Lars Hoffmann, Pauline Schröter, Alexander Groß,
Svenja Mareike Schmid-Kühn & Petra Stanat (Hg.)

Das unvergleichliche Abitur

Entwicklungen – Herausforderungen – Empirische Analysen

IQB
Institut zur Qualitätsentwicklung im Bildungswesen

universität koblenz
weiter:denken

wbv

Das unvergleichliche Abitur

Entwicklungen – Herausforderungen – Empirische Analysen

Lars Hoffmann, Pauline Schröter, Alexander Groß,
Svenja Mareike Schmid-Kühn & Petra Stanat (Hg.)

Lars Hoffmann, Pauline Schröter, Alexander Groß,
Svenja Mareike Schmid-Kühn & Petra Stanat (Hg.)

Das unvergleichliche Abitur

Entwicklungen – Herausforderungen – Empirische Analysen

wbv

2022 wbv Publikation
ein Geschäftsbereich der
wbv Media GmbH & Co. KG, Bielefeld

Gesamtherstellung:
wbv Media GmbH & Co. KG, Bielefeld
wbv.de

Umschlagfoto:
© picture-alliance/dpa/Felix Kästle

Bestell-Nr. I70568
ISBN (Print): 9783763970568
ISBN (E-Book): 9783763972494
DOI: 10.3278/9783763972494

Printed in Germany

Diese Publikation ist frei verfügbar zum
Download unter **wbv-open-access.de**

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Inhalt

Vorwort

Lars Hoffmann, Pauline Schröter, Alexander Groß,
Svenja Mareike Schmid-Kühn & Petra Stanat

Die Allgemeine Hochschulreife ist der höchste und gleichzeitig auch meistdiskutierte Schulabschluss in Deutschland. Im letzten regulären Prüfungsjahr vor Ausbruch der Corona-Pandemie 2019 schlossen 277.308 Schüler:innen ihre Schullaufbahn in Deutschland mit dem Abitur ab (Statistisches Bundesamt, 2020, Tab. 6.2), was einem Anteil der gleichaltrigen Schulabsolvent:innen von 40,2 Prozent entspricht. Im Vergleich zum Jahr 2001 ist dieser Anteil um etwa 15 Prozentpunkte angestiegen (ebd., Tab. 6.7 (i)) – das Abitur hat also auch in jüngerer Zeit noch einmal deutlich an Bedeutung gewonnen. Dieser Anstieg ist sowohl auf sich verändernde elterliche Bildungsaspirationen als auch auf gestiegene Erwartungen und Anforderungen der Berufs- und Arbeitswelt zurückzuführen (z. B. Neumann & Trautwein, 2019). Welche Optionen Schüler:innen für ihre weiteren Bildungs- und Berufswege haben, wird maßgeblich durch das Erreichen bzw. Nicht-Erreichen der Allgemeinen Hochschulreife beeinflusst (z. B. Kahnert, Eickelmann, Lorenz & Bos, 2015; Maag Merki, 2012), wobei zunehmend auch alternative Wege zur Hochschulberechtigung zur Verfügung stehen und genutzt werden (Autorengruppe Bildungsberichterstattung, 2020).

Dennoch ist die Bedeutung des Abiturs für den Zugang zur Hochschulbildung, die von einem hohen Anteil von Eltern für ihre Kinder angestrebt wird, weiterhin groß und die Abiturnote vor allem auch bei zulassungsbeschränkten Fächern in hohem Maße relevant. Entsprechend sind Fragen der Qualität und Vergleichbarkeit des Abiturs Gegenstand wiederkehrender Diskussionen in der (Bildungs-)Politik und Administration, in der Schulpraxis, der Wissenschaft und nicht zuletzt auch der breiten Öffentlichkeit. Kritische Einschätzungen hängen unter anderem damit zusammen, dass aufgrund der föderalen Struktur in Deutschland, die als Kulturhoheit der Länder (Art. 30 GG) vor allem auch in der Bildungspolitik zum Tragen kommt, strukturelle und inhaltliche Entscheidungen, die das Abitur bzw. den Weg zur Allgemeinen Hochschulreife betreffen, den einzelnen Bundesländern obliegen. Neben der allgemeinen Qualität des Abiturs wird daher auch diskutiert, inwieweit dieses Abschlusszertifikat über die Länder hinweg vergleichbar bzw. gleichwertig ist, was trotz verschiedener Standardisierungsbemühungen weiterhin fraglich erscheint.

Eine Sichtung der medialen Berichterstattung zum Abitur macht deutlich, dass Fragen der Qualität und Vergleichbarkeit der Allgemeinen Hochschulreife bereits seit Jahrzehnten kontinuierlich und mit deutlichem Fokus auf wahrgenommene Defizite immer wieder und in ähnlicher Form öffentlich diskutiert werden (Müller et al., 2022). Hierbei geht es insbesondere um das Anspruchsniveau und den damit verbundenen Wert des Abiturs allgemein (*„Was ist das Abitur wert?“* [ZEIT, 27.04.1990]; *„Was ist das Abitur noch wert?“* [FAZ, 12.12.2016]), um die vermeintlich „inflationäre“ Verteilung

von Bestnoten in den Bundesländern (*„Einser-Inflation und Notenungerechtigkeit“* [FAZ, 10.06.2015]) und um angenommene Unterschiede in den Anforderungen zwischen den Ländern (*„Unmut über ungleiches Abitur“* [SZ, 19.12.1970]; *„Unvergleichbares Abitur“* [FAZ, 19.10.2020]), die in der Konsequenz zu einer ungerechten Notenvergabe führten *(„Mehr Gerechtigkeit für Abiturienten“* [SZ, 17.03.1975]; *„Wie Abiturprüflinge ungleich behandelt werden“* [FAZ, 19.05.2015]). Es wird kritisiert, dass die Bundesländer zu großen Teilen eigene Qualitätslogiken und Bewertungsstandards hätten *(„Abi-Noten: Föderalismus sorgt für Chaos und Ungerechtigkeit“* [DER SPIEGEL, 07.06.2015]) und diese bundeslandspezifisch unterschiedlichen Herangehensweisen zu einem Niveauverlust und einer mangelnden Vergleichbarkeit des von allen Ländern vergebenen Abschlusses der Allgemeinen Hochschulreife führten.

Angesichts der wiederkehrenden, häufig stark normativ geprägten Diskussion über Qualität und Vergleichbarkeit des Abiturs ist es überraschend, dass es hierzu – verglichen mit anderen bildungspolitisch relevanten Themen – wenig empirische Forschung gibt: Vor dem Hintergrund der Debatte über die Aussagekraft des Abiturzeugnisses und der vermeintlichen Noteninflation liegen ländervergleichende Analysen von bildungsstatistischen Daten vor, die von der Kultusministerkonferenz bereitgestellt werden (u. a. Abiturdurchschnittsnoten, prozentualer Anteil nicht bestandener Abiturprüfungen und Anteil der Absolventinnen und Absolventen mit der Abschlussnote 1,0; vgl. KMK, 2022) (z. B. Neumann & Trautwein, 2019). Zwar wird in diesem Kontext auch die mangelnde Vergleichbarkeit von Leistungs- und Bewertungsstandards thematisiert, Studien zu voruniversitären Leistungen von Schüler:innen am Ende der gymnasialen Oberstufe wurden bislang jedoch kaum durchgeführt (für einen Überblick vgl. Leucht et al., 2016). Die wenigen Analysen, die hierzu vorliegen, haben sich entweder auf einzelne Bundesländer konzentriert (z. T. verbunden mit Analysen zu Leistungsunterschieden zwischen verschiedenen Kursniveaus und unterschiedlichen zum Abitur führenden Schularten) oder haben lediglich einzelne Länder miteinander verglichen. Veröffentlichungen, die sich dezidiert mit der Abiturprüfung oder der gymnasialen Oberstufe beschäftigen, sind größtenteils im Zuge der Einführung eines Zentralabiturs in mehreren Bundesländern zu Beginn der 2000er-Jahre entstanden. Sie untersuchen unter anderem organisationale Verarbeitungsstrategien im Umgang mit zentralen Abiturprüfungen (z. B. Intensität und Ausrichtung der Lehrkräftekooperation), die Unterrichtsgestaltung im Kontext zentraler Prüfungen (z. B. Intensität der Prüfungsvorbereitung, Vereinbarkeit von Standardisierung und Maßnahmen zur individuellen Förderung, Bezugsnormorientierung) sowie die Wahrnehmung und die Akzeptanz des Zentralabiturs aus Sicht schulischer Akteure (für einen Überblick vgl. Maag Merki, 2012; Klein et al., 2014). Hinsichtlich der mit zentralen Prüfungen verbundenen Intention, Qualität und Vergleichbarkeit zu sichern (vgl. Klein et al., 2009), sind Abiturprüfungsaufgaben sowie Korrektur- und Bewertungsvorgaben konstitutive Merkmale und folglich ebenfalls Gegenstand empirischer Studien, wobei sich diese Forschung bislang überwiegend auf die Analyse von Aufgabenmerkmalen konzentriert (für einen Überblick vgl. Kühn, 2016). Zur Ausgestaltung von Korrektur- und Bewertungsvorgaben im Abitur wurde bislang hingegen

kaum geforscht und die vorliegenden Befunde beziehen sich ausschließlich auf das Fach Deutsch (Diesdorn-Liesen, 2016; Köster, 2006; Zabka & Stark, 2010).

In den vergangenen Jahren wurden verschiedene Maßnahmen auf den Weg gebracht, die länderübergreifend eine höhere Qualität und Vergleichbarkeit der Leistungsanforderungen für das Erreichen der Allgemeinen Hochschulreife sowie der Ausgestaltung der gymnasialen Oberstufe und der Abiturprüfungen gewährleisten sollen. Dazu gehören insbesondere

- die Weiterentwicklung der *Einheitlichen Prüfungsanforderungen in der Abiturprüfung* (EPA) zu *Bildungsstandards für die Allgemeine Hochschulreife* als verbindliche Zielvorgaben für den Unterricht in der Sekundarstufe II und als Grundlage für die Abiturprüfungen in allen Ländern in den Fächern Deutsch, Mathematik, Fortgeführte Fremdsprache (Englisch, Französisch) sowie Biologie, Chemie und Physik (KMK, 2012a, 2012b, 2012c, 2020a, 2020b, 2020c),
- der Aufbau *Gemeinsamer Abituraufgabenpools der Länder* in den Fächern Deutsch, Mathematik, Englisch und Französisch (verfügbar seit dem Prüfungsjahr 2017) sowie Biologie, Chemie und Physik (verfügbar ab dem Prüfungsjahr 2025), denen alle Länder Aufgaben zum Einsatz in den landeseigenen Abiturprüfungen entnehmen können und für die normierende Wirkungen auf landeseigene Abiturprüfungsaufgaben und auf den Unterricht in der Sekundarstufe II erwartet werden, sowie
- die Weiterentwicklung der *Vereinbarung zur Gestaltung der gymnasialen Oberstufe und der Abiturprüfung* als Rahmenvorgabe für alle Bundesländer (KMK, 2021).

Diese Entwicklungen haben durch das Urteil des Bundesverfassungsgerichts vom 19. Dezember 2017 zur Studienplatzvergabe für das Fach Humanmedizin (BVerfG, 2017) und durch die Entscheidung der KMK, dass die Länder ab dem Jahr 2023 (Deutsch, Englisch, Französisch und Mathematik) bzw. ab dem Jahr 2025 (Biologie, Chemie und Physik) jeweils mindestens 50 % der Abiturprüfungsaufgaben aus den Gemeinsamen Abituraufgabenpools entnehmen sollen, zusätzlichen Schub erhalten.

Die skizzierten Debatten um die Qualität und Vergleichbarkeit der Allgemeinen Hochschulreife in Deutschland und die darauf bezogenen Maßnahmen, die in jüngerer Zeit auf den Weg gebracht wurden, bilden den Ausgangspunkt für den vorliegenden Herausgeberband. Dieser zielt darauf ab, aus wissenschaftlich-analytischer Perspektive über historische und aktuelle Entwicklungen des Abiturs zu informieren, den aktuellen Forschungsstand zusammenzufassen, neuere Analysen darzustellen und Forschungslücken zu identifizieren, die zukünftig bearbeitet werden sollten. Zum einen wird aufgezeigt, welche Schritte in der Vergangenheit unternommen wurden, um die Vergleichbarkeit des Abiturs innerhalb Deutschlands zu erhöhen. Zum anderen wird dargestellt, in welchen Bereichen und mit welchen Instrumenten die Angleichung der Abituranforderungen der Bundesländer gegenwärtig weiterentwickelt wird und welche Herausforderungen dabei bestehen bzw. zukünftig daraus erwachsen werden. Der Schwerpunkt der Darstellung liegt dabei auf den schriftlichen Abiturprüfungen, die am Ende der gymnasialen Oberstufe durchgeführt werden. Um

die Prüfungssysteme der Bundesländer international einzuordnen, werden zudem die Prüfungssysteme anderer Staaten skizziert, wobei der Fokus auf unterschiedlichen Ansätzen des Umgangs mit Vergleichbarkeit von Anforderungen und Bewertungen liegt. Auch die empirischen Analysen zum Abitur, über die im Band berichtet wird, beziehen sich auf Fragen der Vergleichbarkeit. Damit soll ein Beitrag zur empirischen Fundierung, Differenzierung und Versachlichung der fortdauernden öffentlichen Diskussionen über das Abitur geleistet werden, die oft durch starke subjektive Überzeugungen und vereinfachende Vorstellungen von Vergleichbarkeit geprägt sind.

Der Herausgeberband ist in zwei Teile gegliedert. Im Fokus von *Teil I, Abitur und Abiturprüfungen in Deutschland*, stehen die im Untertitel des Bandes benannten Themenschwerpunkte „Entwicklungen" und „Herausforderungen" beim Abitur. Beginnend mit dem preußischen Abiturreglement des Jahres 1788 zeichnet Klaus Klemm im *ersten Beitrag* die mehr als zweihundertjährige Geschichte der Allgemeinen Hochschulreife nach. Diese wird in ihrer Berechtigungsfunktion für den Universitätszugang historisch verortet und es wird aufgezeigt, wie sich die Zugänge und Barrieren zum Studium sowie die curriculare Orientierung des Abiturs im Laufe der Geschichte gewandelt haben. Darüber hinaus verdeutlicht der Beitrag, dass einige der historischen Entwicklungslinien eine bemerkenswerte Kontinuität aufweisen. So finden sich im Laufe der Geschichte wiederholt Vereinbarungen zur gegenseitigen Anerkennung von Abschlüssen oder zur Harmonisierung von Rahmenbedingungen des Abiturs.

Die jüngeren Entwicklungen des Abiturs und der Abiturprüfung in Deutschland werden von Lars Hoffmann, Pauline Schröter und Petra Stanat im *zweiten Beitrag* des Herausgeberbandes dargestellt, der die wesentlichen Annäherungs- und Standardisierungsprozesse der Länder der letzten 30 Jahre skizziert. Der Schwerpunkt des Beitrags liegt auf den Gemeinsamen Abituraufgabenpools der Länder, aus denen die Länder seit dem Prüfungsjahr 2017 Abituraufgaben entnehmen und in ihren Abiturprüfungen einsetzen können. Im Beitrag werden die Grundlagen der Abituraufgabenpools, das Vorgehen bei deren Entwicklung und die damit verbundenen Herausforderungen erörtert.

Im Zuge der wiederkehrenden Diskussion über die (mangelnde) Vergleichbarkeit des Abiturs stehen auch die Heterogenität der länderspezifischen Gestaltungsvarianten der Struktur der gymnasialen Oberstufe und der formal-organisatorischen Rahmenbedingungen für die Abiturprüfung im Fokus. Im Rahmen des *dritten Beitrags* stellen Svenja Mareike Schmid-Kühn und Alexander Groß die Ergebnisse einer Dokumentenanalyse vor, in der für die gymnasiale Oberstufe und die Abiturprüfung relevante, institutionell verantwortete Vorgaben (z. B. Abiturprüfungsordnungen) aus allen Bundesländern vergleichend analysiert wurden. Insgesamt zeigt sich, dass trotz sichtbarer Bemühungen um eine stärkere Vergleichbarkeit der Rahmenbedingungen in den letzten Jahren weiterhin eine erhebliche Heterogenität hinsichtlich der Strukturen der gymnasialen Oberstufen und der Abiturprüfungsverfahren in Deutschland vorliegt.

Ein Seitenblick auf die Prüfungssysteme anderer Staaten bietet der von Pauline Schröter, Lars Hoffmann und Svenja Mareike Schmid-Kühn verfasste *vierte Beitrag*. Mit Irland, Australien, den Niederlanden und der Schweiz wurden vier Länder ausgewählt, die sich im Hinblick auf die Struktur ihres Bildungssystems (Zentralismus vs. Föderalismus) und die Organisation ihrer Abschlussprüfungen am Ende der Sekundarstufe II (zentrale vs. dezentrale Prüfungsverfahren) unterscheiden. Im Beitrag werden die wesentlichen Elemente der Prüfungssysteme dieser Länder in konziser Form beschrieben und mit Auszügen aus Interviews unterlegt, in denen Expert:innen der vier Länder zu den Charakteristika ihres jeweiligen Prüfungssystems befragt wurden. Der Schwerpunkt dieser Interviews lag auf der Frage, welche Verfahren und Ansätze andere Länder gewählt haben, um die Vergleichbarkeit von Abschlussnoten zu gewährleisten und welchen Herausforderungen sie sich dabei gegenübersehen. Hier zeigt sich eine große Vielfalt des Umgangs mit Fragen der Vergleichbarkeit und der Herausforderungen, die damit verbunden sind.

