCA-LAU-Untersuchung ergaben jedoch, dass die Mathematikleistungsunterschiede zwischen den Hamburger und Baden-Württemberger Abiturienten nur zu einem geringen Teil auf den höheren Leistungskursanteil in Baden-Württemberg zurückgeführt werden können (Nagy et al. 2007).

4.2 Erhebungsinstrumente

Leistungstests in Mathematik und Englisch. Bei der Frage nach der Vergleichbarkeit von Bewertungsmaßstäben kommt den eingesetzten Leistungstests eine zentrale Rolle zu. Sie stellen gewissermaßen das „objektive“ Kriterium für die Untersuchung von Unterschieden in der Leistungsbewertung dar. Entsprechend voraussetzungsreich ist ihr Einsatz, da implizit davon ausgegangen wird, dass die Testinhalte in den zu vergleichenden Gruppen, also an den Hamburger und Baden-Württemberger Oberstufen, in hinreichendem und

Tab. 1: Schülerinnen und Schüler an allgemeinbildenden Gymnasien nach Bundesland und Kursniveau in Mathematik und Englisch (in Klammern Angaben in Prozent)

		Mathematik		Englisch		
	Schüler gesamt[1]	**Grundkurs**	**Leistungskurs**	**Grundkurs**	**Leistungskurs**	**abgewählt**
Baden-Württemberg	3526	2199 (64,4)	1217 (35,6)	1602 (46,2)	1362 (39,2)	507 (14,6)
Hamburg	3734	3058 (81,9)	676 (18,1)	2255 (60,4)	1237 (33,1)	242 (6,5)

[1] Die Angaben zur Kurszugehörigkeit in Baden-Württemberg addieren sich aufgrund zum Teil fehlender Angaben nicht immer zu den Gesamtschülerzahlen auf.

vergleichbarem Maße unterrichts- und prüfungsrelevant sind. Nur unter dieser Annahme lassen sich belastbare Aussagen zu Bewertungsunterschieden treffen.

Die Leistungen in *Mathematik* wurden mit dem Test zur voruniversitären Mathematik aus der *Third International Mathematics and Science Study* (TIMSS/III; vgl. Baumert et al. 2000; Nagy et al. 2007) erfasst. Der Test umfasste insgesamt 68 Aufgaben aus den Stoffgebieten Zahlen/Gleichungen/Funktionen, Analysis, Geometrie, Aussagenlogik/Beweise und Wahrscheinlichkeitsrechnung/Statistik. Sowohl in TOSCA als auch in LAU-13 kamen vier unterschiedliche Testhefte zum Einsatz, in denen sechs Aufgabencluster systematisch rotiert wurden (*Multi-Matrix-Design*, vgl. Watermann et al. 2004). Der Test wurde auf Grundlage der Item-Response-Theory (IRT) unter Rückgriff auf die Plausible-Value-Technik skaliert (zum genauen Vorgehen bei der Testskalierung siehe Nagy et al. 2007).

Die Testkonzeption des TIMSS-Oberstufentests sah ausdrücklich einen starken Bezug zum Oberstufencurriculum der Teilnehmerländer vor. Die Lehrplan- und Unterrichtsvalidität des Tests wurde in TIMSS mittels Lehrplananalysen, Experten- und Fachleiterbefragungen geprüft (zu Einzelheiten der Testvalidierung vgl. Klieme 2000). Es zeigte sich, dass die Testaufgaben sowohl die Lehrpläne der Grund- und Leistungskurse als auch die Unterrichtsinhalte (realisiertes Curriculum) in hohem Maße widerspiegelten. Die positiven Validitätsbelege konnten in TOSCA und LAU-13 auf der Grundlage erneuter Experten- und Fachleiterbefragungen repliziert werden (vgl. Nagy et al. 2007). Auch die um mögliche Referenzgruppeneffekte bei der Benotung bereinigten Zusammenhänge zwischen Testleistungen und den Fachnoten für das Schulhalbjahr 13/1 (jeweils getrennt nach Kursniveau innerhalb der Schulen standardisiert[2]) arrondierten dieses Bild. Die resultierenden Korrelationen fielen sowohl in Hamburg (GK: $r=0{,}62$, LK: $r=0{,}67$) als auch in Baden-Württemberg (GK: $r=0{,}52$, LK $r=0{,}60$) substanziell aus. Zusammengenommen kann damit festgehalten werden, dass der TIMSS-Test zur voruniversitären Mathematik in beiden Bundesländern eine hohe Validität aufwies und damit in hinreichendem Maß zur Untersuchung von Bewertungsunterschieden geeignet ist.

Zur Erfassung der Leistungen in *Englisch* diente in TOSCA und LAU-13 eine validierte Kurzform des Test of English as a Foreign Language (TOEFL, vgl. Jonkmann et al. 2007). Der vom *Educational Testing Service* (ETS) in Princeton, New Jersey, entwickelte TOEFL umfasst die Bereiche Hörverstehen, Grammatik und Orthografie sowie Wortschatz und Leseverständnis und wird insbesondere von amerikanischen Universitäten

dazu verwendet, die Englisch-Fähigkeiten von Studienbewerbern, deren Muttersprache nicht Englisch ist, auf ein ausreichendes Niveau hin zu überprüfen.

Anders als der TIMSS-Test zur voruniversitären Mathematik erhebt der TOEFL von seiner Konzeption her keinen direkten Anspruch auf curriculare Validität bezüglich der Inhalte des Englisch-Oberstufenunterrichts. Mit Bezug auf die Lehrpläne der einzelnen Bundesländer sowie die EPA für das Fach Englisch kann jedoch davon ausgegangen werden, dass die Förderung der mit dem TOEFL erfassten rezeptiv-kommunikativen Fremdsprachenkompetenzen in den Zielkorridor des Englischunterrichts in der Oberstufe fällt (vgl. Köller u. Trautwein 2004). Die um mögliche Referenzgruppeneffekte bereinigten Korrelationen zwischen den TOEFL-Leistungen und Englisch-Fachnoten im Schulhalbjahr 13/1 (Hamburg GK: $r=0{,}52$, LK: $r=0{,}53$; Baden-Württemberg GK: $r=0{,}54$, LK: $r=0{,}58$) fielen nur leicht niedriger aus als die entsprechenden Korrelationen in Mathematik, was ebenfalls als Beleg für den hinreichenden Unterrichtsbezug des eingesetzten TOEFL gewertet werden kann (vgl. Jonkmann et al. 2007).

Fach- und Prüfungsnoten. Neben den Testleistungen in Mathematik und Englisch wurden auch die Fachnoten für das Schulhalbjahr 13/1 (ausgedrückt in Punkten von 0 bis 15), die Abiturprüfungsnoten (ebenfalls in Punkten) und die Abiturgesamtnoten (herkömmliche 6er-Notenmetrik) der Schülerinnen und Schüler erhoben. Dabei konnten wir auf die Angaben aus den Schülerakten zurückgreifen. Einschränkend ist hierbei anzumerken, dass uns aus Baden-Württemberg die schriftlichen Prüfungsnoten nicht einzeln vorlagen, sondern lediglich die kombinierte Angabe aus schriftlicher Prüfung und mündlicher Nachprüfung (sofern abgelegt). Wie jedoch Kontrollanalysen mit den Hamburger Daten zeigten, korrelierten die schriftlichen Prüfungsnoten mit den kombinierten (schriftlich + mündlich) Prüfungsnoten über alle Hamburger Schülerinnen und Schüler nahezu perfekt miteinander ($r>0{,}99$), sodass nicht mit bedeutsamen Verzerrungen zu rechnen ist.

4.3 Umgang mit fehlenden Werten

Dem Umgang mit fehlenden Werten kommt in der vorliegenden Untersuchung eine besondere Bedeutung zu, da die Teilnahmequoten in Hamburg (95%, Vollerhebung mit verpflichtender Teilnahme) und Baden-Württemberg (80,2%) unterschiedlich hoch ausfielen (vgl. Lüdtke et al. 2007a). Um möglichen systematischen Verzerrungen durch den partiellen Stichprobenausfall vorzubeugen, wurden fehlende Leistungswerte in Mathematik und Englisch unter Zuhilfenahme von leistungsrelevanten und für nahezu alle Schülerinnen und Schüler verfügbaren Hintergrundinformationen (z. B. Noten, Kursbelegung, Leistungsmittelwerte der Schulen) geschätzt. Im Falle der Mathematikleistungen erfolgte dies im Rahmen der Testskalierung (*Plausible-Value*-Technik, zum genauen Vorgehen vgl. Nagy et al. 2007). Fehlende Werte bei den Englischleistungen sowie den Fach- und Prüfungsnoten wurden durch das *Multiple Imputation*-Verfahren (vgl. Lüdtke et al. 2007a, 2007b) ersetzt. Soweit nicht anders ausgewiesen, basieren alle Analysen auf den kombinierten Ergebnissen aus fünf imputierten Datensätzen.

4.4 Statistische Analysen

In den zentralen Analysen zu Bewertungsunterschieden zwischen Baden-Württemberg und Hamburg haben wir verschiedene Vorhersagemodelle für die Fach- und Prüfungsnoten spezifiziert. Da die Daten eine Mehrebenenstruktur aufweisen (Schülerinnen und Schüler sind geschachtelt in Kursen und Schulen) und aufgrund möglicher schulkontextueller Einflüsse nicht von unabhängigen Beobachtungen ausgegangen werden kann, wurden Mehrebenenanalysen durchgeführt (vgl. Raudenbush u. Bryk 2002). Es wurden lineare Mehrebenenmodelle mit dem Programmpaket Mplus 4.2 (Muthén u. Muthén 1998–2008) spezifiziert, in denen als Prädiktoren die individuelle Testleistung auf Schülerebene (Level 1) und die Bundeslandzugehörigkeit sowie die aggregierte Testleistung auf Schulebene (Level 2) aufgenommen wurden.

Für die Mehrebenenanalysen wurden die individuellen Testleistungen der Schülerinnen und Schüler in Mathematik und Englisch jeweils am Testwert von 500 Punkten zentriert und anschließend auf Basis der Streuung in der Untersuchungsstichprobe standardisiert ($x' = x - 500/SD$). Im TIMSS-Test stellt der Wert von 500 Punkten den internationalen Mittelwert dar (Baumert et al. 2000). Im TOEFL-Test ist ein Wert von 500 Punkten für Bewerber mit nichtenglischer Muttersprache Voraussetzung, um an einer prestigeärmeren amerikanischen Universität zum Studium zugelassen zu werden (vgl. Jonkmann et al. 2007). Die vorgenommene Zentrierung ermöglicht somit die inhaltliche Interpretation der aus den Mehrebenenanalysen resultierenden Achsenabschnittsparameter. Die Analysen geben Auskunft darüber, mit welcher Note die Leistung eines Baden-Württemberger Schülers mit 500 Testpunkten bewertet wurde und wie die entsprechende Note bei gleicher Testleistung in Hamburg ausfällt (vgl. Abschnitte 5.2 und 5.3). Für die Analyse des Einflusses der mittleren Leistungsstärke der Referenzgruppe auf die Notenvergabe wurden die individuellen Testleistungen der Grund- und Leistungskursschüler jeweils getrennt auf Schulebene aggregiert.

5 Ergebnisse

5.1 Deskriptive Befunde

Tabelle 2 gibt einen Überblick über die an den allgemeinbildenden Gymnasien in Baden-Württemberg und Hamburg erreichten Testleistungen in Mathematik und Englisch sowie die entsprechenden Fachnoten (ausgedrückt in Punkten auf der 15er-Punktemetrik) für das Schulhalbjahr 13/1. Die Grund- bzw. Leistungskurszugehörigkeit in Mathematik und Englisch bleibt zunächst unberücksichtigt. Ergänzend sind die erzielten Abiturgesamtnoten (6er-Notenmetrik) in beiden Bundesländern aufgeführt. Die dargestellten Effektstärken geben die Bundeslandunterschiede jeweils ausgedrückt in Standardabweichungen an und ermöglichen dadurch eine bessere Vergleichbarkeit der auf den unterschiedlichen Metriken abgetragenen Leistungsindikatoren (Tab. 2).

Die mittlere Abiturgesamtnote an den allgemeinbildenden Gymnasien fiel in Baden-Württemberg besser aus als in Hamburg. Die Differenz zwischen beiden Bundesländern entsprach einem Unterschied von etwa einer fünftel Standardabweichung. Die mittleren

Tab. 2: Abiturgesamtnote, Fachnote Schulhalbjahr 13/1 (in Punkten) und Fachleistung Mathematik und Englisch an allgemeinbildenden Gymnasien, getrennt nach Bundesland (Mittelwerte, Standardabweichungen und Effektstärken)

	Abiturgesamtnote	**Fachnote Mathematik (in Punkten)**	**Testleistung Mathematik**	**Fachnote Englisch[1] (in Punkten)**	**Testleistung Englisch**
Baden-Württemberg	2,38 (0,67)	8,18 (3,64)	501,90 (79,45)	8,99 (2,88)	525,74 (50,71)
Hamburg	2,52 (0,64)	7,94 (3,48)	424,32 (75,27)	8,63 (2,86)	521,03 (56,60)
Effektstärke d	*0,211*[*,2]	*0,066*	*1,002*[*]	*0,126*[*]	*0,088*

[*] Mit der Effektstärke korrespondierender Mittelwertvergleich signifikant auf dem $p<0{,}05$-Niveau.

[1] Nur für Schüler, die Englisch im Grund- bzw. Leistungskurs belegt haben (TOSCA $N=2.982$, LAU-13 $N=3.492$).

[2] Effektstärke umgepolt: positive Effektstärke zugunsten kleinerer Werte auf der Abiturgesamtnote.

Mathematikfachnoten fielen wie erwartet in Baden-Württemberg etwas besser aus, der Punkteunterschied erwies sich allerdings nicht als statistisch signifikant. Angesichts der in der nächsten Spalte aufgeführten Unterschiede in den erzielten Mathematikleistungen von mehr als einer Standardabweichung ist dieser deskriptive Befund bereits ein deutlicher Beleg für die beschränkte Aussagekraft der Fachnoten. In Englisch fanden sich etwas bessere Noten für die Baden-Württemberger Gymnasiasten, obwohl die Testleistungen nicht statistisch signifikant von denen der Hamburger Gymnasiasten abwichen.