In der öffentlichen Debatte zu Qualität und Vergleichbarkeit des Abiturs fällt auf, dass die vorgebrachten Argumente nur selten auf empirischer Evidenz basieren, was nicht zuletzt am oben angesprochenen Mangel empirischer Studien zu diesem Themenfeld liegen dürfte. Vor diesem Hintergrund trägt der *zweite Teil* des vorliegenden Bandes aktuelle *empirische Untersuchungen zum Abitur und zu Abiturprüfungsaufgaben* zusammen. Dabei kommen unterschiedliche Forschungsmethoden und verschiedene, im Untertitel des Bandes benannte „empirische Analysen" zur Anwendung, die zur Beantwortung der jeweiligen Fragestellungen geeignet sind.

Im *fünften Beitrag* stellen Lars Hoffmann, Pauline Schröter und Petra Stanat die konzeptionellen Grundlagen und zentralen Ergebnisse der Evaluation vor, mit der das Institut zur Qualitätsentwicklung im Bildungswesen (IQB) an der Humboldt-Universität zu Berlin den Aufbau der Gemeinsamen Abituraufgabenpools der Länder begleitet. Diese Evaluation ist derzeit noch formativ angelegt und dient primär dazu, den Teams, die mit der Aufgabenentwicklung betraut sind, Rückmeldungen über die Qualität der Poolaufgaben zu geben.

Die Frage nach dem Erfolg der mit den Gemeinsamen Abituraufgabenpools der Länder einhergehenden Standardisierungsbemühungen bildet den Ausgangspunkt des im *sechsten Beitrag* von Alexander Groß und Svenja Mareike Schmid-Kühn vorgestellten, qualitativ-explorativen Forschungsprojekts. Im Rahmen einer Interviewstudie wird der gesamte Implementationsprozess der Abituraufgabenpools untersucht, wobei alle von der Einführung betroffenen Akteure verschiedener Systemebenen in die Analyse miteinbezogen werden. Der Beitrag stellt die Grundkonzeption des Forschungsprojekts sowie erste Untersuchungsergebnisse auf Ebene der Kultusministerkonferenz vor. Hierbei deutet sich ein Spannungsfeld zwischen gemeinsamer Standardisierungsbestrebungen auf der einen und der Anwendung individueller Ländertraditionen auf der anderen Seite an.

Die beiden darauffolgenden Beiträge beschäftigen sich mit den Abiturprüfungen im Fach Deutsch. Im *siebten Beitrag* stellen Michael Kämper-van den Boogaart und Sabine Reh wichtige historische Entwicklungslinien des deutschen Abituraufsatzes

dar und zeichnen dabei die wesentlichen bildungspolitischen und fachdidaktischen Diskurse der Vergangenheit zu dieser speziellen Form der Prüfungsaufgabe nach. Der Beitrag entstand im Kontext eines Forschungsprojekts, in dessen Rahmen Abituraufsätze, die im Zeitraum zwischen 1882 und 1972 verfasst worden sind, einer bildungshistorischen und fachdidaktischen Analyse unterzogen werden. Der *achte Beitrag* bezieht sich auf die schriftlichen Abiturprüfungen im Fach Deutsch der Gegenwart und wurde von Pauline Schröter, Hannelore Söldner, Lars Hoffmann, Anja Riemenschneider, Jörg Jost und Dorothee Wieser verfasst. Es werden darin zentrale Befunde zweier aktueller Studien dargestellt, die sich vor allem mit der Frage befassen, wie objektiv und reliabel Lehrkräfte Abituraufsätze im Fach Deutsch bewerten. In diesen quantitativ angelegten Studien wird zudem untersucht, inwieweit sich mit verschiedenen Varianten von Bewertungsvorgaben die Objektivität und die Reliabilität der Bewertung schriftlicher Abiturprüfungen im Fach Deutsch erhöhen lässt.

Im *neunten Beitrag* untersuchen Lars Hoffmann, Nicolas Hübner, Marco Neumann und Pauline Schröter, in welchem Maße sich die in den Halbjahren der gymnasialen Oberstufe erzielten Kursnoten von den Abiturprüfungsnoten unterscheiden. Für ihre quantitativen Analysen ziehen die Autor:innen Daten aus drei Quellen heran (Daten aus der Evaluation des IQB zu den Gemeinsamen Abituraufgabenpools der Länder, aus dem Nationalen Bildungspanel für das Land Thüringen und aus der BERLIN-Studie) und können so der Frage nach Unterschieden zwischen Kurs- und Prüfungsnoten mit unterschiedlichen Schwerpunktsetzungen und Perspektiven nachgehen. Dies umfasst unter anderem länderübergreifende und landesspezifische Analysen, die Untersuchung fächerspezifischer Befundmuster, die Analyse der Effekte von Oberstufenreformen auf Kurs- und Prüfungsnoten und die Betrachtung von Differenzen auf höherer Ebene, nämlich zwischen den im „Kursblock" für die Bildung der Abiturgesamtnote eingebrachten Leistungen einerseits und den im „Prüfungsblock" erzielten Ergebnissen andererseits. Das Befundmuster der Analysen weist insgesamt darauf hin, dass die in den Abiturprüfungen erzielten Noten im Mittel etwas schlechter ausfallen als die Kursnoten der Qualifikationsphase.

Der von Aiso Heinze, Irene Neumann und Christoph Deeken verfasste *zehnte Beitrag* beschäftigt sich mit Fragen der Passung zwischen den in der gymnasialen Oberstufe vermittelten Kompetenzen im Fach Mathematik und den mathematischen Kompetenzanforderungen zum Beginn eines Hochschulstudiums im MINT-Bereich. Berichtet werden zwei aufeinander aufbauende Studien. In der ersten Studie wurde im Rahmen einer Delphi-Befragung mit Hochschullehrenden erfasst, welche mathematischen Lernvoraussetzungen sie bei Studienanfänger:innen im MINT-Bereich erwarten. In einer zweiten Studie wurden diese Erwartungen der Hochschulseite mit den normativen Vorgaben der Lehrpläne eines Bundeslandes für das Fach Mathematik in der Sekundarstufe II abgeglichen. Die dabei identifizierten Passungslücken können als Grundlage für die Weiterentwicklung der curricularen Vorgaben des Schulsystems einerseits und der Unterstützungsangebote von Hochschulen andererseits genutzt werden.

Der Herausgeberband schließt mit einer *kritischen Zwischenbilanz* von Eckhard Klieme zur Vergleichbarkeit von Abiturnoten. Diese differenzierte Auseinandersetzung mit dem übergeordneten Thema des Bandes schließt nahtlos an die vorhergehenden Beiträge an und bringt deutlich auf den Punkt, dass einfache Vorstellungen von Vergleichbarkeit zu kurz greifen und eine absolute Vergleichbarkeit von Abiturnoten gar nicht erreicht werden kann. Der Autor schätzt die aktuellen Entwicklungen, die auf länderübergreifende Annäherungen abzielen, als prinzipiell sinnvoll ein, mahnt jedoch an, dass diese verstärkt empirisch fundiert und begleitet werden sollten.

Als Herausgeber:innen des vorliegenden Bandes möchten wir uns an dieser Stelle bei einigen Personen bedanken, die zum Gelingen dieses Werks beigetragen haben. An erster Stelle sind dies die Autor:innen der einzelnen Beiträge, denen wir für ihre Mitwirkung und ihre sorgfältige Arbeit beim Verfassen und Überarbeiten der Beiträge herzlich danken. Ebenfalls großer Dank gebührt Tim Desmond, James Tognolini, Marieke van Onna und Franz Eberle, die uns im Rahmen des vierten Beitrags für Interviews zum Prüfungssystem ihres jeweiligen Landes zur Verfügung gestanden und die Zusammenstellung der länderspezifischen Informationen eingehend geprüft haben. Besonderer Dank gilt Molly Carter für die Transkription der Interviews sowie Christoph Reuter für seine konstruktiven Rückmeldungen und wertvollen Recherchen. Weiterhin danken wir Jana Lutz, Janine Euent und Viktor Schenker, die das Projekt als studentische Mitarbeiter:innen unterstützt haben.

Literatur

Autorengruppe Bildungsberichterstattung (Hrsg.) (2020). *Bildung in Deutschland 2020. Ein indikatorengestützter Bericht mit einer Analyse zu Bildung in einer digitalisierten Welt*. Bielefeld: wbv.

BVerfG – Bundesverfassungsgericht (2017). *Urteil des Ersten Senats vom 19. Dezember 2017* (1 BvL 3/14, Rn. 1–253). http://www.bverfg.de/e/ls20171219_1bvl000314.html

Diesdorn-Liesen, V. (2016). *Vergleichbarkeit in der Vielfalt. Leistungsanforderungen und Leistungsfeststellung im Zentralabitur Deutsch*. Wiesbaden: Springer VS.

Grundgesetz der Bundesrepublik Deutschland. GG, Art. 30. Zugriff am 30.05.2022. https://www.gesetze-im-internet.de/gg/art_30.html

Kahnert, J., Eickelmann, B., Lorenz, R. & Bos, W. (2015). Die Steuerungsfunktion von zentralen Abiturprüfungen. In H. J. Abs, T. Brüsemeister, M. Schemmann & J. Wissinger (Hrsg.), *Governance im Bildungssystem* (S. 89–115). Wiesbaden: Springer VS

Klein, E. D., Krüger, M., Kühn, S. M. & van Ackeren, I. (2014). Wirkungen zentraler Abschlussprüfungen im Mehrebensystem Schule. Eine Zwischenbilanz internationaler und nationaler Befunde und Forschungsdesiderata. *Zeitschrift für Erziehungswissenschaft, 16*(1), 7–34.

Klein, E. D.; Kühn, S. M.; Ackeren, I. van & Block, R. (2009). Wie zentral sind zentrale Prüfungen? Zentrale Abschlussprüfungen am Ende der Sekundarstufe II im nationalen und internationalen Vergleich. *Zeitschrift für Pädagogik, 55*(4), S. 596–621.

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (2012a). *Bildungsstandards im Fach Deutsch für die Allgemeine Hochschulreife* (Beschluss der Kultusministerkonferenz vom 18.10.2012). https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Deutsch-Abi.pdf

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (2012b). *Bildungsstandards im Fach Mathematik für die Allgemeine Hochschulreife* (Beschluss der Kultusministerkonferenz vom 18.10.2012). https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Mathe-Abi.pdf

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (2012c). *Bildungsstandards für die fortgeführte Fremdsprache (Englisch/Französisch) für die Allgemeine Hochschulreife* (Beschluss der Kultusministerkonferenz vom 18.10.2012). https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Fortgef-FS-Abi.pdf

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (2020a). *Bildungsstandards im Fach Biologie für die Allgemeine Hochschulreife* (Beschluss der Kultusministerkonferenz vom 18.06.2020). https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2020/2020_06_18-BildungsstandardsAHR_Biologie.pdf

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (2020b). *Bildungsstandards im Fach Chemie für die Allgemeine Hochschulreife* (Beschluss der Kultusministerkonferenz vom 18.06.2020). https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2020/2020_06_18-BildungsstandardsAHR_Chemie.pdf

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (2020c). *Bildungsstandards im Fach Physik für die Allgemeine Hochschulreife* (Beschluss der Kultusministerkonferenz vom 18.06.2020). https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2020/2020_06_18-BildungsstandardsAHR_Physik.pdf

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (2021). *Vereinbarung zur Gestaltung der gymnasialen Oberstufe und der Abiturprüfung* (Beschluss der Kultusministerkonferenz vom 07.07.1972 i. d. F. vom 18.02.2021). https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/1972/1972_07_07-VB-gymnasiale-Oberstufe-Abiturpruefung.pdf

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (2022). *Schnellmeldung Abiturnoten 2021 an Gymnasien, Integrierten Gesamtschulen, Fachgymnasien, Fachoberschulen und Berufsoberschulen -vorläufige Ergebnisse-* (Schuljahr 2020/2021). https://www.kmk.org/fileadmin/Dateien/pdf/Statistik/Dokumentationen/Schnellmeldung_Abiturnoten_2021.pdf

Köster, J. (2006). Das Deutschabitur in Zeiten von Bildungsstandards – Vergleichbarkeit der Prüfungsleistungen und ihrer Bewertung. *Didaktik Deutsch 21*, 78–90.

Kühn, S. M. (2016). Aufgaben in (zentralen) Abschlussprüfungen. Theoretische und empirische Perspektiven auf ein interdisziplinäres Forschungsfeld. In S. Keller & C. Reintjes (Hrsg.), *Aufgaben als Schlüssel zur Kompetenz* (S. 73–92). Münster: Waxmann.

Leucht, M., Kampa, N. & Köller, O. (2016). *Fachleistungen beim Abitur. Vergleich allgemeinbildender und berufsbildender Gymnasien in Schleswig-Holstein*. Münster: Waxmann.

Maag Merki, K. (Hrsg.). (2012). *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Müller, A., Groß, A. & Schmid-Kühn, S. M. (2022). „Hier Einser-Abi, dort durchgefallen." Eine Analyse der öffentlichen Debatte über Qualität und Vergleichbarkeit des Abiturs in den Medien. *SchulVerwaltung, Ausgabe Hessen & Rheinland-Pfalz, 26*(5), 136–138.

Neumann, M. & Trautwein, U. (2019). Sekundarbereich II und der Erwerb der Hochschulzugangsberechtigung. In O. Köller, M. Hasselhorn, F. W. Hesse, K. Maaz, J. Schrader, H. Solga, K. Spieß & K. Zimmer (Hrsg.), *Das Bildungswesen in Deutschland. Bestand und Potenziale* (S. 533–564). Bad Heilbrunn: Klinkhardt, UTB.

Statistisches Bundesamt (Hrsg.). (2020). Allgemeinbildende Schulen – Fachserie 11 Reihe 1 – Schuljahr 2019/2020. Zugriff am 30.05.2022. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Schulen/Publikationen/Downloads-Schulen/allgemeinbildende-schulen-2110100207005.html

Zabka, T. & Stark, T. (2010). Aufgabenstellungen und Erwartungshorizonte als Steuerungsinstrumente. *Der Deutschunterricht, 16*(1), 19–29.

Teil I:
Entwicklungen und Strukturen

1 Die Geschichte der Allgemeinen Hochschulreife in Deutschland – Kontinuitäten im Wandel

Klaus Klemm

Zusammenfassung

Die Darstellung der Geschichte der Allgemeinen Hochschulreife behandelt die Entwicklung von 1788, dem Jahr, in dem in Preußen ein erstes „Reglement für die Prüfung an gelehrten Schulen" erlassen wurde, bis in unsere Tage. Wesentliche Aspekte dieser Entwicklung werden herausgearbeitet. Einleitend wird die bereits 1788 festgelegte Entscheidung für die Platzierung der Abschlussprüfung der höheren Schulen an diesen Schulen und nicht als Eingangsprüfung an den Hochschulen sowie die damit erfolgte dauerhafte Etablierung eines Berechtigungssystems dargestellt. Dem folgt ein Abschnitt, in dem der Zugang zu den abiturführenden Schulen und den im Laufe der Zeit sich verändernden Zugangsbarrieren analysiert wird: Zugangsbarrieren, die die Hochschulen vor realer oder vermuteter Überfüllung schützen sollen und die viele Jahre lang Mädchen und junge Frauen vom Studium fernhalten und die bis heute Kinder aus sozial schwächeren Familien treffen. Diesen eher auf die „Mengensteuerung" beim Hochschulzugang ausgerichteten Analysen steht ein weiterer Abschnitt gegenüber, der die curriculare Orientierung der abiturführenden Bildungsgänge der Schulen und ihrer Abiturprüfungen zum Gegenstand hat: Verfolgt wird die Entwicklung vom altsprachlich orientierten humanistischen Gymnasium über die Gleichstellung naturwissenschaftlich bzw. neusprachlich ausgerichteter Gymnasien bis hin zur curricular breit gefächerten gymnasialen Oberstufe der abiturführenden Schulen. Ein abschließender Abschnitt verweist auf Kontinuitäten der zweihundertjährigen Entwicklung des Abiturs.

1 Einleitung

Es war 1765, dass ein 16-jähriger Teenager, vom Vater zum Jurastudium gedrängt, an der Universität zu Leipzig eintraf. Es war Vorschrift, dass man sich den Professoren vorstellte; darin bestand das Auswahl- und Aufnahmeverfahren. Selten wurde jemand abgewiesen, schließlich lebten die Professoren von den Hörergeldern. Lustlos und uninteressiert folgte der junge Mann den Vorlesungen, zunächst in Leipzig, anschließend in Straßburg, wo er 1771 den Lizenziatentitel erwarb.

Später, der junge Mann – sein Name: Johann Wolfgang Goethe – war inzwischen Autor des Bestsellers „Die Leiden des jungen Werthers" geworden, wandte er sich

einem mittelalterlichen Stoff zu, der ihn schon früh beschäftigt hatte: der Geschichte des Wissenschaftlers, der sich mit dem Teufel verbündet. Erste Szenen entstanden. Hier nun war Gelegenheit zu einer burlesken Einlage, in der Goethe satirisch seine Universitätserfahrung lebendig werden lässt. Es ist die Szene, in der sich ein junger Mann, neu in der Stadt und verwirrt von den vielen Eindrücken, beim berühmten Gelehrten namens Faust vorstellt. Jedoch empfängt ihn nicht der Gelehrte selbst, sondern dessen neuer Gefährte Mephisto, im Talar, also der Amtstracht des Gelehrten. Mephisto berät den jungen Mann mit Auskünften über die studentische Lebensweise; aus diesen wird deutlich, wie sehr Wissenschaft und städtisches Leben auf des jungen Mannes Geld erpicht sind. Er beeindruckt den jungen Mann mit Ausführungen zur Basiswissenschaft Philosophie und auch mit lebenspraktischen, ein wenig schlüpfrigen Ratschlägen zur Berufsausübung als Arzt. Der junge Mann bemerkt nicht, dass Mephisto sich über wissenschaftliche Methodik und auch über Eitelkeit und Vorteilssuche der Lehrenden lustig macht. Dankbar und voller Ehrfurcht verabschiedet er sich schließlich. Er ist jetzt zum Studium aufgenommen.

2 Das preußische Abiturreglement von 1788 als Startschuss

Über derartige universitäre „Aufnahmegespräche", wie sie zur Zeit des jungen Goethe noch gängig waren, schreibt Bölling, dass die

> „Zulassungspraxis zu jener Zeit chaotisch war. Jeder junge Mann, der sich durch eine Lateinschule oder Privatlehrer hinreichend vorbereitet fühlte und die Unterstützung seiner Eltern hatte, konnte zur Universität ziehen [...]. Beim Dekan der gewählten Fakultät mussten sie sich zur Prüfung pro immatriculatione vorstellen, die jedoch keine ernsthafte Hürde darstellte, weil der Dekan oft schon durch die Zahl der Prüflinge überfordert war. Zudem standen die Professoren selbst unter gesellschaftlichem Druck, weil jede Abweisung wegen der Abhängigkeit von Hörergeldern auch ihr schmales Einkommen verringerte [...]". (Bölling, 2010, S. 26)

Mit Blick auf gravierende Qualitätsmängel der universitären Studien, die seit der Mitte des 18. Jahrhunderts in Preußen verstärkt beklagt wurden (vgl. auch Wolter, 1989, S. 7 f.) und durch das damalige Aufnahmeverfahren der Universitäten begünstigt, zumindest aber nicht gehindert wurden, schlug 1787 Hoffmann, der Kanzler der Universität Halle, eine Veränderung des Aufnahmeverfahrens künftiger Studenten vor. In seinem Vorschlag heißt es:

> „Die Erfahrung lehrt, dass sich unter den jungen Leuten, welche die Universitäten beziehn, beständig eine nicht geringe Anzahl von solchen Subjecten befindet, die nicht allein in den beyden sogenannten gelehrten Sprachen, sondern auch in den übrigen Vorkenntnissen, die sie von den Schulen mitbringen sollten, so unwissend sind, dass ihre Unwissenheit bald Mitleiden, und bald Widerwillen erregen muß [...]" (Schwartz, 1910, S. 67)

Um dem abzuhelfen, solle an jeder Universität eine Prüfungskommission eingesetzt werden mit der Aufgabe, die jungen Leute vor der Aufnahme eines Studiums „öffent-

lich zu prüfen, und diejenigen, welche allzu unwissend in den auf der Universität nöthigen Vorkenntnissen befunden würden, zur Schule, oder zu ihren Eltern zurückzuweisen“ (Schwartz, 1910, S. 68).

Die Kritik an der vielfach nicht gegebenen Eignung zum Universitätsstudium wurde in dieser Zeit von Klagen über die zu hohe Zahl der Studierenden begleitet. Ein anonymer Autor schrieb 1788 in der „Berlinischen Monatsschrift“ in einem Beitrag „Ueber die zu große Anzahl der Studierenden“: Vergleiche man die Anzahl der Bewerber

> „mit der von den wirklichen Aemtern, und den selbst nach der größten Mortalität berechneten Vakanzen: so bleibt doch keine Hoffnung übrig, alle diese jungen Leute, oder auch nur den größten Teil derselben, auf eine Art versorgt zu sehen, die mit den vielen Aufopferungen, die ihre Vorbereitung kostet, in einigem Verhältniß stände“ (Zitiert nach Kamp, 1988, S. 45).

Auf dem Hintergrund derartiger Klagen über fehlende Studiervoraussetzungen sowie einer Überfüllung der akademischen Berufe befasste sich das 1787 eingerichtete „Oberschulkollegium“, dem der preußische König die Oberaufsicht über das gesamte preußische Schulwesen übertragen hatte, mit dem Vorschlag Hoffmanns. Es holte dazu bei drei Universitäten und bei einer kleinen Zahl von „Schulmännern“ Gutachten ein. Auf der Basis der Anfang 1788 eingegangenen Gutachten wurde am 23. Dezember 1788 ein „Reglement für die Prüfung an gelehrten Schulen“ erlassen – das in der Literatur als „Abiturreglement von 1788“ bezeichnet wird (der Text des Reglements findet sich bei Schwartz, 1910, S. 122–128). Im Text dieses Reglements wird der erfolgreiche Prüfungsabschluss mit einem „Zeugniß der Reife“ bescheinigt. Der Begriff „Abitur“ taucht im Reglement nicht auf; wohl aber werden dort die erfolgreichen Absolventen als „Abiturientes“ bezeichnet.

Die folgenden Aspekte des Reglements sind für die spätere Entwicklung des preußischen Abiturs von Bedeutung:

- Abweichend von Hoffmanns Vorschlag sollen die Eingangsprüfungen nicht von den aufnehmenden Hochschulen, sondern von den abgebenden Schulen durchgeführt werden. Eine Ausnahme davon wurde denjenigen Studieninteressierten eingeräumt, die sich durch Privatunterricht auf ein Universitätsstudium vorbereitet hatten. Für sie bestand die Möglichkeit einer Aufnahmeprüfung an den Universitäten (vgl. das entsprechende „Reglement für die Prüfung an Universitäten“ (ebd., S. 128–133). Dieser alternative Zugangsweg bestand noch bis 1834 (Paulsen, 1885, S. 568).
- Bemerkenswert an diesem Abiturreglement ist, dass darin der Übergang aus der Schule in die Universität einheitlich geregelt wurde, dass aber zugleich darauf verzichtet wurde, den „zur Universität vorbereitenden Schulen in einem einheitlichen Lehrplan ein gleiches Ziel“ (Schwartz, 1910, S. 65) vorzugeben. Vorgeschrieben waren lediglich schriftliche Prüfungsarbeiten sowie mündliche Examina über die erworbenen Kenntnisse in den alten Sprachen, in den neueren Sprachen (besonders in der Muttersprache) sowie in „wissenschaftlichen Kenntnissen, vornehmlich historischen“ (ebd., S. 126). Schwartz kommentiert diesen

weitgehenden Verzicht auf einen landesweit gültigen Lehrplan so: Einen solchen Lehrplan „in Kürze zu erreichen, war unmöglich. So begnügte man sich mit dem Examen." (ebd., S. 65)

- In einer weiteren Bestimmung blieb das Abiturreglement hinter den auf eine qualitativ bessere schulische Vorbereitung der künftigen Studenten gerichteten Wünschen zurück. Die an den Schulen abgelegte Prüfung bescheinigte dem Absolventen seine „bey der Prüfung befundene Reife oder Unreife zur Universität". Weiter heißt es dann im Reglement: „Es ist jedoch hiebey unsere Absicht nicht, die bürgerliche Freyheit in so fern zu beschränken, daß es nicht ferner jedem Vater und Vormund frey stehen sollte, auch einen unreifen und unwissenden Jüngling zur Universität zu schicken" (Schwartz, 1910, Bd. I, S. 123). Das Ziel der Steuerung einer „Überfüllung" konnte so ebenso wenig wie das der Qualitätssicherung erreicht werden. Erst ab 1834 war für alle Studienanfänger eine erfolgreich abgeschlossene Abiturprüfung Zugangsvoraussetzung zum Universitätsstudium.
- Einer Gruppe allerdings wurde diese „Freiheit" nicht gewährt: „Zugleich wollen wir hierdurch ausdrücklich verordnen, daß nur diejenigen Jünglinge ein öffentliches Stipendium oder anderweitiges Benificium auf der Universität erhalten, und genießen können, welche das Zeugniß der Reife erhalten" (Schwartz, 1910, Bd I, S. 127).
- Grundsätzlich blieben Mädchen vom Zugang zu Universitäten ausgeschlossen.

Der Blick auf die hier skizzierte Entstehung sowie die zentralen Elemente des preußischen Abiturreglements lässt insgesamt drei Regelungsbereiche mit jeweils unterschiedlich weit gehenden Bestimmungen erkennen: die Platzierung der Abiturprüfung (Abschnitt 3), die Begrenzung des Zugangs zu Universitäten (Abschnitt 4) sowie eine curriculare Orientierung (Abschnitt 5). Die folgende Darstellung verfährt nun so, dass diese drei Teilbereiche abschnittweise dargestellt werden. Abschließend werden im 6. Abschnitt dieses Beitrags die Kontinuitäten im Verlauf der Geschichte der Allgemeinen Hochschulreife aufgezeigt.

3 Zur Platzierung der Abiturprüfung am Ende der Schulzeit

1788 wurde mit dem Abiturreglement eine für Preußen und etwa gleichzeitig für andere Länder des späteren deutschen Reiches (vgl. dazu den Abschnitt 5.1 dieses Beitrages) folgenreiche Weichenstellung vollzogen: Sieht man von einer bis 1834 währenden Übergangszeit ab, wurde die Hochschulzugangsberechtigung auf der Grundlage einer Abiturprüfung ausschließlich an den „gelehrten Schulen", an den Gymnasien, vergeben. Dadurch, dass das Abitur als schulische Abschlussprüfung an die Stelle einer universitären Aufnahmeprüfung trat, erhielten die Höheren Schulen bei der Vergabe der Hochschulzugangsberechtigung nahezu ein Monopol. Absolventen (zunächst waren dies ausschließlich junge Männer) erwarben von da an mit einer bestan-

denen Abiturprüfung das Recht zum Universitätsstudium. Anders als in anderen Ländern wie zum Beispiel in den Vereinigten Staaten konnten und können seither diese Schulabsolventen und -absolventinnen mit einer bestandenen Abiturprüfung nicht von Universitäten abgewiesen werden. Das damit installierte Berechtigungswesen fand nur da eine Begrenzung, wo die Kapazitätsgrenzen der Universitäten die Zahl der aufzunehmenden Studienberechtigten limitierte (vgl. dazu ausführlicher Tenorth, 1975, S. 222 ff.). Aber auch in diesen Fällen, in denen ein Numerus clausus die Zahl der aufzunehmenden Studierenden begrenzte, blieb die Mitwirkung der Schulen beim Hochschulzugang bestehen: über die Vergabe des Abiturzeugnisses als grundlegende Voraussetzung des Hochschulzugangs und zusätzlich über die darin vergebenen Noten – wenn auch wie zum Beispiel beim „Medizinertest" eingeschränkt. Aus der Tatsache, dass die Studienberechtigung an einer abgebenden Schule und nicht an einer aufnehmenden Universität vergeben wurde, folgte zwangsläufig seitens der deutschen Länder die Notwendigkeit einer wechselseitigen Anerkennung der Abiturprüfung.

Während die Vergabe des Abiturs als unverzichtbare Voraussetzung für ein universitäres Studium seit dem frühen 19. Jahrhundert unangetastet geblieben ist, weitete sich das Spektrum der Schulen, an denen dieses Abitur erworben werden konnte, im Verlauf der etwa zweihundert Jahre seither kontinuierlich aus. Zu den Schulen der „ersten Stunde", also zu den altsprachlichen Gymnasien, kamen weitere Schultypen mit abiturführenden Bildungswegen: Das waren zunächst die Realgymnasien und die Oberrealschulen, die 1900 das Recht zur Feststellung der Hochschulreife erhielten, dann wenige Jahre danach im ersten Jahrzehnt des zwanzigsten Jahrhunderts die Höheren Mädchenschulen, die Lyzeen. In der zweiten Hälfte des zwanzigsten Jahrhunderts erhielten eine Reihe anderer weiterführender allgemeinbildender Schulen das Recht, eine Abiturprüfung durchzuführen: Gesamtschulen, Integrierte Sekundarschulen, Oberschulen, Stadtteilschulen und Gemeinschaftsschulen konnten gleichfalls eine Hochschulzugangsberechtigung vergeben. An ihre Seite traten noch die beruflichen Gymnasien.

Die im Laufe der Entwicklung eingeleitete Entmonopolisierung des Gymnasiums hat dazu geführt, dass von den 40 Prozent eines Altersjahrgangs, die heute die allgemeine Hochschulreife, das „klassische" Abitur, erwerben, die dahin führende Abiturprüfung an den folgenden Schultypen abgelegt haben (eigene Berechnungen nach KMK, 2020 und Statistisches Bundesamt, 2020b, 2019):

- zu 72 Prozent an Gymnasien,
- zu 17 Prozent an beruflichen Schulen,
- zu 9 Prozent an Gesamtschulen (darunter fallen auch Gemeinschaftsschulen, Stadtteilschulen, Oberschulen, integrierte Sekundarschulen) und zu jeweils 1 Prozent an Freien Waldorfschulen und an Schulen des zweiten Bildungsweges.

Hinzu kam und kommt bis heute die kleine Zahl derer, die auf dem Wege einer externen Prüfung ihr Abitur erlangen konnten, und die neuerdings wachsende Zahl derer, die auf der Basis einer beruflichen Ausbildung ohne Abitur eine Studienzulassung erhalten können.

4 Zur Begrenzung des Zugangs zu Universitäten

Der Weg zum Universitätsstudium ist im Verlauf der Jahre vom ausgehenden 18. Jahrhundert bis in unsere Gegenwart immer wieder versperrt oder zumindest erschwert worden: zum Teil zur Abwehr einer tatsächlichen oder doch unterstellten „Überfüllung" der Universitäten, zum Teil aber auch infolge der Zugehörigkeit zum weiblichen Geschlecht oder zu einer sozial schwachen Gruppe der Bevölkerung.

4.1 Überfüllungskrisen

2004 konstatiert Titze in seinem Aufsatz zu „Bildungskrisen und sozialer Wandel 1780–2000": „Mangel befördert Integration. Überfüllung befördert Differenzierung" (Titze, 2004, S. 366). Mit Blick auf Zeiten der Überfüllung heißt es bei ihm: „Die Phasen der Berufsüberfüllung standen unter dem Vorzeichen der Differenzierung und Abgrenzung. [...] Die wichtigsten politischen Eingriffe zur Normierung von Bildungsprozessen und die Durchsetzung von Grenzwerten der Berechtigung wurden in solchen Phasen vorgenommen" (Titze, 2004, S. 366). Dafür ist das 1788 beschlossene „Reglement für die Prüfung an den Gelehrten Schulen" ein frühes Beispiel. Dieses Reglement beginnt mit einer Passage, die seine Intention unmissverständlich klarstellt: „Es ist bisher vielfältig bemerkt worden, daß so viele zum Studieren bestimmte Jünglinge ohne gründliche Vorkenntnisse unreif und unwissend zur Universität eilen [...]" (Schwartz,1910, S. 122). Neben dem darin enthaltenen Hinweis auf mangelnde Studierfähigkeit muss dieser Einleitungssatz auch vor dem Hintergrund der Überfüllungsdiskussion in der zweiten Hälfte des 18. Jahrhunderts verstanden werden. Das Abiturreglement sollte die fachlichen Studienvoraussetzungen gewährleisten und zugleich der Sicherung eines Gleichgewichts zwischen der Zahl der Examinierten und den zu besetzenden Posten dienen (vgl. dazu ausführlicher Kamp, 1988, und da den Abschnitt „Die Überfüllungsdebatte des 18. Jahrhunderts", S. 43–60).

Die damit gegebene Steuerungsfunktion beeinflusst die Entwicklung des Abiturs seither. Dies wird ein weiteres Mal, nahezu einhundert Jahre nach dem ersten Abiturreglement, im letzten Viertel des 19. Jahrhunderts überdeutlich. Müller beschreibt die wirtschaftliche Lage der Jahre nach der Reichsgründung von 1871: „Nach einer über zwanzigjährigen wirtschaftlichen Hochkonjunkturphase beginnt 1873 eine der schwersten deutschen Konjunkturkrisen" (Müller, 1977, S. 274). Und weiter: „Die Schulreform im letzten Drittel des 19. Jahrhunderts wird bestimmt und legitimiert durch das Schreckgespenst einer Überproduktion akademischer Qualifikationen. Der Jargon der Überfüllungsideologie wird direkt den wirtschaftlichen Konjunkturdiagnosen der Zeit entnommen. Er findet seine prägnanteste und gefährlichste Formulierung im Schlagwort vom ‚Abiturientenproletariat'" (ebd., S. 278). Die politische Reaktion darauf zeigte sich in der 1892 verfügten drastischen Erhöhung des Schulgeldes, das für den Besuch eines Gymnasiums entrichtet werden musste. Dazu hatte der Reichskanzler Bismarck 1890 in einer Eingabe an den preußischen König und deutschen Kaiser geraten – mit der Begründung, durch eine Anhebung der Preise die

Unbemittelten so weit wie möglich von höherer Bildung und damit auch vom Abitur fernzuhalten (vgl. dazu Herrlitz & Titze, 1977, S. 358).

Ein weiteres Mal kam es unmittelbar nach der nationalsozialistischen Machtergreifung in Deutschland mit dem bereits am 25.04.1933 erlassenen „Gesetz gegen die Überfüllung deutscher Schulen und Hochschulen" zu einer bildungspolitischen Reaktion auf eine erwartete Überfüllungssituation (Reichsministerium des Inneren, 1933, S. 225). In den Zwanzigerjahren hatte es einen starken Anstieg der Schulabgänger mit einer allgemeinen Hochschulreife gegeben, sodass zu Beginn der Dreißigerjahre infolge dieses Anstiegs, der mit den Folgen der Weltwirtschaftskrise einherging, eine deutliche Verschlechterung der Berufsaussichten für Akademiker befürchtet wurde. Um diese Entwicklung, die dann allerdings nicht eintrat (vgl. dazu Zymek, 1989, S. 190), abzuwehren und um in Verbindung damit den Zugang jüdischer Studentinnen und Studenten zurückzudrängen, wurde das oben erwähnte Gesetz verabschiedet. In § 1 dieses Gesetzes heißt es, durch eine Begrenzung der Zahl der Schülerinnen und Schüler sowie des Zugangs zu Universitäten solle „die gründliche Ausbildung gesichert und dem Bedarf der Berufe" Genüge geleistet werden. Dazu wurden die Landesregierungen in § 2 verpflichtet, die Aufnahmezahl der Schülerinnen und Schüler der Höheren Schulen und der Studenten und Studentinnen festzulegen. Nur 1,5 Prozent aller neu Aufgenommenen durften Deutsche mit einer nichtarischen Abstammung sein. In einer „Anordnung des Reichsministeriums des Inneren über die zahlenmäßige Begrenzung des Zugangs zu den Hochschulen" (Reichsministerium des Inneren, 1934, S. 16 f.) vom 28.12. des gleichen Jahres wurde dann festgelegt, dass im Jahr 1934 insgesamt nur 15.000 Abiturientinnen bzw. Abiturienten ein Hochschulstudium aufnehmen durften, darunter nur 10 Prozent, also 1.500, Frauen.