Tabelle 3 stellt die in Baden-Württemberg und Hamburg erreichten Testleistungen und Fachnoten, getrennt nach Kurszugehörigkeit, gegenüber. Für Mathematik fand sich nach Berücksichtigung der Kurszugehörigkeit in beiden Bundesländern ein nahezu identisches Notenniveau. Die Testleistungsunterschiede zwischen Hamburger und Baden-Württemberger Gymnasiasten fielen im Leistungskurs etwas geringer aus als im Grundkurs. Im Fach Englisch korrespondieren Testleistungen und Fachnoten innerhalb desselben Kursniveaus sehr gut miteinander (vgl. Tab. 3).

5.2 Vorhersage der Fachnoten in Mathematik und Englisch und Referenzgruppeneffekte

Nach den Befunden aus den Tabellen 2 und 3 sind vor allem für den Bereich Mathematik Diskrepanzen zwischen erzielten Testleistungen und den vergebenen Fachnoten sichtbar geworden. Um das Ausmaß dieser Abweichungen zu veranschaulichen und der Rolle von Referenzgruppeneffekten bei der Notenvergabe nachzugehen, haben wir mehrere Mehrebenenmodelle zur Vorhersage der Mathematik- und Englischfachnoten spezifiziert, in denen sukzessive die Bundeslandzugehörigkeit (*Frage der Notenunterschiede*), die individuelle Schülerleistung (*Frage der Bewertungsstrenge*) und die aggregierte Testleistung der Grund- bzw. Leistungskursschüler der Schule (*Frage der Referenzgruppeneffekte*) aufgenommen wurden (vgl. Tab. 4).

In Modell 1 für den Mathematikgrund- und -leistungskurs zeigen sich noch einmal die bereits deskriptiv berichteten, nicht vorhandenen Bundeslandunterschiede in den mittleren Fachnoten. Baden-Württemberger und Hamburger Gymnasiasten erhielten in Mathematik im Mittel identische Fachnoten. Um Unterschiede in der Bewertungsstrenge zu untersu-

Tab. 3: Fachnote Schulhalbjahr 13/1 (in Punkten) und Fachleistung Mathematik und Englisch an allgemeinbildenden Gymnasien, getrennt nach Bundesland und Kursniveau (Mittelwerte, Standardabweichungen und Effektstärken)

		Grundkurs		Leistungskurs	
		Fachnote (in Punkten)	**Testleistung**	**Fachnote (in Punkten)**	**Testleistung**
Mathematik	Baden-Württemberg	7,57 (3,68)	466,47 (61,59)	9,33 (3,28)	565,60 (68,64)
	Hamburg	7,62 (3,46)	404,46 (61,23)	9,37 (3,20)	514,17 (66,89)
	Effektstärke d	*−0,016*	*1,010*[*]	*−0,013*	*0,759*[*]
Englisch	Baden-Württemberg	8,80 (2,92)	517,38 (46,63)	9,22 (2,82)	545,84 (46,55)
	Hamburg	8,02 (2,77)	505,77 (52,06)	9,74 (2,70)	556,46 (46,45)
	Effektstärke d	*0,273*[*]	*0,235*[*]	*−0,191*[*]	*−0,228*[*]

[*] Mit der Effektstärke korrespondierender Mittelwertvergleich signifikant auf dem $p < 0{,}05$-Niveau.

chen, wurde in Modell 2 als nächstes die individuelle Mathematikleistung aufgenommen. Der Intercept von 8,69 Punkten für den Grundkurs bzw. 7,38 Punkten für den Leistungskurs lässt sich aufgrund der vorgenommenen Zentrierung (vgl. Abschnitt 4.4) interpretieren als die Mathematiknote (auf der 15er-Punktemetrik) eines Baden-Württemberger Schülers, der im Mathematiktest ein Ergebnis von 500 Punkten (internationaler Mittelwert im TIMSS-Test) erreicht hat. Lag ein Schüler eine Standardabweichung (87,38 Testleistungspunkte) über dem TIMSS-Mittelwert, fiel seine Note um 2,94 (GK) bzw. 2,64 (LK) Punkte höher aus. Von zentraler Bedeutung für die Frage der Bewertungsstrenge sind die nach Kontrolle der individuellen Mathematikleistung verbleibenden Regressionskoeffizienten für die Bundeslandzugehörigkeit. Bei vergleichbaren Leistungen, in unserem Fall 500 Punkten im Mathematiktest, erhielt ein Hamburger Grundkursschüler eine um 2,16 Punkte höhere (10,85 vs. 8,69 Punkte) und ein Hamburger Leistungskursschüler eine um 1,57 Punkte höhere (8,95 Punkte vs. 7,38 Punkte) Mathematiknote. Diese aus der Bundeslandzugehörigkeit resultierenden Effekte auf die Fachnoten entsprachen einer Größenordnung von $d = 0{,}60$ (GK) bzw. $d = 0{,}48$ (LK) Standardabweichungen und sind als sehr bedeutsam einzustufen.

Für die Untersuchung möglicher Referenzgruppeneffekte bei der Leistungsbewertung wurde im nächsten Schritt auf der Aggregatsebene die mittlere Mathematikleistung der Grund- bzw. Leistungskursschüler einer Schule aufgenommen. Wie Modell 3 entnommen werden kann, hatte die mittlere Leistungsstärke erwartungsgemäß einen negativen Effekt auf die Fachnoten. Bei gleichen individuellen Leistungen erhielten Schüler in leistungsstärkeren Schulen schlechtere Noten (d. h. niedrigere Punktzahlen). Die Bundeslandeffekte gingen durch die Berücksichtigung des mittleren Leistungsniveaus um etwa die Hälfte (GK) bzw. ein Drittel (LK) zurück. Referenzgruppeneffekte bei der Notenvergabe trugen damit entscheidend zu den Bewertungsunterschieden zwischen beiden Bundesländern bei, erklärten diese aber nicht vollständig.

Den ebenfalls in Tabelle 4 enthaltenen Befunden für das Fach Englisch kann entnommen werden, dass sich nach Berücksichtigung der individuellen Testleistungen (Modell 2) nur für den Grundkurs Unterschiede in der Bewertungsstrenge zeigten. Bei vergleichbaren Leistungen erhielten Hamburger Grundkursschüler etwas schlechtere Noten. Der Unterschied fiel jedoch wesentlich geringer aus als in Mathematik. Weiterhin fanden sich

Tab. 4: Vorhersagemodelle für die Fachnote Schulhalbjahr 13/1 (in Punkten) am allgemeinbildenden Gymnasium in Mathematik und Englisch durch individuelle Schülerleistung, Bundeslandzugehörigkeit und mittlere Leistungsstärke der Schule, getrennt nach Grund- und Leistungskurs (Befunde aus Mehrebenenanalysen)

	Mathematik						**Englisch**					
	Grundkurs			**Leistungskurs**			**Grundkurs**			**Leistungskurs**		
	M1	**M2**	**M3**	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**
Intercept	**7,57**[1]	**8,69**	**8,17**	**9,31**	**7,38**	**8,15**	**8,79**	**8,42**	**8,57**	**9,23**	**7,92**	**8,36**
Individuelle Schülerleistung[2]	–	**2,94**	**2,99**	–	**2,64**	**2,75**	–	**1,79**	**1,82**	–	**1,91**	**1,94**
R^2 (innerhalb Schule)	–	*39,6*	*40,4*	–	*44,0*	*46,2*	–	*29,6*	*30,3*	–	*31,3*	*32,1*
Bundesland (Referenz: BW)	0,02	**2,16**	**1,17**	0,05	**1,57**	**0,96**	**−0,81**	**−0,41**	**−0,57**	**0,47**	0,21	**0,30**
Mittlere Schülerleistung	–	–	**−1,40**	–	–	**−1,17**	–	–	**−0,75**	–	–	**−0,68**
R^2 (zwischen Schulen)[3]	–	*66,0*	*73,8*	–	*44,0*	*57,5*	–	*7,1*	*16,0*	–	*2,1*	*9,3*

[1] Auf dem $p<0,05$-Niveau signifikante Parameter fett.