Nach 1945 taucht die nun schon tradierte Klage über eine zu hohe Zahl von Studienberechtigten erneut auf. Mit Blick auf kapazitative Engpässe in zahlreichen Studiengängen beschloss die Westdeutsche Rektorenkonferenz (WRK) 1968 die Einführung eines Numerus Clausus, der den Hochschulzugang für Studiengänge, in denen die Aufnahmekapazität nicht für alle Bewerberinnen und Bewerber reichte, vom Durchschnitt der erreichten Abiturnoten abhängig machte. Diese Orientierung an Abiturnoten unterstellt implizit, dass für vergleichbare schulische Leistungen vergleichbare Noten vergeben werden. Bereits die erste länderspezifische Auswertung der PISA-Studie (Baumert et al. 2002) entzog dieser Annahme die Grundlage: Der Ländervergleich belegt bei den Schülerinnen und Schülern der neunten Jahrgangsstufe der Gymnasien erhebliche Leistungsunterschiede zwischen den einzelnen Bundesländern – Leistungsunterschiede, die sich im regelmäßig durch die Kultusministerkonferenz veröffentlichten Notendurchschnitt der Abiturzeugnisse der einzelnen Bundesländer nicht wiederfinden (vgl. zuletzt KMK 2021a).

Ganz unabhängig von der an der Aufnahmekapazität der Hochschulen orientierten Forderung nach Zugangsbeschränkungen wird bis in die jüngste Zeit auch immer wieder die Studierfähigkeit zahlreicher Studienberechtigter thematisiert. So kritisiert z. B. Nida-Rümelin in seiner Arbeit zum „Akademisierungswahn" die „unbegrenzte Ausweitung des Hochschulzugangs" (Nida-Rümelin, 2014, S. 20) und die zu erwar-

tende „Bildungskatastrophe“, die bei einer „Fortsetzung des eingeschlagenen Weges“ zu erwarten sei – eines Wegs, der dazu geführt habe, dass 2013 „die Zahl derjenigen, die ein Studium aufnahmen, höher war als die Zahl derjenigen, die eine Lehre begannen“ (Nida-Rümelin, 2014, S. 21 – vgl. zu der Kritik an diesen Daten Klemm, 2016). Besorgt formuliert er: „Das Spezifikum eines wissenschaftlichen Studiums, nämlich die Forschungsorientierung, ginge verloren, und eine allgemeine oberflächliche Kompetenzorientierung würde Fachwissen generell entwerten“ (Nida-Rümelin, 2014, S. 22). Seine Empfehlung lautet – auch mit einem Verweis auf ein Lehrkräfteüberangebot, das 2013 von der KMK in einer Fehleinschätzung der späteren Entwicklung erwartet wurde (Nida-Rümelin, 2014, S. 198 f.): „Stopp des Akademisierungswahns, das heißt kein weiterer Anstieg, sondern eine Verminderung der Studienanfängerquoten zukünftiger Jahrgänge“ (Nida-Rümelin, 2014, S. 128). Die Instrumente, die dazu hätten eingesetzt werden können und zu denen Nida-Rümelin sich nicht äußert, wären eine Erschwerung des Zugangs zu abiturführenden Bildungswegen, eine stärkere interne Selektion im Verlauf dieser Bildungswege oder eine deutliche Erschwerung der Hochschulreifeprüfung gewesen. Flankiert wird seine Verfallsklage durch Erklärungen aus den Universitäten. So wird der Präsident der Hochschulrektorenkonferenz, Peter-André Alt, in der FAZ vom 18.06.2019 mit dem Satz zitiert: „Wir leben in der Fiktion, dass mit dem Abitur die Voraussetzungen für das Studium erfüllt sind. Die Realität zeigt: Viel zu oft stimmt das nicht“ (Ochmann, 2019). Eine politische Reaktion auf diese Kassandra-Rufe steht einstweilen aus.

4.2 Geschlechtsspezifische Barrieren

Noch einmal soll an dieser Stelle an den ersten Satz des Abiturreglements von 1788 erinnert werden: „Es ist bisher vielfältig bemerkt worden, daß so viele zum Studieren bestimmte Jünglinge ohne gründliche Vorkenntnisse unreif und unwissend zur Universität eilen […]“ (Schwartz 1910, S. 122) Dieser Satz mit seiner Beschränkung auf „Jünglinge“ charakterisiert die Entwicklung der Höheren Schulen für mehr als einhundert Jahre. Erst im Verlauf des ersten Jahrzehnts des zwanzigsten Jahrhunderts entstand in den einzelnen Reichsländern ein Höheres Schulwesen für Mädchen, das den Frauen den Weg zur Immatrikulation ebnete: 1900 in Baden, 1903 in Bayern, 1904 in Württemberg, 1906 in Sachsen und dann 1908 auch in Preußen (Kraul, 1991, S. 289). In Preußen führte dieser Weg – beginnend mit sechs Lebensjahren – nach sieben oder acht Schuljahren in eine Studienanstalt, die sich in ihrer curricularen Ausrichtung an den seit 1900 in Preußen gleichberechtigt nebeneinander angebotenen gymnasialen Typen orientierte. Sie boten Oberrealschulkurse, realgymnasiale Kurse und gymnasiale Kurse an, die gleichermaßen zur Hochschulreife führten (Kraul, 1991, S. 286 f.).

Dieses in Preußen erst 1908 gewonnene Recht der Mädchen zur gleichberechtigten Teilhabe an abiturführenden Bildungsgängen und an universitären Studien währte nicht lange: Mit dem bereits erwähnten „Gesetz gegen die Überfüllung deutscher Schulen und Hochschulen“ wurde 1933 das Recht der Frauen zum Hochschulzugang massiv eingeschränkt. Innerhalb der generell erfolgten Reduzierung der Zahl

der Studienanfängerinnen und -anfänger auf 15.000 je Jahr wurden den jungen Frauen lediglich 10 Prozent der Studienplätze eingeräumt (vgl. Abschnitt 4.1). Verbunden war diese Zurückstellung der Frauen auch mit curricularen Restriktionen (vgl. im Folgenden den Abschnitt 5.3). Die 1938 geschaffene Oberschule für Mädchen kannte nur zwei Formen: die hauswirtschaftliche Form (mit Englisch) und die sprachliche Form (mit Englisch und Latein bzw. einer weiteren lebenden Fremdsprache). Die daneben bestehende Sonderform, das Gymnasium mit Latein, Griechisch und Englisch als Pflichtsprachen, blieb den Jungen vorbehalten.

In den Jahren nach 1945 verschwanden diese geschlechtsspezifischen Barrieren grundsätzlich. Auch wenn Peisert 1967 die Ungleichheit der Bildungschancen noch mit dem „griffigen" Wort vom „katholischen Arbeitermädchen auf dem Lande" beschrieb, einer Kunstfigur, die die konfessions-, schicht-, geschlechts- und regionalspezifische Ungleichheiten „bündelte", gelang den Mädchen und jungen Frauen seither eine fulminante Auf- und Überholjagd: 2018/19 waren von allen Schulabsolventen mit einer allgemeinen Hochschulreife 55 Prozent junge Frauen (eigene Berechnungen nach KMK 2020 und Statistisches Bundesamt 2019 sowie 2020b).

4.3 Schichtspezifische Barrieren

Der früheste Hinweis auf herkunftsbezogene Barrieren beim Zugang zu Universitäten findet sich wiederum im ersten preußischen Abiturreglement. Während in den Jahren bis 1834 ein Universitätsstudium auch ohne eine bestandene Abiturprüfung begonnen werden konnte, erhielten nur diejenigen, die die Abiturprüfung bestanden hatten, ein Stipendium. Söhne aus ärmeren Familien, die ohne ein solches Stipendium nicht studieren konnten, mussten – anders als die aus wohlhabenden Familien – ein Zeugnis der Reife vorlegen. Die Kopplung eines Universitätsstudiums an die ökonomischen Ressourcen der Herkunftsfamilien Studierwilliger hat seither Tradition. Erinnert sei in diesem Zusammenhang an den Rat Bismarcks an Kaiser Wilhelm II., der in der Überfüllungskrise des ausgehenden 19. Jahrhunderts durch eine Schulgelderhöhung die Zahl der Kinder „Unbemittelter" beim Zugang zu höherer Bildung und damit zum Abitur zurückzudrängen vorschlug (vgl. dazu Abschnitt 4.1 dieses Beitrages). Auch wenn das Schulgeld, das bereits 1919 in Artikel 145 der Weimarer Verfassung für die damaligen Volksschulen abgeschafft wurde, dann auch mit deutlicher Verzögerung in der zweiten Hälfte der Fünfzigerjahre in der DDR und in den Ländern der damaligen Bundesrepublik generell abgeschafft wurde, bleibt die selektive Wirkung beschränkter ökonomischer und sozialisatorischer familialer Ressourcen bis in unsere Tage unverkennbar wirksam: So erhielten Kinder aus sozial „starken" Familien im Vergleich zu denen aus sozial „schwachen" Familien bei gleichen Fähigkeiten am Ende der Grundschulzeit mit einer mehr als dreifachen Wahrscheinlichkeit von ihren Lehrkräften eine Empfehlung zum Übergang in ein Gymnasium (Hußmann et al. 2017, S. 245). Unter anderem dies trug dazu bei, dass ausweislich der 21. Sozialerhebung des Deutschen Studentenwerks mit 66 Prozent zwei Drittel der Studierenden „aus einem Elternhaus stammen, in dem Vater und/oder Mutter das Abitur abgelegt haben. [...] Studierende aus Familien, in denen kein Elternteil einen Schulabschluss vorweisen kann, sind die absolute Ausnahme (1 %)" (BMBF 2017, S. 26).

5 Zur curricularen Orientierung

Die Entwicklung der im Abiturexamen schriftlich geprüften Fächer und der diesen Fächern vorgelagerten Lehrpläne lässt sich in insgesamt fünf größere Phasen unterteilen. Diese Phasen gliedern den folgenden Abschnitt.

5.1 Die Monopolstellung des altsprachlichen Gymnasiums in den Jahren von 1788 bis 1890

Die Entwicklung des preußischen Gymnasiums ist in den ersten Jahrzehnten nach der Einführung des Abiturientenexamens 1788 durch die Etablierung eines zum Abitur führenden gymnasialen Lehrplans gekennzeichnet. Ein solcher Lehrplan bildete sich in mehreren Schritten von dem 1812 erlassenen „Edikt über die Abiturientenprüfung“ über das „Reglement für die Prüfung der zu den Universitäten übergehenden Schüler“ von 1834 bis hin zum preußischen Lehrplan für den gymnasialen Unterricht von 1837 heraus. Der von da an durchgängig neunjährige gymnasiale Bildungsgang umfasste 280 Wochenstunden, von denen 46 Prozent den Fächern Latein und Griechisch, 12 Prozent der Mathematik, 9 Prozent der Geschichte und Geographie, 8 Prozent dem Deutschen und 4 Prozent dem Französischen gewidmet waren. Die Abiturprüfungen bestanden aus sechs Prüfungsarbeiten: aus drei Aufsätzen in deutscher, lateinischer und französischer Sprache, der Übersetzung eines Textes ins Lateinische sowie eines Textes aus dem Griechischen und schließlich einer Prüfungsarbeit in Mathematik. Hinzu kam eine in Gruppen durchgeführte mündliche Prüfung zu Themen aus der Geschichte, der Geographie, der Naturbeschreibung, der Philosophie und der Religion (vgl. Reble, 1975, S. 61–68; Bölling, 2010, S. 33 f.).

Die bisher mit Bezug auf Preußen beschriebene Entwicklung findet sich in vergleichbarer Weise in anderen Ländern des späteren deutschen Reichs: So führte Bayern schon 1809 als Voraussetzung zur Zulassung zu einer Universitätsstudium eine an den Gymnasien bestandene „Absolutorialprüfung“ ein (Paulsen, 1885, S. 651 ff.). Ähnlich verfuhren 1811 Württemberg (ebd., S. 664 ff.), 1823 Baden (ebd., S. 670) und 1829 auch Sachsen (ebd., S. 640). Einen ersten Abstimmungsversuch zwischen deutschen Ländern, hier den Mitgliedstaaten des 1815 auf dem Wiener Kongress gegründeten Deutschen Bundes, stellt 1834 das Schlussprotokoll der Wiener Ministerkonferenzen mit den „Sechzig Artikeln“ dar. Unter anderem enthält dieses Dokument mit Artikel 43 eine Vereinbarung, die den Zugang zu Universitäten in den Mitgliedstaaten betraf. Dieser Artikel lautet:

> „Ein Studirender, welcher um die Immatrikulation nachsucht, muß der Commission vorlegen: Wenn er das akademische Studium beginnt, ein Zeugniß seiner wissenschaftlichen Vorbereitung zu demselben und seines sittlichen Betragens, wie solches durch die Gesetze des Landes, dem er angehört, vorgeschrieben ist. Wo noch keine Verordnungen hierüber bestehen, werden sie erlassen werden. Die Regierungen werden einander von ihren über die Zeugnisse erlassenen Gesetzen durch deren Mittheilung an die Bundesversammlung in Kenntniß setzen.“ (Zit. n. Huber 1978)

Zur Benotung der erbrachten Leistungen sowie über die Vergleichbarkeit vergebener Benotungen finden sich keine Verabredungen (Huber, 1978, S. 144).

Erstmalig liegt mit dem Schlussprotokoll der Wiener Ministerkonferenzen ein Dokument vor, in dem sich die deutschen Staaten – Jahrzehnte vor der Reichsgründung von 1871 – verpflichten, die Zulassung zur Universität an ein Zeugnis über eine angemessene Vorbereitung zu binden und diese dann auch wechselseitig anzuerkennen. Diese Feststellung wird nicht dadurch infrage gestellt, dass es in der Vereinbarung nicht um curriculare oder qualitative Fragen ging, sondern um politische Kontrolle. Wolter beschreibt dies so: Den die Vereinbarung schließenden Staaten ging es in erster Linie darum, „die Reifeprüfung als Instrument zur Gegensteuerung im Interesse des Restaurationsstaates nutzen zu können, gegen alle Ansprüche auf demokratische Teilnahme und staatsbürgerliche Gleichheit gerichtet" (Wolter, 1989, S. 24).

In Fortsetzung dieses Ansatzes zu einer wechselseitigen Anerkennung der Abiturprüfung einigten sich nach der Reichsgründung die zum Deutschen Reich gehörenden Länder bei Fortbestand der bestehenden Länderhoheit in Bildungsfragen 1874 auf eine „Gegenseitige Anerkennung der Maturitätszeugnisse der Gymnasien in den Staaten des Deutschen Reiches" (Zentralblatt für die gesamte Unterrichtsverwaltung in Preußen, 1874, 144). Für die Anerkennung mussten insbesondere die folgenden Voraussetzungen gegeben sein:

- Die Schuldauer des Gymnasiums dauert neun Jahre (Punkt 1 der Vereinbarung).
- Schriftliche Prüfungsarbeiten sind ein deutscher Aufsatz, eine lateinische Arbeit und die Lösung mathematischer Aufgaben. Es bleibt den Anordnungen der einzelnen Staaten überlassen, darüber hinaus auch eine Übersetzung ins Deutsche, Griechische, Französische u. a. zu verlangen. Mündliche Prüfungen kommen ergänzend hinzu (Punkte 5 und 7 der Vereinbarung).
- Bei der Erteilung des Zeugnisses der Reife gelten die Anforderungen, die das preußische Prüfungsreglement aufgestellt hat (Punkt 6 der Vereinbarung). Hinweise zu den Benotungen und zu deren Vergleichbarkeit finden sich nicht (ebd.).

Damit war im Reichsgebiet eine Entwicklung abgeschlossen, die in allen Reichsländern dem „humanistischen Gymnasium" bei der Vergabe des Abiturs eine Monopolstellung sicherte.

5.2 Die curriculare Diversifizierung des Abiturs

Die humanistische Ausrichtung des Abiturs erfuhr allerdings bereits wenige Jahre später auf der Reichsschulkonferenz von 1890 eine deutliche Schwächung. In seiner Eröffnungsansprache forderte Kaiser Wilhelm II: „Wir müssen als Grundlage für das Gymnasium das Deutsche nehmen; wir sollen nationale junge Deutsche erziehen und nicht junge Griechen und Römer [...].Wir müssen das Deutsche zur Basis machen. Der deutsche Aufsatz muss der Mittelpunkt sein, um den sich alles dreht" (Reble, 1975, S. 103). Diese eindeutige kaiserliche Vorgabe führte dann dazu, dass in der Lehrplanreform von 1892 die alten Sprachen deutlich an Gewicht verloren und im preußischen Gymnasialabitur nur noch vier schriftliche Prüfungsleistungen erbracht werden mussten: ein deutscher Aufsatz, eine Übersetzung ins Lateinische sowie eine aus dem Griechischen und eine Prüfungsarbeit in Mathematik.

Nach diesem „Zurückfahren“ des Gewichts der alten Sprachen führte dann 1900 in Preußen eine zweite Schulkonferenz dazu, dass die mathematisch-naturwissenschaftlich ausgerichtete Oberrealschule und das neusprachlich geprägte Realgymnasium künftig mit dem humanistischen Gymnasium gleichberechtigt zur allgemeinen Hochschulreife führte. Im humanistischen Gymnasium mussten im Abitur vier schriftliche Arbeiten (Deutsch, Mathematik, Übersetzung aus dem Deutschen ins Lateinische, Übersetzung aus dem Griechischen ins Deutsche) erbracht werden. Im Realgymnasium und in der (lateinlosen) Oberrealschule waren dies insgesamt fünf Prüfungsarbeiten: in beiden Schultypen jeweils eine schriftliche Prüfung in Deutsch und in Mathematik. Im Realgymnasium kamen ein französischer oder englischer Aufsatz oder eine französische oder englische Arbeit/Übersetzung sowie eine Arbeit in Physik hinzu. In der Oberrealschule wurden die schriftlichen Prüfungsfächer Deutsch und Mathematik ergänzt durch einen französischen oder englischen Aufsatz, eine Arbeit/bzw. Übersetzung im Englischen oder Französischen sowie eine Arbeit in Physik oder Chemie (vgl. zu den Lehrplänen Reble, 1975, S. 113 f. und zu den Abiturfächern Bölling, 2010, S. 47). Damit war in Preußen und in der Nachfolge auch in den übrigen Reichsländern (vgl. Bölling, 2010, S. 49) eine – wie es Wolter formuliert – curriculare „Diversifizierung“ (Wolter, 1989, S. 35) der höheren Bildung durchgesetzt: Im Abitur konnte Studierfähigkeit mit der tradierten humanistischen, einer neusprachlichen oder einer naturwissenschaftlich geprägten Bildung erworben werden.

Diese strukturelle und curriculare Ausrichtung überdauerte in ihren Grundzügen die Jahre des Weltkriegs und auch die der Weimarer Republik. In den Beratungen der Reichsschulkonferenz von 1920 konnte sich Paul Oestreich als Vertreter des „Reichsbundes entschiedener Schulreformer“ mit seiner Forderung „[d]ie Reifeprüfung ist alsbald aufzuheben“ (Reichsministerium des Inneren, 1921, S. 927 f.) nicht durchsetzen. Die Mehrheit der auf der Reichskonferenz vertretenen etwa 650 Bildungsexperten votierte dafür, „daß die Zuerkennung von Berechtigungen in erster Linie an die Schlussbeurteilung durch die Schule gebunden ist“ (ebd., S. 829). Diesem Beschluss vorangestellt war die Forderung:

> „Die gegenseitige Anerkennung der Prüfungen und Zeugnisse der Schüler muß auf Grund entsprechender grundsätzlicher Vereinheitlichung der Lehrpläne sowie der Bestimmungen über die Klassenfrequenzen, die Schüleraufnahme, die Versetzungen und die Prüfungen auf das ganze Reich ausgedehnt werden, ebenso müssen die etwaigen Berechtigungen überall die gleichen sein.“ (ebd., S. 829)

Auch wenn die Beratungen der Reichsschulkonferenz nicht unmittelbar zu Umsetzungen in der Schulpolitik führten, kann doch festgestellt werden, dass sie der künftigen Entwicklung der Gymnasien in den Reichsländern während der Jahre der Weimarer Republik den Weg wiesen. Schon 1922 einigten sich die Länder in den „Vereinbarungen der Länder über die gegenseitige Anerkennung der Reifezeugnisse der höheren Schulen“ auf Grundzüge der weiteren Entwicklung des Höheren Schulwesens (Führ, 1970, S. 289 ff.). Für die gegenseitige Anerkennung der Reifezeugnisse gelten die folgenden Voraussetzungen:

- Die Anerkennung des Reifezeugnisses erstreckt sich nur auf Gymnasien, Realgymnasien, Oberrealschulen; es gewährt in allen Reichsländern alle Berechtigungen.
- Der Lehrgang muss neun Jahre umfassen.
- Gegenstand der Reifeprüfung am Ende des neunten gymnasialen Schuljahrgangs sind Deutsch, Geschichte und Mathematik; ferner bei den Gymnasien Latein, Griechisch, Französisch oder Englisch, an den Realgymnasien Lateinisch, Französisch, Englisch und Naturwissenschaften, bei den Oberrealschulen Französisch, Englisch und Naturwissenschaften. Die Prüfung besteht aus einem schriftlichen und einem mündlichen Teil. Der schriftliche Teil ist immer Deutsch und Mathematik, in den Gymnasien zusätzlich Latein und Griechisch, in den Realgymnasien Latein und Französisch oder Englisch, in den Oberrealschulen Französisch oder Englisch und Naturwissenschaften. Über die Notenvergabe und über die Vergleichbarkeit der vergebenen Benotungen enthält die Vereinbarung keine Vorgaben (ebd.).

Das mit Abstand größte Reichsland Preußen bewegte sich mit seinen Regelungen im Rahmen dieser Vereinbarung. In seiner Reifeprüfungsordnung legte Preußen 1926 fest: Wie schon in der Zeit vor dem ersten Weltkrieg blieben in den jetzt vier Hauptformen der Höheren Schulen (humanistisches Gymnasium, Realgymnasium, Oberrealschule und – neu hinzugekommen – Deutsche Oberschule) Deutsch und Mathematik schriftliche Prüfungsfächer. Im Humanistischen Gymnasium kamen Latein und Griechisch hinzu, im Realgymnasium Französisch und Englisch (wahlweise konnte eine der neuen Sprachen durch Latein ersetzt werden), in der Oberrealschule Französisch oder Englisch sowie ein naturwissenschaftliches Fach, in der Deutschen Oberschule Französisch oder Englisch sowie Geschichte oder Geografie (ausführliche zu den Lehrplänen Reble, 1975, S. 150; zu den Prüfungsfächern Bölling, 2010, S. 79).

5.3 Das Höhere Schulwesen im Dritten Reich

In den wenigen Jahren zwischen der nationalsozialistischen Machtergreifung 1933 bis zum Ausbruch des zweiten Weltkriegs kam es in Deutschland nicht zu neuen Abiturprüfungsordnungen. Allerdings erlaubt die 1938 vorgenommene Neuordnung des von da an auf eine achtjährige Schulzeit begrenzten Höheren Schulwesens Rückschlüsse auf die künftig geplanten Schwerpunktsetzungen auch in den Abiturprüfungen (vgl. zu Folgendem: Reichs- und Preußisches Ministerium für Wissenschaft, Erziehung und Volksbildung, 1938, S. 23–31). Die Höheren Schulen waren künftig untergliedert in zwei Hauptformen und in eine Sonderform. Die beiden Hauptformen waren die Oberschule für Jungen und die Oberschule für Mädchen. Die Oberschule für Jungen kannte einen naturwissenschaftlich-mathematischen Zweig und einen sprachlichen Zweig (in beiden Zweigen wurden – mit unterschiedlichem Gewicht – Englisch und Latein unterrichtet). Die Oberschule für Mädchen bot zwei Formen: die hauswirtschaftliche Form (mit Englisch) und die sprachliche Form (mit Englisch und Latein bzw. einer weiteren lebenden Fremdsprache). Neben diesen beiden Hauptfor-

men gab es eine Sonderform, das Gymnasium für Jungen mit Latein, Griechisch und Englisch als Pflichtsprachen. In all diesen Formen wurden Deutsch und Mathematik – wenn auch unterschiedlich stark gewichtet – als Pflichtfächer unterrichtet. Diese erst 1938 verankerte Grundstruktur war in den folgenden Kriegsjahren einem andauernden Erosionsprozess unterworfen – bis hin dazu, dass schon 1942 in östlichen Teilen Deutschlands die schriftlichen Reifeprüfungen ausgesetzt wurden (Bölling, 2010, S. 88).

5.4 Abiturordnungen im Nachkriegsdeutschland

Die Jahre nach 1945 sind durch fundamental unterschiedliche Entwicklungen in den Besatzungszonen und dann in der Deutschen Demokratischen Republik einerseits und in der Bundesrepublik Deutschland andererseits geprägt. Beide deutsche Staaten mussten den Um- und Wiederaufbau ihrer auf universitäre Studien vorbereitenden Schulen auf der Basis infrastrukturell weitgehend zerstörter, personell ausgezehrter und ideologisch hoch belasteter Schulen vorantreiben.

Die Entwicklung in der Deutschen Demokratischen Republik

Die Grundlage der Entwicklung des Schulsystems der späteren DDR wurde bereits 1946 mit dem „Gesetz zur Demokratisierung der deutschen Schule“ gelegt. In § 3 dieses Gesetzes wurde der vierstufige Aufbau des Bildungssystems geregelt: Auf den Kindergarten folgten eine achtjährige Grundschule, eine vierjährige Oberstufe, die zur Hochschulreife führte, und schließlich die Hochschule. Im „Gesetz über die sozialistische Entwicklung des Schulwesens in der Deutschen Demokratischen Republik“ (1959) wurde diese Struktur so weiterentwickelt, dass alle Kinder und Jugendlichen bis zur achten Jahrgangsstufe gemeinsam die Polytechnische Oberschule besuchen mussten. Danach verblieb der weitaus größere Teil von ihnen in der insgesamt zehnjährigen Polytechnischen Oberschule, während ein kleinerer Teil in die vierjährige Erweiterte Oberschule überwechselte, um dort ein Abitur zu erlangen. 1965 wurde dann die bis zum Ende des Staates bestehende Regelung im „Gesetz über das einheitliche sozialistische Bildungssystem“ verankert (Michael & Schepp, 1993, S. 364–378): Danach waren die beiden grundlegenden Bestandteile des einheitlichen sozialistischen Bildungssystems die zehnklassige allgemeinbildende Polytechnische Oberschule sowie die zur Hochschulreife führenden Bildungseinrichtungen. Die wichtigste der zur Hochschulreife führenden Einrichtungen war die Erweiterte Polytechnische Oberschule (EOS), die im Anschluss an die Polytechnische Oberschule in zwei Jahren zum Abitur führte. In der EOS wurden von den je Schuljahr 32 bis 33 Pflichtwochenstunden drei bis vier in deutscher Sprache und Literatur, fünf in Mathematik, sieben bis neun in Naturwissenschaften, vier in Russisch als erster und drei in einer zweiten Fremdsprache unterrichtet. In der abschließenden zentralen Reifeprüfung wurden – neben einer Reihe mündlicher Prüfungen – insgesamt drei fünfstündige Klausuren in Deutsch, Mathematik und einer der Naturwissenschaften geschrieben. Hinzu kam eine eineinhalbstündige Klausur in Russisch. Bemerkenswert ist – gerade auch im Vergleich zu der Entwicklung im Westen Deutschlands – die starke mathematisch-naturwissenschaftliche Ausrichtung der Reifeprüfung in der DDR.

Die Entwicklung in der Bundesrepublik Deutschland bis 1972

Während die Entwicklung der Schulen der DDR mit diesen Bestimmungen nicht nur eine klare Abkehr von der Schule im Nationalsozialismus, sondern auch von der Schule der Weimarer Republik vollzog, knüpfte die Schulpolitik in der Bundesrepublik an die der Weimarer Jahre an. Der Vorgabe der von den Alliierten erlassenen „Grundsätze für die Demokratisierung des deutschen Bildungswesens" aus dem Jahr 1947 wurde nicht gefolgt. Dort hieß es unter Punkt 4: „Die Abschnitte der Elementarbildung und der weiterführenden Bildung sollten zwei aufeinanderfolgende Stufen der Unterweisung bilden, nicht zwei Wege oder Abschlüsse der Unterweisung (nebeneinander), die teilweise übereinstimmen" (Michael & Schepp, 1993, S. 338). Tenorth konstatiert zum Umgang der westdeutschen Bundesländer mit den „Grundsätzen" der Alliierten knapp: „Die von den Alliierten versuchte Neuordnung wird für die Westzone nicht verwirklicht. Dafür sind politische und soziale sowie ökonomische Faktoren ebenso verantwortlich zu machen wie mangelnde sachliche wie politische Kompetenz der Planer" (Tenorth, 1975, S. 131). Und er formuliert weiter: „Die Standards der Hochschulreife gliedern das Bildungssystem, trennen die Bildungsinstitutionen nach der Berechtigung, die sie verleihen, und wirken über die Eingangsforderungen der Gymnasien auch in die allgemeine Grundschule ein" (ebd., S. 132).

Bezüglich des Abiturs ist die Entwicklung der zum Abitur führenden Schulen in der Bundesrepublik bis zur Neugestaltung der gymnasialen Oberstufe durch die KMK von 1972 durch die folgenden „Meilensteine" geprägt:

- Die Tübinger Beschlüsse von 1951, in denen es im Abschnitt „Erste Resolution: Lehrkräfte der Höheren Schule" heißt: „Die Durchdringung des Wesentlichen der Unterrichtsgegenstände hat den unbedingten Vorrang vor jeder Ausweitung des stofflichen Bereichs. Die Zahl der Prüfungsfächer im Abitur sollte eingeschränkt, die Prüfungsmethoden sollten mehr auf Verständnis als auf Gedächtnisleistung abgestellt werden" (Reble, 1975, S. 157).
- Das Düsseldorfer Abkommen von 1955 (ebd., S. 160 ff.) stellt zur Langform des Gymnasiums, das nach der vierjährigen Grundschule neun Jahrgangsstufen umfasst, in den §§ 8 und 9 fest, dass drei Schultypen möglich sind: das altsprachliche, das neusprachliche und das mathematisch-naturwissenschaftliche Gymnasium. Darüber hinaus heißt es in § 14: „Die in den vertragsschließenden Ländern ausgestellten Reifezeugnisse werden nach Maßgabe der jeweiligen Vereinbarungen der Kultusministerkonferenz gegenseitig anerkannt" (Tenorth, 1975).
- Der Tutzinger Maturitätskatalog von 1958 stellt das Ergebnis eines Gesprächskreises von Mitgliedern der Kultusministerkonferenz und der Westdeutschen Rektorenkonferenz dar, der sich 1958 in Tutzing traf und – wie Tenorth unterstreicht – aus der Sicht der Universitäten Hochschulzugangskriterien formulierte (Tenorth, 1975, S. 142). Der „Maturitätskatalog" formuliert für die Hochschulreife insgesamt neun inhaltliche Anforderungen (vgl. Bölling, 2010, S. 102 f.): einwandfreies Deutsch, Verständnis der Meisterwerke der deutschen Literatur und der Weltliteratur, Einführung in eine Fremdsprache, Kenntnisse der Elementarmathematik, Einführung in die Hauptphänomene der Physik, Zugang zu

biologischen Betrachtungsweisen, Geschichte, Verständnis für philosophische Einleitungsfragen, Orientierung über die Christenlehre sowie Einführung in ethische Grundfragen.

- Die Saarbrücker Rahmenvereinbarung von 1960 (Reble, 1975, S. 166 ff.) benennt in Abschnitt III die Gegenstände der Reifeprüfung: Im schriftlichen Teil sind dies in den drei gymnasialen Schultypen jeweils Deutsch und Mathematik, im altsprachlichen Gymnasium dazu noch Latein und Griechisch (oder Französisch), im neusprachlichen Gymnasium zwei Pflichtfremdsprachen, im mathematisch-naturwissenschaftlichen Gymnasium zusätzlich Physik und eine Pflichtfremdsprache. Die Gegenstände der mündlichen Prüfungen sind die vier Fächer der schriftlichen Prüfung, Gemeinschaftskunde sowie ein weiteres Fach.

5.5 Gymnasiale Bildung und das Abitur seit 1972: Vom Typengymnasium zum „Wahlbetrieb“

Mit der Neugestaltung der gymnasialen Oberstufe durch die KMK wurde 1972 ein grundlegender Wandel der Entwicklung der Gymnasien eingeleitet (vgl. den Text bei Reble, 1975, S. 201 ff.). Sie schafft die im Prinzip seit 1900 etablierten unterschiedlichen Gymnasialtypen ab. In Absatz 3.1 der „Vereinbarung zur Neugestaltung der gymnasialen Oberstufe in der Sekundarstufe II“ heißt es dazu: „Der Unterricht in der Oberstufe wird nach Begabung und Leistung differenziert; die Oberstufe wird nicht mehr nach Gymnasialtypen gegliedert“ (ebd.). Mit der Abschaffung der im Prinzip seit 1900 etablierten Gymnasialtypen soll – so heißt es im ersten Abschnitt „Zielsetzung“ – die Stufe des Übergangs zur Hochschule so strukturiert werden, dass „sowohl eine gemeinsame Grundausbildung für alle Schüler gewährleistet ist als auch der individuellen Spezialisierung Raum gegeben ist“ (ebd.) Dieses Ziel soll dadurch erreicht werden, dass der Unterricht in der zweijährigen Qualifikationsphase den Schülerinnen und Schülern in einem Pflicht- und in einem Wahlbereich geboten wird. Der Pflichtbereich umfasst drei Aufgabenfelder: das sprachlich-literarisch-künstlerische, das gesellschaftswissenschaftliche und das mathematisch-naturwissenschaftlich-technische; hinzu kommen Religion und Sport. Der Wahlbereich dient in Verbindung mit dem Pflichtbereich der Schwerpunktbildung. Sowohl im Pflichtbereich wie auch im Wahlbereich wird der Unterricht auf „Grundkurs-“, bzw. auf „Leistungskursniveau“ erteilt. Jede Schülerin und jeder Schüler muss zwei Leistungskurse belegen: Einer davon muss eine Fremdsprache, Mathematik oder eine Naturwissenschaft sein, der zweite muss aus dem breiteren Spektrum von Pflicht- und Wahlangeboten gewählt werden. In der Abiturprüfung werden die Schülerinnen und Schüler in vier Fächern geprüft. Dabei müssen Kenntnisse in den Aufgabenfeldern des Pflichtbereichs, vertiefte Kenntnisse in den gewählten Leistungsfächern nachgewiesen werden. Die Prüfung besteht aus drei Klausuren in den beiden gewählten Leistungskursen und in einem weiteren Grundkurs sowie aus einer mündlichen Prüfung in einem weiteren Grundkursfach. Für die Beruflichen Gymnasien wurde zusätzlich geregelt, dass das für diese Schulen profilgebende berufsbezogene Fach auf erhöhtem Anforderungsniveau unterrichtet und geprüft werden muss (KMK 2021b). Zwar finden sich in

der KMK-Vereinbarung sehr differenzierte Vorgaben dazu, wie die in der Qualifikationsphase und in der Abiturprüfung erbrachten Leistungen und erreichten Noten in eine Gesamtqualifikation einfließen, doch wird die Frage der Vergleichbarkeit von Noten nicht einmal im Ansatz thematisiert. Zur „Sicherung der Gleichwertigkeit der schulischen Ausbildung, der Vergleichbarkeit der Schulabschlüsse sowie der Durchlässigkeit des Bildungswesens in der Bundesrepublik Deutschland" hatte die Kultusministerkonferenz 1978 eine „Vereinbarung über Einheitliche Prüfungsanforderungen in der Abiturprüfung" (EPA) geschlossen; diese Vereinbarung wurde 2007 zuletzt aktualisiert (KMK 2007). Darauf, dass diese „einheitlichen Prüfungsanforderungen" nicht zu einer Vergleichbarkeit der Notengebung in den einzelnen Bundesländern geführt haben, wurde bereits im Abschnitt 4.1 (Überfüllungskrisen) verwiesen.