[2] Individuelle Schülerleistung in Mathematik und Englisch an 500 Punkten zentriert und auf Basis der Stichprobenstreuung standardisiert.

[3] Inkrementelles R^2 nach Kontrolle der individuellen Schülerleistung.

auch für Englisch Effekte der mittleren Leistungsstärke auf die Fachnote. Anders als in Mathematik führte die Aufnahme der mittleren Leistung jedoch nicht zu einer Verringerung der Bundeslandunterschiede, sondern tendenziell zu einer Vergrößerung. Die mittlere Leistung scheint damit in erster Linie Bewertungsunterschiede zwischen den Schulen innerhalb der Bundesländer zu erklären, nicht jedoch zwischen den Bundesländern.

5.3 Vergleichbarkeit von Prüfungsnoten

Zur Prüfung der Frage der besseren Vergleichbarkeit von Prüfungsnoten aus zentralen Abiturprüfungen haben wir in Analogie zu den in Tabelle 4 aufgeführten Analysen für die Fachnoten Vorhersagemodelle für die schriftlichen Prüfungsnoten spezifiziert (vgl. Tab. 5). Die Ergebnisse aus den Tabellen 4 und 5 sind direkt miteinander vergleichbar, da jeweils die gleichen Prädiktoren berücksichtigt wurden. Wir beschränken uns in unseren Analysen auf die Leistungskursschüler, da hier die schriftliche Prüfung für alle Schüler verpflichtend ist und somit am ehesten von repräsentativen Befunden ausgegangen werden kann. In Modell 1 für die Prüfungsnoten in Mathematik findet sich bereits ein deutlicher Unterschied zu den Ergebnissen für die Fachnoten. Fielen die Fachnoten in beiden Bundesländern praktisch identisch aus, erhielten Hamburger Abiturienten im Mittel schlechtere Prüfungsnoten als die Abiturienten in Baden-Württemberg. Die Prüfungsnoten korrespondierten damit wesentlich besser mit den erfassten Testleistungen; allerdings nicht vollständig, wie sich an Modell 2 erkennen lässt. Nach Kontrolle der individuellen Schülerleistung fielen die Prüfungsnoten in Hamburg im Mittel immer noch um 0,46 Punkte besser aus. Der entsprechende Unterschied bei den Fachnoten (vgl. Tab. 4) fiel mit 1,57 Punkten jedoch wesentlich höher aus.

Auch bei den Prüfungsnoten fanden sich Hinweise auf referenzgruppenspezifische Bewertungsmaßstäbe (vgl. Modell 3). Bei vergleichbarer individueller Leistung hatte die aggregierte Schülerleistung einen negativen Effekt auf die Prüfungsnote. Der in Modell 2 ausgewiesene Bundeslandeffekt ist nach Berücksichtigung des mittleren Leistungs-

Tab. 5: Vorhersagemodelle für die schriftliche Prüfungsnote (in Punkten) in Leistungskursen am allgemeinbildenden Gymnasium in Mathematik und Englisch durch individuelle Schülerleistung, Bundeslandzugehörigkeit und mittlere Leistungsstärke der Schule (Befunde aus Mehrebenenanalysen)

	Mathematik			Englisch		
	M1	**M2**	**M3**	**M1**	**M2**	**M3**
intercept	**9,97**[1]	**7,66**	**7,98**	**9,13**	**7,38**	**7,99**
Individuelle Schülerleistung [2]		**3,06**	**3,12**		**2,23**	**2,28**
R² (innerhalb Schule)		*45,3*	*46,2*		*32,5*	*33,6*
Bundesland (Referenz: BW)	**−1,35**	**0,46**	0,21	0,12	−0,20	−0,09
Mittlere Schülerleistung			**−0,48**			**−0,83**
R² (zwischen Schulen)[3]		*7,6*	*13,3*		*2,3*	*14,4*

[1] Auf dem $p < 0{,}05$-Niveau signifikante Parameter fett.

[2] Individuelle Schülerleistung in Mathematik und Englisch an 500 Punkten zentriert und auf Basis der Stichprobenstreuung standardisiert.

[3] Inkrementelles R² nach Kontrolle der individuellen Schülerleistung.

niveaus nicht mehr statistisch signifikant und scheint somit in erster Linie das Ergebnis von Referenzgruppeneffekten bei der Leistungsbewertung darzustellen. Insgesamt scheinen – zumindest für Mathematik – die landesspezifischen Zentralprüfungen damit in gewissem Maß auch eine Annäherung länderübergreifender Bewertungsmaßstäbe zu bewirken.

Für die Prüfungsnoten in Englisch fanden sich in Abweichung zu den Fachnoten keine Bundeslandunterschiede im mittleren Notenniveau. Die Prüfungsnoten korrespondierten damit in etwas geringerem Maß mit den erfassten Englischleistungen als die Fachnoten. Nach Kontrolle der individuellen Englischleistung (vgl. Modell 2) ließ sich jedoch, wie bei den Fachnoten, kein statistisch signifikanter Effekt der Bundeslandzugehörigkeit nachweisen.

5.4 Praktische Implikationen eingeschränkt aussagekräftiger Abiturnoten und hypothetische Konsequenzen bei der Studienzulassung

Die bisherigen Analysen haben gezeigt, dass Bedenken hinsichtlich der Vergleichbarkeit von Abiturleistungen zwischen den Bundesländern durchaus berechtigt sind. Abschließend wollen wir mögliche praktische Konsequenzen dieser eingeschränkten Vergleichbarkeit für die Abnehmerseite (Hochschulen und Ausbildungsbetriebe) und die Abiturienten verdeutlichen.

Wenngleich Schulnoten, wie eingangs geschildert, eine Vielzahl an Funktionen übernehmen, dienen sie der *Abnehmerseite* in erste Linie als Selektionskriterium für die Auswahl von Bewerbern um einen Studien- oder Ausbildungsplatz. Von den im Abschlusszeugnis aufgeführten Noten wird dabei implizit auf dahinterstehende Kompetenzen in einem oder mehreren Fächern geschlossen. Dieser Schluss setzt jedoch starke Annahmen zur Vergleichbarkeit der Noten über unterschiedliche Schülergruppen hinweg voraus.