Schon in den ersten Jahren nach dem KMK-Beschluss über die Neugestaltung der gymnasialen Oberstufe setzten Novellierungen des Beschlusses ein – Novellierungen, die zum einen auf eine Reduzierung des Wahlbereichs und auf eine größere Verbindlichkeit bei der Belegung einzelner Unterrichtsangebote abzielten und zum anderen der Tatsache Rechnung trugen, dass, nach der Vereinigung der beiden deutschen Staaten 1990, neue Gegebenheiten eingetreten waren. Da die neu hinzugekommenen Länder bis zum Beitritt zur Bundesrepublik insgesamt zwölfjährige Bildungsgänge kannten und z.T. auch beibehalten wollten, wurde in der KMK vereinbart, dass die einzelnen Länder – unbeschadet der Zahl der Schuljahre – ab der Jahrgangsstufe 5 bis zum Abitur insgesamt 265 Wochenstunden erteilen müssen (KMK, 2021b, S. 4). Hinsichtlich der Verbindlichkeit einzelner Fächer wurde Folgendes festgelegt:

- Im sprachlich-künstlerischen Aufgabenfeld müssen die Schülerinnen und Schüler in der Qualifikationsphase jeweils vier Schulhalbjahre in Deutsch und in der fortgeführten Fremdsprache belegen.
- Im gesellschaftswissenschaftlichen Aufgabenfeld gilt dies für Geschichte oder ein anderes gesellschaftswissenschaftliches Fach mit einem festen Anteil von Geschichte.
- Im mathematisch-naturwissenschaftlich-technischen Aufgabenfeld müssen jeweils vier Schulhalbjahre in Mathematik und in Biologie oder Chemie oder Physik belegt werden (KMK, 2021b, S. 6).

Die Abiturprüfung umfasst vier bis fünf Prüfungsfächer. Es müssen mindestens drei schriftliche und mindestens ein mündliches Prüfungsfach geprüft werden. Pflichtfächer der schriftlichen Abiturprüfung sind mindestens zwei Fächer mit erhöhtem Anforderungsniveau, darunter mindestens eines der Fächer Deutsch, Fremdsprache, Mathematik oder eine Naturwissenschaft. Gegenstand der mündlichen Prüfung muss ein Fach sein, das nicht schon schriftlich geprüft wurde.

Insgesamt entwickelt sich die Oberstufe des abiturführenden gymnasialen Bildungsgangs nach 1972 von der zunächst stärkeren Ermöglichung individuellerer Schwerpunktsetzungen wieder verstärkt hin zu einer für alle gleichermaßen geltenden größeren Verbindlichkeit.

6 Geschichte der Allgemeinen Hochschulreife: Kontinuitäten im Wandel

Beim Rückblick auf die mehr als zweihundertjährige Geschichte der allgemeinen Hochschulreife werden Entwicklungslinien mit teilweise bemerkenswerten Kontinuitäten erkennbar:

Schon bei der Einführung der ersten Abiturprüfung und seither immer wieder spielte das Interesse an der Herstellung eines tendenziellen Gleichgewichts zwischen der Zahl der zum Studium Zugelassenen und den Beschäftigungsmöglichkeiten künftiger Absolventinnen und Absolventen eine hervorragende Rolle – verbunden mit der Sorge um einen Qualitätsverlust schulischer und universitärer Bildung als Folge einer zu expansiven Bildungsbeteiligung. Die angestrebte „Mengensteuerung" ist während der langen Geschichte des Abiturs auch immer wieder verbunden gewesen mit Abschließungstendenzen: viele Jahre lang gegenüber Mädchen und jungen Frauen, während der Jahre der nationalsozialistischen Herrschaft auch zulasten jüdischer Studentinnen und Studenten und von Beginn an bis in unsere Tage mit Zurückweisung der Kinder sozial schwacher Familien – im 19. Jahrhundert explizit formuliert, in den Jahren danach zwar nicht intendiert, so doch vielfach ungewollt hingenommen.

Neben der intendierten „Mengensteuerung" hat sich die im preußischen Abiturreglement getroffene Entscheidung, die Reifeprüfung an den abgebenden Höheren Schulen und nicht an den aufnehmenden Universitäten durchzuführen, als über die Maßen stabil erwiesen. Mit ihr verbunden war die Installierung des Berechtigungssystems: Wer die Abiturprüfung bestanden hatte, erwarb das Recht, an jeder Universität ein Hochschulstudium aufzunehmen. Daraus folgte die Notwendigkeit der wechselseitigen Anerkennung von Hochschulreifeprüfungen. Implizit waren damit auch die in den einzelnen Ländern für die erbrachten Leistungen erteilten Noten anerkannt, ohne dass die Vergleichbarkeit dieser Noten hätte belegt werden müssen.

Im Verlauf der Geschichte der Hochschulreifeprüfung wiederholten sich Vereinbarungen zu deren wechselseitiger Anerkennung; Damit verbunden sind kontinuierlich wiederkehrende Verabredungen zur Inanspruchnahme von Bildungszeit vom Ende der dem Gymnasium vorgeschalteten Schule bis zum Abitur. Die schon 1837 beschlossene Stundentafel für das preußische Gymnasium sah einen auf neun Jahre verteilten Lehrplan mit insgesamt 280 Wochenstunden vor. Dieses Zeitmuster erwies sich – von den Ausnahmen im nationalsozialistischen Deutschland und dann in der DDR abgesehen – als bemerkenswert stabil. Nach der Vereinigung der beiden deutschen Staaten wollten einige ostdeutsche und auch einzelne westdeutsche Bundesländer eine Schulzeit von in den meisten Ländern vier Grundschuljahren und weiteren acht Jahren bis zum Abitur beibehalten bzw. einführen. So kam es in der Kultusministerkonferenz zu der Vereinbarung, dass der Weg nach der Grundschule bis zum Abitur nicht weniger als 265 Unterrichtsstunden umfassen müsse.

Schließlich kann festgestellt werden, dass sich die in den Abiturprüfungen präsenten Unterrichtsfächer deutschlandweit sehr vergleichbar entwickelten. In all den Jahren dominierten die Fächer Deutsch, Mathematik und Fremdsprachen (zunächst

die alten und dann die lebenden Sprachen) das Prüfungsgeschehen. Die Naturwissenschaften erhielten erst spät und zunächst nur in einem einzelnen Gymnasialtyp den „Rang" eines Abiturprüfungsfaches. Fachgebiete wie Rechts- oder Wirtschaftswissenschaft fanden bis heute keinen oder allenfalls einen marginalen Eingang in die Reihe der im Abitur prüfungsrelevanten Fächer.

Literatur

Baumert, J., Artelt, C., Klieme, E., Neubrand, M, Prenzel, M., Schiefele, U., Schneider, W., Tillmann, K.-J. & Weiß, M. (Hrsg.). (2002). *PISA 2000 – Die Länder der Bundesrepublik Deutschland im Vergleich*. Opladen: Leske + Budrich.

BMBF – Bundesministerium für Bildung und Forschung (2017). *Die wirtschaftliche und soziale Lage der Studierenden in Deutschland 2016*. 21. Sozialerhebung des Deutschen Studentenwerkes. Bonn: o. V.

Bölling, R. (2010). *Kleine Geschichte des Abiturs*. Paderborn: Ferdinand Schöningh.

Führ, Chr. (1970). *Zur Schulpolitik der Weimarer Republik. Die Zusammenarbeit von Reich und Ländern im Reichsschulausschuss (1919–1923) und im Ausschuss für das Unterrichtswesen (1924–1933)*. Darstellung und Quellen. Weinheim: Beltz.

Herrlitz, H.-G. & Titze, H. (1977). Überfüllung als bildungspolitische Strategie. Zur administrativen Steuerung der Lehrerarbeitslosigkeit in Preußen 1870–1914. In: Herrmann, U. (Hrsg.), *Schule und Gesellschaft im 19. Jahrhundert. Sozialgeschichte der Schule im Übergang zur Industriegesellschaft* (S. 348–370). Weinheim: Beltz.

Huber, E. R. (Hrsg.) (1978). *Dokumente zur deutschen Verfassungsgeschichte. Band 1: Deutsche Verfassungsdokumente 1803–1850* (3., neu bearb. Aufl.). Stuttgart: W. Kohlhammer.

Hußmann, A. u.a. (Hrsg.). (2017). *Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.

Kamp, N. (1988). *Das Abiturreglement von 1788. Zur Diskrepanz von Schulverwaltungsanspruch und Wirklichkeit*. Unveröffentlichte Dissertation, Universität – Gesamthochschule Essen.

Klemm, K. (2016). Akademikerschwemme? Anmerkungen zu den Zahlen der Studienanfänger. Schulmanagement. *Die Fachzeitschrift für Schul- und Unterrichtsentwicklung, 1/2016*, 25–26.

KMK (2007). *Vereinbarung über Einheitliche Prüfungsanforderungen in der Abiturprüfung*. (Beschluss der Kultusministerkonferenz vom 01.06.1979 i. d. F. vom 20.09.2007).

KMK (2020). *Schüler, Klassen, Lehrer und Absolventen der Schulen 2009 bis 2018*. Berlin: o. V.

KMK (2021a). *Abiturnoten 2020 (und frühere Jahrgänge)*.Berlin.

KMK (2021b). *Vereinbarung zur Gestaltung der gymnasialen Oberstufe und der Abiturprüfung (Beschluss der Kultusministerkonferenz vom 07.07.1972 i. d. F. vom* 18.02.2021).

Kraul, M. (1991). Höhere Mädchenschulen. In: Berg, Chr. (Hrsg.), *Handbuch der deutschen Bildungsgeschichte, Band IV, 1870–1918. Von der Reichsgründung bis zum Ende des Ersten Weltkriegs* (S. 279–303). München: C. H. Beck.

Michael, B. & Schepp, H.-H. (1993). *Die Schule in Staat und Gesellschaft. Dokumente zur deutschen Schulgeschichte im 19. Und 20. Jahrhundert*. Göttingen: Muster-Schmidt.

Müller, D. K. (1977). *Sozialstruktur und Schulsystem. Aspekte zum Strukturwandel des Schulwesens im 19. Jahrhundert*. Göttingen: Vandenhoeck & Ruprecht.

Nida-Rümelin, J. (2014). *Der Akademisierungswahn. Zur Krise beruflicher und akademischer Bildung*. Hamburg: Edition Körber-Stiftung.

Ochmann, M. (2019). *Trotz Abiturs nicht reif für den Hörsaal*. FAZ. https://www.faz.net/aktuell/rhein-main/studenten-brauchen-aufbaukurse-16308768.html

Paulsen, F. (1885). *Geschichte des Gelehrten Unterrichts auf den deutschen Schulen und Universitäten vom Ausgang des Mittelalters bis zur Gegenwart*. Leipzig: Veit & Comp.

Peisert, H. (1967). *Soziale Lage und Bildungschancen in Deutschland*. München: Piper.

Reble, A. (Hrsg.). (1975). *Zur Geschichte der Höheren Schule. Band II (19. und 20. Jahrhundert)*. Bad Heilbrunn: Julius Klinkhardt.

Reichsministerium des Inneren (1921). *Die Reichsschulkonferenz 1920. Ihre Vorgeschichte und Vorbereitung und ihre Verhandlungen*. Leipzig: Quelle und Meyer.

Reichsministerium des Inneren (1933). *Reichsgesetzblatt – Teil I, Jahrgang 1933*. Berlin: Reichsverlagsamt.

Reichsministerium des Inneren (1934). *Reichsministerialblatt der inneren Verwaltung*, 62. Jahrgang. Berlin: o. V.

Reichs- und Preußisches Ministerium für Wissenschaft, Erziehung und Volksbildung (1938). *Erziehung und Unterricht in der Höheren Schule*. Berlin: Weidmannsche Verlagsbuchhandlung.

Schwartz, P. (1910). *Die Gelehrtenschulen Preußens unter dem Oberschulkollegium (1787–1806) und das Abiturexamen. Erster Band*. Berlin: Weidmannsche Buchhandlung.

Statistisches Bundesamt (2019). *Bildung und Kultur. Berufliche Schulen. Schuljahr 2018/19*, Fachserie 11, Reihe 2. Wiesbaden: o. V.

Statistisches Bundesamt (2020a). *Privatschulen in Deutschland – Fakten und Hintergründe*. Wiesbaden: o. V.

Statistisches Bundesamt (2020b). *Bildung und Kultur. Allgemeinbildende Schulen. Schuljahr 2018/19*, Fachserie 11, Reihe 1. Wiesbaden: o. V.

Tenorth, H-E. (1975). *Hochschulzugang und gymnasiale Oberstufe in der Bildungspolitik von 1945–1973*. Bad Heilbrunn: Julius Klinkhardt.

Titze, H. (2004). Bildungskrisen und sozialer Wandel 1780–2000. In: *Geschichte und Gesellschaft, 30*(2), 339–372.

Wolter, A. (1989). *Von der Elitenbildung zur Bildungsexpansion. Zweihundert Jahre Abitur (1788–1988)*. Oldenburg: Bibliotheks- und Informationssystem der Universität Oldenburg.

Zentralblatt für die gesamte Unterrichtsverwaltung in Preußen – 1859–1934. http://www.bbf.dipf.de/cgi-opac/catalog.pl?t_digishow=x&zid=2a1811

Zymek, B. (1989). Schulen. In: Langewiesche, D. & Tenorth, H.-E. (Hrsg.), *Handbuch der deutschen Bildungsgeschichte Band V. 1918–1945. Die Weimarer Republik und die nationalsozialistische Diktatur* (S. 155–208). München: C. H. Beck.

2 Jüngere Entwicklungen bei Abitur und Abiturprüfungen in Deutschland

Lars Hoffmann, Pauline Schröter & Petra Stanat

Zusammenfassung

Blickt man auf die Entwicklungen des Abiturs und der Abiturprüfungen in Deutschland in den letzten drei Dekaden, wird ein Prozess erkennbar, der durch eine zunehmende Standardisierung verschiedener Aspekte der Prüfungsorganisation innerhalb einzelner Länder sowie – in jüngerer Zeit – durch eine Annäherung der Länder bei den landesspezifischen Regelungen zu den Abiturprüfungen gekennzeichnet ist. Der vorliegende Beitrag, der historisch an die Ausführungen im ersten Beitrag dieses Bandes anknüpft, gibt einen Überblick über einige wesentliche Etappen dieses Standardisierungsprozesses. Dabei wird auf die ersten vier Etappen – die Neuordnung der gymnasialen Oberstufe, die Einführung des Zentralabiturs, die Überarbeitung der Einheitlichen Prüfungsanforderungen in der Abiturprüfung und die Entwicklung der Bildungsstandards für die Allgemeine Hochschulreife – eher knapp eingegangen. Die fünfte und aktuellste Etappe – die Entwicklung von Gemeinsamen Abituraufgabenpools der Länder – wird hingegen etwas genauer in den Blick genommen.

1 Einleitung

In Anknüpfung an den ersten Beitrag dieses Bandes hat der vorliegende Beitrag das Ziel, einen Überblick über die jüngeren Entwicklungen des Abiturs und der Abiturprüfungen in Deutschland zu geben. Dabei wird grob der Zeitraum der letzten 30 Jahre in den Blick genommen. In diesem Zeitraum fanden mehrere Reformen der gymnasialen Oberstufen der Länder statt, die entweder direkt die Abiturprüfungen betrafen oder sich zumindest mittelbar auf diese auswirkten. In der Rückschau lässt sich dabei ein Prozess erkennen, der durch eine zunehmende Standardisierung verschiedener Aspekte der Prüfungsorganisation innerhalb einzelner Länder sowie – in jüngerer Zeit – durch eine Annäherung der Länder bei den landesspezifischen Regelungen zu den Abiturprüfungen und Prüfungsaufgaben gekennzeichnet ist.

Die Ursachen und Motoren dieser Standardisierungs- und Annäherungsprozesse können an dieser Stelle nur skizziert werden. In den von uns betrachteten Zeitraum fallen historische Wendepunkte wie die deutsche Wiedervereinigung sowie globale, auch heute noch wirksame allgemeine Trends wie Migration, Internationalisierung und Globalisierung, die einen tiefgreifenden gesellschaftlichen Wandel in unterschiedlichen Bereichen auslösten. Als eine Begleiterscheinung solcher Entwicklun-

gen standen zum Beispiel Universitäten und Hochschulen vor der Frage, wie mit den in der DDR oder in anderen Staaten erworbenen Hochschulzugangsberichtigungen umzugehen ist. Diese Frage, wie auch die Öffnung von Studiengängen für Bewerberinnen und Bewerber ohne Abitur, machten deutlich, dass ein gewisser Standardisierungsbedarf im Sinne einer präziseren Definition von allgemeiner Hochschulreife und Studierfähigkeit sowie ihrer konstituierenden Elemente besteht (Blossfeld et al., 2011).

Als zentraler Treiber des Standardisierungs- und Annäherungsprozesses der letzten Jahre ist aber zweifellos die beständige, bis heute andauernde Debatte um die vermeintlich mangelnde Vergleichbarkeit der Abiturprüfungen der Länder sowie der dabei erzielten Noten hervorzuheben. Kritisch diskutiert wurde hierbei etwa das mit der Oberstufenreform von 1972 eingeführte System aus Grund- und Leistungskursen, das umfangreiche Wahloptionen bei der Fächerbelegung in der Qualifikationsphase gestattete. Zudem bot es in vielen Ländern die Möglichkeit, zentrale Kernfächer wie Deutsch, Mathematik oder die erste Fremdsprache entweder gänzlich abzuwählen oder zumindest nicht in die Abiturgesamtnote einzubringen (Neumann, 2010). Weitere Kritikpunkte betrafen etwa Länderunterschiede in den inhaltlichen Vorgaben der Lehrpläne und Curricula für die gymnasiale Oberstufe, die trotz der im Jahr 1979 eingeführten Einheitlichen Prüfungsanforderungen in der Abiturprüfung (EPA) der KMK bestanden, sowie den geringen Standardisierungsgrad der schriftlichen Abiturprüfungen, die bis Mitte der 2000er-Jahre in vielen Ländern nicht zentral gestellt wurden (Klein et al., 2009). Im Mittelpunkt der Debatte um die Vergleichbarkeit der Abiturprüfungen der Länder steht jedoch die bis heute immer wieder diskutierte Annahme, dass die im Abitur erzielten Noten nur bedingt mit dem tatsächlichen Kompetenzniveau der Schülerinnen und Schüler korrespondieren. Insbesondere wird vermutet, dass es je nach Land, und zum Teil auch je nach Schule, unterschiedlich anspruchsvoll sei, gute Abiturnoten zu erzielen (z. B. Blossfeld et al., 2011). Tatsächlich ist diese Vermutung sehr plausibel: Es liegen zwar keine aktuellen Studien vor, die derartige Länderunterschiede bei der Benotung im Abitur in einer bundesweiten Perspektive und für eine größere Anzahl von Fächern systematisch untersucht hätten; jedoch konnten in der Vergangenheit zumindest für das Fach Mathematik deutliche Länderunterschiede im mittleren Kompetenzniveau in der gymnasialen Oberstufe festgestellt werden (Nagy et al., 2007). Solche Unterschiede dürften auch heute noch bestehen. So zeigen die Ergebnisse der vom Institut zur Qualitätsentwicklung im Bildungswesen (IQB) in regelmäßigen Abständen durchgeführten Bildungstrendstudien, dass sich die am Ende der Sekundarstufe I in den Fächern Deutsch, Englisch und Mathematik sowie in den naturwissenschaftlichen Fächern im Mittel erzielten Kompetenzen zum Teil erheblich zwischen den Ländern unterscheiden (Böhme & Hoffmann, 2016; Holtmann et al., 2019; Mahler & Kölm, 2019; Schipolowski & Sachse, 2016). Dies ist auch dann der Fall, wenn nur die Schülerinnen und Schüler betrachtet werden, die ein Gymnasium besuchen bzw. die unabhängig von der besuchten Schulart die Allgemeine Hochschulreife anstreben. Diese im Rahmen von Schulleistungsstudien ermittelten Kompetenzunterschiede zwischen den Schülerinnen und Schü-

lern verschiedener Länder, die bis zum Erreichen der Allgemeinen Hochschulreife kaum vollständig auszugleichen sein dürften, spiegeln sich nicht in den Abiturnoten wider. So zeigt ein Blick in die von der KMK jährlich veröffentlichten Abiturnotenstatistiken, dass sich die erzielten Notenmittel kaum zwischen den Ländern unterscheiden: In den Jahren vor der Corona-Pandemie lag die mittlere Abiturdurchschnittsnote in den meisten Ländern zwischen 2,3 und 2,5.[1] Ergänzend ist in diesem Zusammenhang zu erwähnen, dass zwischen den Ländern erhebliche Unterschiede im Anteil der Schulabsolventinnen und -absolventen mit allgemeiner Hochschulreife bestehen. Im Abgangsjahr 2019 variierte dieser Anteil etwa zwischen 32,1 Prozent (in Bayern und Sachsen-Anhalt) und 54,5 Prozent (in Hamburg) (Statistisches Bundesamt, 2020). Dabei ist davon auszugehen, dass Länder mit sehr geringen Abiturientenquoten durch eine höhere leistungsbezogene Selektivität der Schülerschaft gekennzeichnet sind, was sich zumindest in der Tendenz in einem höheren mittleren Leistungsniveau der Abiturientinnen und Abiturienten niederschlagen sollte. In Ländern mit sehr hohen Absolventenquoten und geringerer leistungsbezogener Selektivität der Schülerschaft dürfte das mittlere Leistungsniveau der Abiturientinnen und Abiturienten hingegen tendenziell niedriger ausfallen. Doch, wie bereits erwähnt, spiegeln sich dahingehende Länderunterschiede nicht in den mittleren Abiturdurchschnittsnoten der Länder wider.

In der öffentlichen Debatte wird die scheinbar geringe Vergleichbarkeit von Abiturnoten nicht selten als Gerechtigkeitsproblem wahrgenommen (Blossfeld et al., 2011; Stanat et al., 2016). Auf der einen Seite wird argumentiert, dass Schülerinnen und Schüler aus leistungsstarken Ländern, in denen es vermeintlich schwieriger ist, gute Noten zu erzielen, bei der Studienplatzvergabe benachteiligt seien. Entsprechend wiesen auch die Richter des Bundesverfassungsgerichtes in Karlsruhe im Jahr 2017 im Rahmen ihrer Begründung des viel beachteten Urteils zur verfassungsrechtlichen Prüfung der Studienplatzvergabe im Fach Medizin darauf hin, dass Abiturnoten nur eingeschränkt länderübergreifend vergleichbar seien (BVerfG, 2017). Auf der anderen Seite befürchten Abiturientinnen und Abiturienten aus vermeintlich leistungsschwächeren Ländern, dass ihr Abschluss als weniger wertvoll wahrgenommen und möglicherweise von einigen anderen Ländern nicht anerkannt werden könnte.[2,]

Es liegt auf der Hand, dass die skizzierte Debatte um die mangelnde Vergleichbarkeit von Abiturnoten auch die Bildungspolitik der letzten drei Dekaden beschäftigte und den eingangs erwähnten Standardisierungs- und Annäherungsprozess der Länder bei den Regelungen zu den Abiturprüfungen maßgeblich motivierte. Im Folgenden werden vier wesentliche Etappen dieses Prozesses beschrieben. Auf die ersten drei Etappen – die Neuordnung der gymnasialen Oberstufe, die Einführung des Zentralabiturs sowie die Überarbeitung der EPA und die Entwicklung der Bildungsstandards für die Allgemeine Hochschulreife soll dabei nur knapp eingegangen werden, da diese bereits in anderen Publikationen dargestellt wurden (vgl. z. B. Kühn, 2012;

1 https://www.kmk.org/dokumentation-statistik/statistik/schulstatistik/abiturnoten.html [15.07.2021].
2 https://www.spiegel.de/spiegel/beim-abitur-entscheidet-herkunft-statt-leistung-a-1145711.html [19.01.2022]; https://www.welt.de/politik/deutschland/article129189233/Die-gefaehrliche-Entwertung-des-deutschen-Abiturs.html [19.01.2022].

Neumann, 2010; Stanat et al., 2016). Etwas genauer wird hingegen die vierte Etappe – die Entwicklung von Gemeinsamen Abituraufgabenpools der Länder – in den Blick genommen.

2 Die Neuordnung der gymnasialen Oberstufe

Eine zentrale Zäsur in der Geschichte der gymnasialen Oberstufe stellt zweifelsohne die bereits im ersten Beitrag des vorliegenden Bandes erwähnte Reform des Jahres 1972 dar (Bölling, 2010; Neumann, 2010). Im Ergebnis dieser Reform trat an die Stelle des Unterrichts in festen Klassenverbänden ein auch heute noch in vielen Ländern vorzufindendes Kurssystem, das seinerzeit umfangreiche Wahlmöglichkeiten zur individuellen Profilierung bot. Grundlegend war dabei die Einführung zweier Anforderungsniveaus: Im Rahmen eines komplexen Regelsystems, das zwischen einem Pflicht- und einem Wahlbereich unterschied, waren die Schülerinnen und Schüler nun aufgefordert, zwei Fächer auszuwählen, die sie als fünf- oder sechsstündige Leistungskurse belegen wollten. Der Unterricht in diesen Leistungskursen fand dabei auf erhöhtem Niveau statt und sollte „ein vertieftes wissenschaftspropädeutisches Verständnis und erweiterte Spezialkenntnisse“ (KMK, 1972, S. 14) vermitteln. Darüber hinaus mussten die Schülerinnen und Schüler weitere Fächer wählen, die sie in Form von zwei- oder dreistündigen Grundkursen belegten. Diese Fächer wurden auf grundlegendem Niveau unterrichtet, das primär auf die Sicherung einer vertieften Allgemeinbildung abzielte. Am Ende der gymnasialen Oberstufe wurden die Schülerinnen und Schüler in jeweils vier Fächern geprüft. In den beiden Leistungskursfächern und in einem Grundkursfach erfolgte die Abiturprüfung schriftlich, in einem weiteren Grundkursfach mündlich. Die Bewertung der in der Prüfung und in der Qualifikationsphase erzielten Leistungen erfolgte anhand des ebenfalls im Zuge der Reform eingeführten und auch heute noch verwendeten Punktesystems (Boggasch, 2011; KMK, 1972; Neumann, 2010).

Die reformierte Oberstufe und ihre länderspezifischen Umsetzungen wurden immer wieder stark kritisiert (z. B. KMK-Expertenkommission, 1995; Neumann, 2010; Neumann et al., 2012; Tenorth, 1996). Weite Teile der Kritik bezogen sich dabei auf die bereits eingangs erläuterten Probleme der mangelnden Vergleichbarkeit von Abiturnoten. Beispielsweise wurde problematisiert, dass sich die auf der Oberstufenvereinbarung basierenden Oberstufenverordnungen der Länder im Laufe der Jahre zunehmend auseinanderentwickelt hätten und dadurch verschiedene Systeme entstanden seien, die sich etwa im Hinblick auf die Regelungen zu Belegpflichten für Kernfächer wie Deutsch, Mathematik oder die erste Fremdsprache erheblich unterschieden. Zudem wurde argumentiert, dass die Wahlmöglichkeiten zu weit gingen, die Fächerzusammenstellung beliebig sei und eine zu starke Spezialisierung (in den jeweils gewählten Leistungskursfächern) zulasten der Vermittlung und Anwendung grundlegender Konzepte (in den jeweils anderen Fächern) stattfand. Ähnlich wie heute wurde auch ein allgemeiner Niveauverfall beim Abitur beklagt, der zu mangelnder Studierfähigkeit und geringer Allgemeinbildung der Abiturientinnen und Abiturienten führe. Zu-

dem wurde verschiedentlich die Vermutung geäußert, dass die Wahl von Grund- und Leistungskursen weniger interessengeleitet sei als vielmehr taktischem Kalkül folge, also primär darauf ziele, eine möglichst gute Abiturnote zu erreichen.

Obwohl einige dieser Kritikpunkte nicht empirisch fundiert waren (Baumert & Köller, 2000; KMK-Expertenkommission, 1995; Neumann, 2010; Neumann et al., 2012; Roeder & Gruehn, 1996; Schnabel & Gruehn, 2000), veranlassten sie die KMK Mitte der 1990er-Jahre, eine aus Bildungsforschern, Schulleiterinnen, Elternvertretern, Wirtschaftsexperten und Hochschullehrenden bestehende Expertenkommission einzusetzen, um eine „Weiterentwicklung der Prinzipien der Gymnasialen Oberstufe und des Abiturs" prüfen zu lassen (Tenorth, 1996). Diese Kommission erarbeitete Empfehlungen, die 1995 in den „Mainzer-Beschlüssen" (KMK, 1995 – 274. Plenarsitzung der KMK) aufgegriffen wurden. Im Rahmen einer von der KMK beschlossenen „Richtungsentscheidung zur Weiterentwicklung der Prinzipien der gymnasialen Oberstufe und des Abiturs" votierten die Länder für eine generelle Belegpflicht in den Kernfächern Deutsch, Mathematik und einer Fremdsprache (KMK, 1995; Neumann, 2010). Diese Richtungsentscheidung wurde in die revidierte KMK-Vereinbarung für die gymnasiale Oberstufe des Jahres 1997 aufgenommen und dort weiter präzisiert.

Eine noch viel weitreichendere Revision der reformierten Oberstufe wurde bereits zwei Jahre später im Rahmen der 287. Plenarsitzung der Kultusministerkonferenz in Husum vorgenommen. Auf Initiative von Baden-Württemberg und Bayern (Neumann, 2010) hatten sich die Länder darauf geeinigt, größere Spielräume bei der Ausgestaltung der Oberstufenmodelle zu ermöglichen. Insbesondere war es nun gestattet, mehr als drei Leistungskurse verbindlich zu belegen. Zudem durfte der Stundenumfang von Leistungskursen statt wie bislang mindestens fünf Wochenstunden nun auch nur vier Wochenstunden umfassen. Darüber hinaus gaben die Husumer Beschlüsse den Ländern die Möglichkeit, die Anzahl der Abiturprüfungen, die von den Schülerinnen und Schülern am Ende der gymnasialen Oberstufe jeweils zu absolvieren sind, von vier auf fünf Fächer zu erhöhen (KMK, 1999).

Die Husumer Beschlüsse ermöglichten es den Ländern, die bisherigen Kurssysteme zu stärker kanonförmigen Kernfach- bzw. Profilsystemen umzugestalten (Kampa et al., 2016; Neumann, 2010; Trautwein et al., 2010). Der Rückbau des Kurssystems und die Abkehr vom Prinzip der frühen Spezialisierung hatten dabei insbesondere zum Ziel, die allgemeine Studierfähigkeit und die vertiefte Allgemeinbildung der Abiturientinnen und Abiturienten zu sichern. Innerhalb der einzelnen Länder sollte zudem mit den Kernfächern eine Angleichung des Leistungsniveaus in zentralen Domänen sowie eine bessere Vergleichbarkeit von Abiturzeugnissen herbeigeführt werden (Neumann, 2010).

Empirische Befunde zu den Wirkungen der Reform liegen insbesondere für Baden-Württemberg vor, wo die Neuordnung der gymnasialen Oberstufe bereits zum Schuljahr 2001/2002 erfolgte. Die Ergebnisse der Studie TOSCA-Repeat zeigten unter anderem, dass die Reform insgesamt zu einer Erhöhung der Unterrichtszeit in den beiden untersuchten Fächern Mathematik und Englisch geführt hatte (Neumann et al., 2012). Ein Vergleich der Kohorten aus den Jahren 2002 und 2006 ließ ferner darauf schließen, dass im Fach Mathematik eine moderate Verbesserung der mathe-

matischen Kompetenzen sowie eine Verringerung der Leistungsstreuung bewirkt wurde (Nagy et al., 2010). Für das Fach Englisch hingegen wurden im Kohortenvergleich keine bedeutsamen Leistungsunterschiede gefunden (Jonkmann et al., 2010).

In den Folgejahren machte ein Großteil der Länder von den Möglichkeiten der Husumer Beschlüsse Gebrauch und ordnete ihre gymnasialen Oberstufen neu (Trautwein et al., 2010), wobei sich die dabei resultierenden Systeme zum Teil deutlich voneinander unterscheiden. Auch wurde in einigen Ländern bereits zum Kurssystem zurückgekehrt bzw. wird eine Rückkehr aktuell erwogen. In Summe haben die infolge der Husumer Beschlüsse durchgeführten Reformen also zu einer weiteren Diversifizierung der Oberstufensysteme der Länder geführt. Mit den stark reduzierten Wahlmöglichkeiten erhöhte sich zwar die Vergleichbarkeit von Abiturzeugnissen innerhalb der jeweiligen Länder, die Vergleichbarkeit von Abiturnoten zwischen den Ländern wurde hingegen infolge der Diversifizierung der Oberstufensysteme weiter reduziert.

3 Die Einführung des Zentralabiturs

Zeitlich parallel zur skizzierten Neuordnung der gymnasialen Oberstufe wurde mit der Einführung des Zentralabiturs in vielen Ländern eine weitere Reform durchgeführt, welche direkt die schriftlichen Abiturprüfungen und die dabei gestellten Aufgaben betraf. In den westdeutschen Ländern waren zentral organisierte Abiturprüfungen zuvor nur in drei Ländern implementiert: in Bayern (seit 1946) und Baden-Württemberg (seit 1946 bzw. 1952), wo sich das Zentralabitur auf eine bis ins 19. Jahrhundert zurückgehende Tradition stützt, sowie im Saarland (seit 1945), dessen Prüfungssystem nach dem Zweiten Weltkrieg maßgeblich vom zentralistischen Steuerungsprinzip der französischen Besatzungsmacht geprägt wurde (Bölling, 2010; van Ackeren, 2007). Die ostdeutschen Länder hatten (mit Ausnahme von Brandenburg) nach dem Beitritt zur Bundesrepublik am Prinzip des Zentralabiturs festgehalten, das in der DDR seit 1959 existierte. In allen übrigen Ländern wurde das Zentralabitur zwischen den Jahren 2005 und 2008 eingeführt (Bölling, 2010; Klein et al., 2009). Eine Ausnahme hiervon stellt Rheinland-Pfalz dar, wo bis heute ein Großteil der Aufgaben der schriftlichen Abiturprüfungen schulintern entwickelt und eingesetzt wird.