Abbildung 2 gibt Auskunft darüber, von welchen Unterschieden in den erfassten Mathematik- und Englischkompetenzen bei Baden-Württemberger und Hamburger Abiturienten mit vergleichbaren Fachnoten auszugehen wäre. Wie die Befunde für Englisch zeigen, sind die hinter den einzelnen Notenbereichen stehenden Englischleistungen weitgehend vergleichbar. Ein anderes Bild zeigt sich für Mathematik. Über den gesamten Notenbereich hinweg liegen die Mathematikleistungen der Baden-Württemberger Gymnasiasten deutlich über dem der Hamburger Abiturienten. Gemessen an der Streuung in der Gesamtstichprobe reichen die Leistungsdifferenzen bis an eine Standardabweichung heran. Eine leistungsbezogene Vergleichbarkeit ist hier damit kaum noch gegeben.

Die in Abb. 2 dargelegten Diskrepanzen zwischen Noten und Fachleistungen haben für die *Bewerberseite,* je nachdem, welche Merkmale – Noten oder standardisierte Leistungstests – zu Auswahlzwecken eingesetzt werden, unterschiedliche Konsequenzen. Um diese aufzuzeigen, haben wir auf Grundlage der vorliegenden Daten verschiedene hypothetische Szenarien der Studienzulassung an einer fiktiven Hochschule simuliert und dabei unterschiedliche Zulassungskriterien angesetzt. Neben der Abiturgesamtnote werden die Fachnoten und Testleistungen in Mathematik und Englisch berücksichtigt. Für unser Szenario haben wir die Stichprobe der Baden-Württemberger Abiturienten durch eine nachträgliche Gewichtung an die Stichprobengröße der Hamburger Abiturienten angepasst. Verglichen werden somit zwei Bundesländer mit gleicher Anzahl von Studienbewerbern

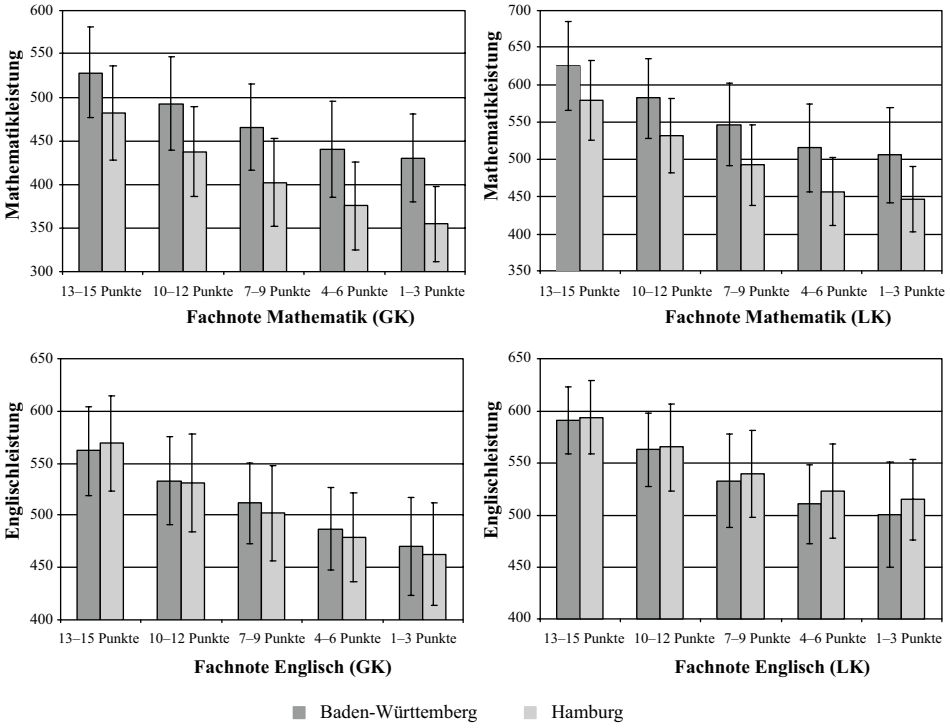


Abb. 2: Fachnoten Schulhalbjahr 13/1 (in Punkten) und korrespondierende mittlere Fachleistung in Mathematik und Englisch an allgemeinbildenden Gymnasien nach Bundesland und Kursniveau (Mittelwerte ± eine Standardabweichung, Berechnungsgrundlage: erster imputierter Datensatz)

aus allgemeinbildenden Gymnasien. Unser Szenario weicht insoweit von den aktuellen Regelungen des Hochschulrahmengesetzes ab, als dass wir zum einen davon ausgehen, dass die Hochschule die Gesamtheit der zukünftigen Studierenden eigenständig auswählt und nicht nur die nach ZVS-Zulassung (Numerus clausus unter Heranziehung der Abiturgesamtnote) und Wartezeit verbleibenden 60%. Zum anderen sieht das Szenario keine Kombination verschiedener Leistungsindikatoren vor, sondern betrachtet die verschiedenen Zulassungskriterien jeweils separat. Die im Folgenden berichteten Ergebnisse erheben daher keineswegs den Anspruch eines absolut realistischen Szenarios. Sie sind jedoch durchaus geeignet, mögliche Auswirkungen der gestärkten Autonomie der Hochschulen bei der Bewerberauswahl zu verdeutlichen, die bei der Studierendenauswahl zukünftig in stärkerem Maß von standardisierten Studieneingangstests Gebrauch machen werden.

Tabelle 6 enthält die auf Baden-Württemberg und Hamburg entfallenden prozentualen Anteile der zum Studium zugelassenen Schüler in Abhängigkeit vom angesetzten Zulassungskriterium. Bei Heranziehung der Abiturgesamtnote würde der Anteil der Baden-Württemberger Gymnasiasten unter allen zugelassenen Bewerbern je nach gewähltem Cut-Off-Kriterium (die besten 10%, 20% bzw. 30%) zwischen 56,3 und 61,0% betragen und damit zwischen 1,29- und 1,57-mal so hoch ausfallen wie der Anteil der Hamburger Gymnasiasten. Die ausschließliche Zugrundelegung der Mathematikfachnote würde zu

Tab. 6: Hypothetisches Szenario der Studienzulassung unter Zugrundelegung unterschiedlicher Zulassungskriterien (prozentuale Anteile der zugelassenen Baden-Württemberger und Hamburger Abiturienten allgemeinbildender Gymnasien bei angenommener gleicher Populationsgröße, Berechnungsgrundlage: erster imputierter Datensatz)

Zulassungskriterium	**Die besten …**	**ca. 10%**	**ca. 20%**	**ca. 30%**
Abiturgesamtnote	BW	61,0%	56,8%	56,3%
	HH	39,0%	43,2%	43,7%
		1,57	*1,31*	*1,29*
Fachnote Mathematik	BW	55,0%	53,9%	54,0%
	HH	45,0%	46,1%	46,0%
		1,22	*1,17*	*1,18*
Mathematikleistung	BW	84,1%	78,5%	76,0%
	HH	15,9%	21,5%	24,0%
		5,28	*3,65*	*3,16*
Fachnote Englisch (ohne abgewählt)	BW	51,9%	51,5%	50,7%
	HH	48,1%	48,5%	49,3%
		1,08	*1,06*	*1,03*
Englischleistung	BW	46,0%	48,1%	50,3%
	HH	54,0%	51,9%	49,7%
		0,85	*0,93*	*1,01*

etwas höheren Anteilen der Hamburger Gymnasiasten führen als die Abiturgesamtnote. Gravierende Unterschiede ergäben sich, wenn die im TIMSS-Oberstufentest erreichten Mathematikleistungen – etwa in Form eines fachspezifischen Studieneingangstests – als ausschließliches Zulassungskriterium herangezogen würden. Bei Beschränkung auf die besten 10% in den Mathematikleistungen würde die Gruppe der zugelassenen Bewerber zu fast 85% aus Baden-Württemberger Gymnasiasten bestehen und damit mehr als fünfmal so hoch ausfallen wie der Anteil der Hamburger Gymnasiasten. Auch für die besten 20 bzw. 30% im Mathematiktest würden die Baden-Württemberger Gymnasiasten mehr als drei Viertel der zugelassenen Bewerber ausmachen. Ein ganz anderes Bild ergäbe sich bei ausschließlicher Beschränkung auf die Englischfachnoten bzw. die im TOEFL-Test erbrachten Leistungen, aus denen zum Teil höhere Zulassungsquoten für die Hamburger Gymnasiasten resultieren würden. Insgesamt wird deutlich, dass die zukünftig deutlich erweiterten Spielräume der Hochschulen bei der Studienzulassung je nach zugrunde gelegtem Zulassungskriterium sehr unterschiedliche Konsequenzen nach sich ziehen können (Tab. 6).

6 Diskussion

Im Zentrum des vorliegenden Beitrags stand die länderübergreifende Untersuchung von Leistungs- und Bewertungsunterschieden zwischen Hamburger und Baden-Württemberger Abiturienten. Auf der Basis standardisierter Leistungstests wurden Befunde zur Vergleichbarkeit von Noten aus dem Abiturzeugnis von Schülerinnen und Schülern aus

unterschiedlichen Bundesländern vorgelegt, wobei wir uns auf die Fach- und Prüfungsnoten in Mathematik und Englisch beschränkt haben.

Während sich für Englisch nur geringe Unterschiede in der Bewertungsstrenge zeigten, ergaben unsere Analysen für Mathematik deutliche Bewertungsunterschiede zwischen den Bundesländern. Leistungsunterschieden in der Größenordnung von einer drei Viertel Standardabweichung im Leistungskurs und einer ganzen Standardabweichung im Grundkurs standen jeweils nahezu identische mittlere Fachnoten gegenüber. Die Leistungsunterschiede in Mathematik spiegelten sich damit nicht in erwartbarer Weise in den erteilten Fachnoten wider. Folglich erhielten Schülerinnen und Schüler bei gleichem Leistungsniveau je nach Bundeslandzugehörigkeit unterschiedliche Fachnoten. Der Punktevorteil auf der 15er-Punktemetrik für die Hamburger Abiturienten betrug etwa 1,5 Punkte im Leistungskurs und mehr als 2 Punkte im Grundkurs. Auf der herkömmlichen 6er-Notenmetrik entspricht dies einem Notenvorteil von etwa einer halben Notenstufe im Leistungskurs bzw. zwei Dritteln einer Notenstufe im Grundkurs zugunsten der Hamburger Abiturienten.

Als wie bedeutsam sind diese Befunde einzustufen? Unter dem Blickwinkel der Verteilungsgerechtigkeit lässt sich diese Frage neben der von uns vorgenommenen Projektion der Studienzulassung vor allem mit Bezug auf den Einfluss der einzelnen Fachnote auf die nach wie vor besonders verteilungsrelevante Abiturgesamtnote beantworten. Nach den maßgebenden Regelungen aus der für die untersuchten Abiturientenjahrgänge geltenden KMK-Oberstufenvereinbarung floss die Fachnote im Leistungskurs insgesamt mit dem Faktor 8 in die Gesamtqualifikation ein,[3] bei der ein Maximalwert von 840 Punkten erreichbar war. Entsprechend würde ein Bewertungsunterschied von 1,5 Punkten in einem Leistungskursfach mit 12 Punkten in der Gesamtqualifikation zu Buche schlagen. Für ein Grundkursfach flossen in der Regel die Kursergebnisse aus vier Halbjahren mit ein, was in unserem Fall einen Unterschied von 8 Punkten (2 Punkte mal 4) ausmachen würde. Das sind die zu erwartenden Auswirkungen von Bewertungsunterschieden in einem einzelnen Fach, die jedoch nicht ohne weiteres auf andere Fächer übertragen werden können. Dies zeigen unsere Analysen für das Fach Englisch ganz deutlich. Aussagen zu Unterschieden in der Bewertungsstrenge lassen sich nur fachspezifisch treffen.

Ein bedeutender Teil der vorgefundenen Bewertungsunterschiede in Mathematik ließ sich auf Referenzgruppeneffekte bei der Leistungsbewertung zurückführen. In Übereinstimmung mit bereits vorliegenden Befunden (vgl. Ingenkamp 1995; Trautwein u. Baeriswyl 2007) wurden individuelle Schülerleistungen in leistungsstarken Lerngruppen schlechter bewertet als in leistungsschwächeren Lerngruppen. Damit zeigt sich hier sehr deutlich der Spagat, den Schulnoten in Hinblick auf ihre vielfältigen Funktionen bewältigen müssen. Während hinsichtlich der Selektions- bzw. Allokationsfunktion ganz klar objektive, referenzgruppenübergreifende Kriterien im Vordergrund stehen, hat sich die Benotung „um des Gelingens der pädagogischen Arbeit willen auch an lokalen, in der Regel lerngruppenspezifischen Referenzmaßstäben zu orientieren“ (Baumert et al. 2003, S. 322). Gleichwohl steht unter dem Gesichtspunkt der Verteilungsgerechtigkeit außer Frage, dass die Spielräume für die regionale und die institutionelle Flexibilität von Bewertungsmaßstäben nur begrenzt sein können. Hier scheint ein nahezu unauflöslicher Widerspruch zu bestehen, es sei denn, Leistungsunterschiede zwischen formal äquivalenten Lerngruppen tendieren gegen Null, was wiederum eine realitätsferne Annahme ist. Dennoch kann in Bemühungen um eine Reduzierung starker Leistungsunterschiede

zwischen Lerngruppen ein Ansatzpunkt gesehen werden, um Bewertungsunterschiede zwischen Schülerinnen und Schülern aus unterschiedlichen Lerngruppen zu reduzieren.

Mit Blick auf die beschränkte Vergleichbarkeit von Fachnoten sind wir der Frage nachgegangen, inwieweit Prüfungsnoten aus zentral vorgegebenen Abschlussprüfungen ein stärkeres Maß an Vergleichbarkeit bieten. Unsere Analysen für das Fach Mathematik liefern deutliche Hinweise darauf, dass die jeweils auf Bundeslandebene vorgegebenen zentralen Abiturprüfungen zu einer Reduzierung von Bewertungsunterschieden zwischen den Bundesländern beitragen können. Hamburger Abiturienten erhielten in Mathematik im Mittel niedrigere Prüfungsnoten. Die Prüfungsnoten korrespondierten relativ gut mit den gezeigten Mathematikleistungen. Der Punktevorteil für die Hamburger Gymnasiasten betrug im Leistungskurs bei gleicher Leistung lediglich 0,5 Punkte auf der 15er-Punktemetrik. Unsere Befunde für Mathematik stützen damit die Sicht, dass Noten aus zentralen Abschlussprüfungen unter dem Aspekt der Verteilungsgerechtigkeit ein objektiveres Auswahlkriterium darstellen als referenzgruppenbezogene Fachnoten. Allerdings ist an dieser Stelle nochmals hervorzuheben, dass sich unsere Analysen nur auf zwei Bundesländer und die Leistungskurse in zwei Fächern bezogen. Für eine Generalisierbarkeit unserer Befunde bedarf es weiterer Untersuchungen in anderen Bundesländern und weiteren Fächern. Zu beachten ist außerdem, dass unsere Schlussfolgerungen wesentlich davon abhängen, dass die von uns verwendeten „objektiven" Leistungstests tatsächlich in beiden Bundesländern in ähnlicher und umfassender Weise bewertungsrelevante Fähigkeiten erfassen – für diese Annahme fanden wir, wie oben berichtet, einige Hinweise, endgültig verifizieren lässt sie sich jedoch nicht.

Aber auch unabhängig davon dürfen unsere Befunde nicht uneingeschränkt als Aufforderung zur Stärkung bzw. Ausweitung zentraler Prüfungskomponenten verstanden werden; denn die stärkere Gewichtung der auf den ersten Blick leistungsgerechteren Prüfungsnoten kann eine andere Form der Benachteiligung nach sich ziehen: Und zwar in der Art, dass Schülerinnen und Schüler mit gleichen Lernvoraussetzungen beim Zugang zu attraktiven Bildungswegen durch ungünstige schulische Lernbedingungen stärker benachteiligt werden als zuvor, da das geringere Leistungsniveau nicht mehr im gleichen Maß durch die referenzgruppenbezogene Notenvergabe kompensiert wird.

Insgesamt gesehen scheint eine einfache bzw. vollständige Lösung des Spannungsverhältnisses von überschulischer und länderübergreifender objektiver Leistungsbewertung auf der einen und dem pädagogisch sinnvollen Umgang mit der Notenvergabe auf der anderen Seite nicht in Sicht. Das Abiturzeugnis stellt in dieser Hinsicht insofern einen Balanceversuch dar, als dass sowohl Fach- als auch Prüfungsnoten in die Gesamtqualifikation einfließen. Die sich aus den zahlreichen Kursergebnissen und Prüfungskomponenten zusammensetzende Abiturgesamtnote hat sich in vielen Untersuchungen (vgl. Köller u. Baumert 2002; Nagy 2006; Trost et al. 1998) als bester Prädiktor für den Studienerfolg erwiesen. Verglichen mit einzelnen Fach- bzw. Prüfungsnoten weist sie ein wesentlich höheres Aggregationsniveau auf. In diesem Zusammenhang möchten wir abschließend herausstellen, dass sich aufgrund unserer Befunde keine Aussagen zur Objektivität und prognostischen Validität der Abiturgesamtnote treffen lassen. Hierzu bedarf es umfangreicherer Untersuchungen in den von uns nicht berücksichtigten Leistungsdomänen sowie dem Einbezug von Studien- und Berufserfolgsmaßen, wie dies für Baden-Württemberg im Rahmen der längsschnittlichen Weiterführung der TOSCA-Studie erfolgt. Mit den

zukünftig auf den nationalen Bildungsstandards basierenden Ländervergleichen und dem sich gegenwärtig im Vorbereitungsstadium befindenden Nationalen Bildungspanel deuten sich darüber hinaus eine Reihe aussichtsreicher Möglichkeiten an, die es zu nutzen gilt, um die Befundlage zur Vergleichbarkeit und der Aussagekraft von Abiturnoten auf eine breitere empirische Basis zu stellen.

Anmerkungen

1 Die Punkte auf der 15er-Punktemetrik lassen sich wie folgt in die 6er-Notenmetrik überführen: 0 Punkte = Note 6, 1–3 Punkte = Note 5, 4–6 Punkte = Note 4, 7–9 Punkte = Note 3, 10–12 Punkte = Note 2, 13–15 Punkte = Note 1.

2 Durch die vorgenommene Standardisierung bilden die Korrelationen den Zusammenhang zwischen Testleistungen und Fachnoten innerhalb der einzelnen Schulen ab, wobei für Grund- und Leistungskursschüler jeweils getrennte Korrelationen ermittelt wurden.

3 Dabei flossen ein: die doppelt gewichteten Kursergebnisse aus den ersten drei Halbjahren der Qualifikationsphase sowie in einfacher Wertung das Kursergebnis aus dem vierten Halbjahr und – sofern keine Facharbeit in die Gesamtqualifikation eingebracht wurde – ein weiteres Kursergebnis.

Literatur

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.

Baumert, J., Bos, W., & Lehmann, R. (2000). *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.* Opladen: Leske + Budrich.

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 et al. (Hrsg.), *PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 261–331). Opladen: Leske + Budrich.

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Valtin, R., & Walther, G. (Hrsg.). (2004a). *IGLU. Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich.* Münster: Waxmann.

Bos, W., Lankes, E.-M., Schwippert, K., Thiel, O., Valtin, R., & Voss, A. (2004b). Schullaufbahnempfehlungen von Lehrkräften für Kinder am Ende der vierten Jahrgangsstufe. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, R. Valtin, & G. Walther (Hrsg.), *IGLU. Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich* (S. 191–228). Münster: Waxmann.

Büchel, F., Jürges, H., & Schneider, K. (2003). Die Auswirkungen zentraler Abschlussprüfungen auf die Schulleistung. Quasi-experimentelle Befunde aus der deutschen TIMSS-Stichprobe. *Vierteljahrshefte zur Wirtschaftsforschung, 72,* 238–251.

Ingenkamp, K. (1969). *Zur Problematik der Jahrgangsklasse*. Weinheim: Beltz.

Ingenkamp, K. (Hrsg.). (1995). *Die Fragwürdigkeit der Zensurengebung. Texte und Untersuchungsberichte* (9. Aufl.). Weinheim: Beltz.

Jonkmann, K., Köller, O., & Trautwein, U. (2007). Englischleistungen am Ende der Sekundarstufe II. In U. Trautwein, O. Köller, R. Lehmann, & O. Lüdtke (Hrsg.), *Schulleistungen von Abiturienten. Regionale, schulformbezogene und soziale Disparitäten* (S. 113–142). Münster: Waxmann.

Jürgens, E. (2005). *Leistungen und Beurteilungen in der Schule. Eine Einführung in Leistungs- und Bewertungsfragen aus pädagogischer Sicht* (5. Aufl.). St. Augustin: Academia.

Klein, E.-D., Kühn, S.-M., Ackeren, I.van, & Block, R. (2009). Wie zentral sind zentrale Püfungen? Abschlussprüfungen am Ende der Sekundarstufe II im nationalen und internationale Vergleich. *Zeitschrift für Pädagogik, 55,* 596–621.

Klieme, E. (2000). Fachleistungen im voruniversitären Mathematik- und Physikunterricht. Theoretische Grundlagen, Kompetenzstufen und Unterrichtsschwerpunkte. In J. Baumert, W. Bos, & R. Lehmann (Hrsg.), *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. 57–128). Opladen: Leske + Budrich.

Köller, O., & Baumert, J. (2002). Das Abitur – immer noch ein gültiger Indikator für Studierfähigkeit? *Aus Politik und Zeitgeschichte, B26*, 12–19.

Köller, O., & Trautwein, U. (2004). Englischleistungen von Schülerinnen und Schülern an allgemein bildenden und beruflichen Gymnasien. In O. Köller, R. Watermann, U. Trautwein & O. Lüdtke (Hrsg.), *Wege zur Hochschulreife in Baden-Württemberg. TOSCA – Eine Untersuchung an allgemein bildenden und beruflichen Gymnasien* (S. 285–326). Opladen: Leske + Budrich.

Köller, O., Baumert, J., & Schnabel, K. U. (1999). Wege zur Hochschulreife. Offenheit des Systems und Sicherung vergleichbarer Standards. Analysen am Beispiel der Mathematikleistungen von Oberstufenschülern an integrierten Gesamtschulen und Gymnasien in Nordrhein-Westfalen. *Zeitschrift für Erziehungswissenschaft, 2,* 385–422.

Köller, O., Watermann, R., Trautwein, U., & Lüdtke, O. (Hrsg.). (2004). *Wege zur Hochschulreife in Baden-Württemberg. TOSCA – Eine Untersuchung an allgemein bildenden und beruflichen Gymnasien.* Opladen: Leske + Budrich.

Lehmann, R. H., Peek, R., Gänsfuß, R., & Husfeldt, V. (2002). *Aspekte der Lernausgangslage und der Lernentwicklung – Klassenstufe 9. Ergebnisse einer Längsschnittuntersuchung in Hamburg.* Hamburg: Behörde für Bildung und Sport, Amt für Schule.

Lehmann, R. H., Vieluf, U., Nikolova, R., & Ivanov, S. (2006). *LAU 13. Aspekte der Lernausgangslage und Lernentwicklung – Klassenstufe 13.* Hamburg: Behörde für Bildung und Sport, Amt für Bildung.

Lüdtke, O., Becker, M., Neumann, M., Nagy, G., Jonkmann, K., & Trautwein, U. (2007a). Durchführung und methodische Grundlagen. In U. Trautwein, O. Köller, R. Lehmann, & O. Lüdtke (Hrsg.), *Schulleistungen von Abiturienten. Regionale, schulformbezogene und soziale Disparitäten* (S. 31–42). Münster: Waxmann.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Köller, O. (2007b). Umgang mit fehlenden Werten in der psychologischen Forschung. Probleme und Lösungen. *Psychologische Rundschau, 58(2)*, 103–117.

Maag Merki, K. (2008). Introduction of central A-level-exams. First results on the effects of the implementation. Vortrag auf der Konferenz der Special Interest Group Educational Effectiveness (SIG 18) der European Association of Research on Learning and Instruction (EARLI) in Frankfurt a. M. (28.-30. August 2008).

Maaz, K., Neumann, M., Trautwein, U., Wendt, W., Lehmann, R., & Baumert, J. (2008). Der Übergang von der Grundschule in die weiterführende Schule. Die Rolle von Schüler- und Klassenmerkmalen beim Einschätzen der individuellen Lernkompetenz durch die Lehrkräfte. *Schweizerische Zeitschrift für Bildungswissenschaften, 30*, 519–548.

Muthén, B. O., & Muthén, L. K. (1998–2008). *Mplus User's Guide*. Los Angeles: Muthén & Muthén.

Nagy, G. (2006). Berufliche Interessen, kognitive und fachgebundene Kompetenzen. Ihre Bedeutung für die Studienfachwahl und die Bewährung im Studium. Dissertation, Freie Universität Berlin. URL: http://www.diss.fu-berlin.de/2007/109/ (Zugegriffen 30. Nov. 2008).

Nagy, G., Neumann, M., Becker, M., Watermann, R., Köller, O., Lüdtke, O. et al. (2007). Mathematikleistungen am Ende der Sekundarstufe II. In U. Trautwein, O. Köller, R. Lehmann, & O. Lüdtke (Hrsg.), *Schulleistungen von Abiturienten. Regionale, schulformbezogene und soziale Disparitäten* (S. 71–112). Münster: Waxmann.

Neubrand, M., & Klieme, E. (2002). Mathematische Grundbildung. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele et al. (Hrsg.), *PISA 2000. Die Länder der Bundesrepublik Deutschland im Vergleich* (S. 95–127). Opladen: Leske +Budrich.

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E. et al. (Hrsg.). (2005). *PISA 2003. Der zweite Vergleich der Länder in Deutschland – Was wissen und können Jugendliche*. Münster: Waxmann.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Rheinberg, F. (2001). Bezugsnormen und schulische Leistungsbeurteilung. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 59–71). Weinheim: Beltz.

Sacher, W. (2005). *Leistungen entwickeln, überprüfen und beurteilen. Grundlagen, Hilfen und Denkanstöße für alle Schularten*. Bad Heilbrunn: Klinkhardt.

Schrader, F.-W. (2001). Diagnostische Kompetenz von Eltern und Lehrern. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (2. Aufl., S. 91–96). Weinheim: Beltz.

Schrader, F.-W., & Helmke, A. (2001). Alltägliche Leistungsbeurteilung durch Lehrer. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 45–58). Weinheim: Beltz.

Tent, L. (2001). Zensuren. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (2. Aufl., S. 805–811). Weinheim: Beltz.

Trautwein, U., & Baeriswyl, F. (2007). Wenn leistungsstarke Klassenkameraden ein Nachteil sind. Referenzgruppeneffekte bei Übergangsentscheidungen. *Zeitschrift für Pädagogische Psychologie, 21,* 119–133.

Trautwein, U., Köller, O., Lehmann, R., & Lüdtke, O. (2007). *Schulleistungen von Abiturienten. Regionale, schulformbezogene und soziale Disparitäten*. Münster: Waxmann.

Trost, G., Klieme, E., & Nauels, H.-U. (1998). The relationship between different criteria for admission to medical school and student success. *Assessment in Education, 5*, 247–254.

Watermann, R., Nagy, G., & Köller, O. (2004). Mathematikleistungen in allgemein bildenden und beruflichen Gymnasien. In O. Köller, R. Watermann, U. Trautwein, & O. Lüdtke (Hrsg.), *Wege zur Hochschulreife in Baden-Württemberg. TOSCA – eine Untersuchung an allgemein bildenden und beruflichen Gymnasien* (S. 205–283). Opladen: Leske + Budrich.

Wild, P., & Krapp, A. (2001). Pädagogisch-psychologische Diagnostik. In A. Krapp & B. Weidemann (Hrsg.), *Pädagogische Psychologie* (S. 513–563). Weinheim: Beltz.

Wissenschaftsrat (2004). Empfehlungen zur Reform des Hochschulzugangs. URL: http://www.wissenschaftsrat.de/texte/5920-04.pdf (Zugegriffen 30. Nov. 2008).

Wößmann, L. (2005). The effect heterogeneity of central exams. Evidence from TIMSS, TIMSS-Repeat and PISA. *Education Economics, 13*(2), 143–169.