Die Umstellung von einem dezentralen zu einem zentralen Prüfungssystem erfolgte nicht zuletzt in der Annahme, hierdurch die Qualitätsstandards für die Abiturprüfungen, für die Bewertung von Abiturarbeiten und für den Unterricht in der gymnasialen Oberstufe besser sichern zu können (Klein et al., 2009). Das Zentralabitur sollte einerseits die Vergleichbarkeit von Abiturprüfungen und Abiturnoten erhöhen und andererseits ein hohes Leistungsniveau der Abiturientinnen und Abiturienten sichern (Wößmann, 2003, 2008). Von zentral gestellten Abiturprüfungsaufgaben erhoffte man sich zudem positive Rückwirkungen auf das Lehren und Lernen in der gesamten gymnasialen Oberstufe. Solche Rückwirkungen werden in der Forschungsliteratur zumeist als „Washback-“ bzw. „Backwash-Effekte“ bezeichnet (Cheng et al., 2004) und sollten beim Zentralabitur insbesondere die Aufgabenkultur an den Schulen, vor allem auch die Gestaltung der in den Halbjahren der Qualifikationsphase zu

schreibenden Klausuren, betreffen (Kühn, 2011; Maué & Maag Merki, 2019). Den Lehrkräften sollten die zentral gestellten Abiturprüfungsaufgaben Orientierung für die Entwicklung eigener Klausuraufgaben geben, indem sie exemplarisch zeigen, wie geprüft werden kann, inwieweit Schülerinnen und Schüler die von ihnen in der gymnasialen Oberstufe geforderten Anforderungen bewältigen. Die zentral gestellten Abiturprüfungsaufgaben dienten mithin auch als ein Vehikel, um zentrale Aspekte der Qualität und didaktischen Gestaltung von Prüfungsaufgaben in den Schulen zu implementieren. Sie sollten eine normierende Wirkung auf die Qualifikationsphase haben und somit auch zu einer besseren Vergleichbarkeit der Halbjahresnoten führen, die zusätzlich zu den Prüfungsergebnissen in die Abiturnoten einfließen.

Demgegenüber warnten Kritikerinnen und Kritiker der Umstellung auf ein zentrales Prüfungssystem vor nicht intendierten Washback-Effekten, wie etwa eine Engführung des Unterrichts auf den Prüfungsstoff bzw. eine stark eingeschränkte Themenvarianz („Teaching-to-the-Test-Effekt") (Jäger, 2012; Oerke et al., 2013). Als weitere Nebenwirkung wurde ein Niveauabfall in der Aufgabenkultur durch eine Dominanz von Aufgabenformaten befürchtet, die vor allem reproduktives Lernen voraussetzen, gegenüber Aufgaben, die verständnisorientiertes Lernen und komplexes Problemlösen erfordern (van Ackeren, 2007).

Im Rahmen der Diskussion um die Einführung zentraler Prüfungen wurde zudem kritisch hinterfragt, wie zentral die Zentralabiture der einzelnen Länder tatsächlich seien. So attestierten die Befunde einer internationalen Vergleichsstudie von Klein et al. (2009), in der die zentralen Prüfungssysteme verschiedener Staaten und Länder gegenübergestellt wurden, den deutschen Zentralabituren einen nur geringen bis mittleren Standardisierungsgrad. In der Tat zielte die Einführung des Zentralabiturs hierzulande vor allem auf eine Veränderung des Gremiums, das die Prüfungsaufgaben entwickelt bzw. auswählt. An die Stelle schulinterner Prüfungen, bei der die Prüfungsaufgaben von den Fachlehrerinnen und Fachlehrern der einzelnen Schulen erstellt wurden, rückten schulexterne Prüfungen, bei denen die Entwicklung der Prüfungsaufgaben vom jeweiligen Kultusministerium verantwortet wird. Ein internationaler Vergleich zeigte demgegenüber, dass in einigen anderen Staaten weitere Elemente der Prüfungsorganisation, wie etwa die Prüfungsvorbereitung oder die Korrektur der Prüfungsarbeiten, in hohem Maße standardisiert sind (Klein et al., 2009; Kühn, 2012). Vergleichende Analysen der Prüfungssysteme einzelner Bundesländer verdeutlichten darüber hinaus, dass sich die Zentralabiture zwar oberflächlich ähnelten, im Detail jedoch große Unterschiede bestanden bzw. bis heute bestehen (Klein et al., 2009; Kühn, 2012). Ein aktueller Überblick über diese Heterogenität findet sich in dem von Svenja Mareike Schmid-Kühn und Alexander Groß verfassten dritten Beitrag des vorliegenden Bandes.

In Bremen und Hessen wurde die Einführung des Zentralabiturs von empirischen Evaluationen flankiert, die untersuchten, welche Effekte und ggf. Nebenwirkungen mit der Umstellung des Prüfungsverfahrens verbunden waren (Maag Merki, 2012b). Positive Effekte auf das Leistungsniveau der Schülerinnen und Schüler in den Fächern Englisch und Mathematik zeigten sich dabei nur vereinzelt (Maag Merki, 2012a, 2016), hingegen war in beiden Ländern ein von Kritikerinnen und Kritikern

befürchteter Teaching-to-the-Test-Effekt zu verzeichnen (Jäger, 2012; Maag Merki, 2016). Gleichzeitig konnten für beide Länder Hinweise auf eine größere Standardisierung von Bewertungsmaßstäben im Sinne einer stärkeren Orientierung an den für die Prüfungsaufgaben vorgegebenen Bewertungskriterien („kriteriale Bezugsnormorientierung") nachgewiesen werden (Holmeier, 2012; Maag Merki, 2016).

Insgesamt zeichneten die Ergebnisse in den Ländern Bremen und Hessen also ein gemischtes Bild der Wirkungen einer Umstellung von dezentralen auf zentrale Prüfungssysteme. Darüber hinaus ist anzunehmen, dass die Einführung landesweiter Zentralabiture, aufgrund der erwähnten Heterogenität zwischen den Systemen der einzelnen Länder, die bundesweite Vergleichbarkeit von Abiturnoten kaum erhöht haben dürfte. Nicht zuletzt deshalb sind in der öffentlichen Debatte zur Vergleichbarkeit von Abiturnoten Forderungen nach einem bundesweiten Zentralabitur bis heute überaus populär. So sprachen sich etwa im Jahr 2019 in einer vom Meinungsforschungsinstitut YouGov im Auftrag der Deutschen Presse-Agentur durchgeführten repräsentativen Umfrage rund 80 Prozent der Befragten für bundesweit einheitliche Abiturprüfungen aus.[3] Vonseiten der Politik wurde dieses Thema insbesondere im Jahr 2007 diskutiert, als der damalige Baden-Württembergische Ministerpräsident Günther Oettinger (CDU) einen entsprechenden Vorschlag unterbreitete, der auch von der damaligen Bundesbildungsministerin Annette Schavan (CDU) unterstützt wurde. Dieser Vorschlag konnte sich zwar nicht durchsetzen, motivierte aber einige Länder zur Bildung der sogenannten Gruppe „länderübergreifendes Abitur" (LüA, s. u.) – mit dem Ziel, gemeinsam länderübergreifende Abiturprüfungsaufgaben zu entwickeln.

Auf bildungspolitischer Ebene wird die Einführung eines bundesweiten Zentralabiturs gegenwärtig nicht mehr ernsthaft diskutiert (Stanat et al., 2016). Hier setzt man auf die in Abschnitt 5 des vorliegenden Beitrags dargestellten Gemeinsamen Abituraufgabenpools der Länder.

4 Die Überarbeitung der einheitlichen Prüfungsanforderungen für die Abiturprüfung und die Entwicklung von Bildungsstandards für die allgemeine Hochschulreife

Das enttäuschende Abschneiden von Schülerinnen und Schülern in Deutschland bei internationalen Schulleistungsstudien wie PISA (Baumert et al., 2001; Baumert et al., 2002) oder TIMSS-III (Baumert et al., 2000) beförderte zu Beginn des neuen Jahrtausends ein Umdenken der Bildungspolitik und -verwaltung in Bezug auf die Art und Weise, wie Bildungssysteme zu steuern sind. War man bis dato davon überzeugt, Bildungssysteme am effizientesten über den „Input" steuern zu können, also durch

3 https://de.statista.com/statistik/daten/studie/183258/umfrage/meinung-zur-einfuehrung-eines-bundesweiten-zentralabiturs/ [03.12.2021].

Vorgabe von detaillierten Richtlinien und Regelungen sowie Zuteilung von Ressourcen (z. B. in Form von Lehrplänen, Ausbildungsbestimmungen für Lehrkräfte oder Prüfungsrichtlinien), erfolgte nun ein Paradigmenwechsel hin zu einer stärkeren Outputorientierung. Bei der Steuerung des allgemeinen Schulsystems wurden nun neben den Aspekten, welche die bislang dominierende Inputsteuerung kennzeichnen, nun auch die in den verschiedenen Ebenen des Schulsystems erzielten Bildungserträge verstärkt in den Blick genommen (Grünkorn et al., 2019; Klieme et al., 2007). In diesem Sinne verabschiedete die KMK in den Jahren 2003 und 2004 für den Primarbereich, für den Hauptschulabschluss (HSA) und den Mittleren Schulabschluss (MSA) in zentralen Fächern Bildungsstandards, die seither für alle 16 Länder verbindliche Zielkriterien für zu erreichende Kompetenzen definieren (z. B. KMK, 2004, 2005a, 2005b).

Im Bereich der gymnasialen Oberstufe wurden zwischen den Jahren 2002 und 2004 zunächst die Einheitlichen Prüfungsanforderungen in der Abiturprüfung (EPA) überarbeitet. Wie bereits erwähnt, waren die zunächst auch unter dem Namen „Normbücher" bekannten EPA bereits im Jahr 1979 von der KMK eingeführt worden (Bölling, 2010). Anlass hierfür war ein Urteil des Bundesverfassungsgerichts aus dem Jahr 1972, in dessen Rahmen die Richter eine Vergabe von Studienplätzen nach dem Notendurchschnitt nur dann als zulässig erachteten, wenn die Vergleichbarkeit von Abiturnoten erhöht werde. Vor diesem Hintergrund definierten die Normbücher für zunächst 14 Fächer Anforderungen an die Prüfung sowie an die Konstruktion und Bewertung von Abiturprüfungsaufgaben, an denen sich alle Länder orientieren sollten. Seither sind auch für viele weitere Fächer EPA entwickelt worden (Bölling, 2010). Die im Jahr 2002 von der KMK beschlossene Überarbeitung der EPA zielte auf eine kompetenzorientierte Revision der bisherigen Vorgaben ab. Hierbei wurden für jedes Fach Kompetenzbereiche benannt und mit Angaben zu den jeweils zu beherrschenden fachlichen Anforderungen und Inhalten spezifiziert (Köller, 2007). Die Revision der EPA sollte mithin insbesondere dazu dienen, die in den Abiturprüfungen der einzelnen Länder zu bewältigenden Leistungsanforderungen anzugleichen und somit die länderübergreifende Vergleichbarkeit von Abiturnoten zu erhöhen.

Im Jahr 2007 beschloss die KMK in ihrer 319. Sitzung schließlich, die revidierten EPA für die Fächer Deutsch und Mathematik sowie für die erste Fremdsprache (Englisch bzw. Französisch) zu Bildungsstandards für die Allgemeine Hochschulreife (AHR) weiterzuentwickeln und das IQB mit der Koordination dieses Prozesses zu betrauen (Stanat et al., 2016). Die Bildungsstandards, deren Entwicklung im Jahr 2009 begann, wurden 2012 im Rahmen der 339. Sitzung der KMK verabschiedet. Gleichzeitig hatten die Länder vereinbart, ihre Abiturprüfungen spätestens ab dem Jahr 2017 auf der Grundlage der Standards durchzuführen. Ebenfalls im Jahr 2017 begann die Entwicklung von Bildungsstandards für die AHR in den Fächern Biologie, Chemie und Physik, die dann im Jahr 2020 in der 370. Sitzung der KMK beschlossen wurden. Dabei wurde vereinbart, dass die Abiturprüfungen der Länder in den naturwissenschaftlichen Fächern ab dem Schuljahr 2024/25 auf den verabschiedeten Standards basieren sollten.

Das Kernstück der auf den 339. und 370. Sitzungen der KMK veröffentlichten Dokumente bilden die Bildungsstandards, die bundesweit gültige Zielvorgaben dazu formulieren, über welche Kompetenzen Schülerinnen und Schüler in der Regel verfügen sollen, wenn sie die Schule mit der AHR abschließen. Diese Zielvorgaben knüpfen zum einen an die Bildungsstandards für den MSA und den ihnen zugrunde liegenden Kompetenzstrukturmodelle an und greifen zum anderen die in den EPA benannten fachlichen Anforderungen auf. Analog zu den Bildungsstandards für den Primarbereich, den HSA und den MSA werden die Bildungsstandards für die AHR mit Beispielaufgaben illustriert. Für die Fächer Deutsch und Mathematik sowie für die erste Fremdsprache wurden diese gemeinsam mit den Bildungsstandards veröffentlicht. In Zusammenarbeit des IQB mit fachdidaktischen Kooperationspartnern des Instituts wurden zudem Publikationen mit umfangreichen fachdidaktischen Erläuterungen sowie mit Hinweisen und Beispielen zur Implementation der Standards im Unterricht der gymnasialen Oberstufe erstellt (Becker-Mrotzek et al., 2015; Blum et al., 2015; Tesch et al., 2017). Zu den Bildungsstandards für die AHR in den naturwissenschaftlichen Fächern finden sich illustrierende Aufgaben auf den Internetseiten des IQB.[4]

Gegenüber den Bildungsstandards für den Primarbereich, den HSA und den MSA weisen die Dokumente mit den Bildungsstandards für die AHR die Besonderheit auf, dass sie (ähnlich wie die von ihnen abgelösten EPA) „Hinweise für die Prüfungsdurchführung" beinhalten. Diese Hinweise spezifizieren „insbesondere [...], welche Arten von Aufgaben in der Abiturprüfung gestellt werden können, in welcher Weise die erwarteten Schülerleistungen zu beschreiben und nach welchen Kriterien die erbrachten Abiturprüfungsleistungen zu bewerten sind" (KMK, 2012, S. 22). Mit der Verabschiedung der Bildungsstandards für die AHR wurden also nicht nur bundesweit geltende Vorgaben von Zielkriterien für die AHR formuliert. Vielmehr erfolgte auch eine weitere Angleichung der Prüfungsdurchführung. Beides schuf eine allgemeine Grundlage für die Entwicklung eines gemeinsamen Abituraufgabenpools der Länder, aus denen Prüfungsaufgaben entnommen und in den Abiturprüfungen mehrerer Länder parallel eingesetzt werden können.

Nach Verabschiedung der Bildungsstandards für die AHR waren die Länder bis zum Schuljahr 2014/15 aufgefordert, ihre curricularen Vorgaben bzw. Lehrpläne für die gymnasiale Oberstufe an die in den Bildungsstandards formulierten Zielkriterien anzupassen (KMK, 2015a). Hierbei hatten die Länder zum Teil recht große Gestaltungsspielräume. Diese bestanden in den Fächern Deutsch, Englisch und Französisch zum Beispiel darin, dass die Bildungsstandards in diesen drei Fächern recht allgemeine Angaben zu den jeweils zu beherrschenden Inhalten im Bereich der Literatur machen. Dieser inhaltliche Gestaltungsspielraum wird bis heute recht unterschiedlich genutzt. So geben die Lehrpläne einiger Länder einen fixen Literaturkanon in Form von Lektürelisten für das Fach Deutsch und die modernen Fremdsprachen vor, der definiert, welche literarischen Werke im Unterricht zu behandeln sind. Andere

4 Diese Aufgaben können unter https://www.iqb.hu-berlin.de/abitur/sammlung/naturwissenschaften/ [19.01.2021] eingesehen werden.

Länder formulieren stattdessen sogenannte Themenkorridore, die thematische Schwerpunkte setzen, deren Ausgestaltung (zum Beispiel mit Blick auf die Unterrichtslektüre) aber weitestgehend der jeweiligen Lehrkraft obliegt. Im Fach Mathematik sind die inhaltlichen Vorgaben der Bildungsstandards für die AHR etwas konkreter als in den sprachlichen Fächern, dennoch bestehen auch hier Spielräume bei der Lehrplangestaltung. Im Fach Mathematik betrifft ein weiterer wichtiger Aspekt die Verwendung digitaler Hilfsmittel. Da die Bildungsstandards hierzu nur allgemeine Angaben machen, kommen in den Ländern, und nicht selten sogar innerhalb einzelner Länder, unterschiedliche Arten von Taschenrechnern (d. h. wissenschaftliche Taschenrechner, grafikfähige Taschenrechner oder Computeralgebrasysteme) zum Einsatz.

5 Die Entwicklung von gemeinsamen Abituraufgabenpools der Länder

Wie eingangs erwähnt, liegt der Fokus des fünften und umfangreichsten Abschnitts des vorliegenden Beitrags auf dem aktuellsten Vorhaben der Bundesländer zur Sicherung der Vergleichbarkeit von Prüfungsanforderungen – den Gemeinsamen Abituraufgabenpools der Länder. Nach einem einleitenden Abriss der Entstehungsgeschichte und den konkreten Zielen des Vorhabens wird ausführlicher beschrieben, wie die Aufgaben für die Abituraufgabenpools entwickelt werden. Anschließend wird dargestellt, wie die Länder Aufgaben aus den Pools entnehmen und in ihren schriftlichen Abiturprüfungen einsetzen können. Im Anschluss wird skizziert, wie die Abiturprüfungspools eine Annäherung der Prüfungsvorgaben der Länder befördern sollen. Der Abschnitt schließt mit einem kurzen Ausblick auf zukünftige Vorhaben.

5.1 Entwicklungsgeschichte und Ziele

Bereits in ihrer 319. Sitzung im Jahr 2007, bei der auch die Weiterentwicklung der EPA zu bundesweit einheitlichen Bildungsstandards für die AHR beschlossen worden war, hatte die KMK die Erstellung eines Pools von Aufgaben für die gymnasiale Oberstufe angekündigt. In einem parallelen Prozess fanden sich kurze Zeit später, im Jahr 2008, einige Länder der sogenannten B-Seite (Länder, in denen die CDU/CSU das Kultusministerium innehat) mit dem Ziel zusammen, unter dem Namen „Süd-Abitur" gemeinsame Abiturprüfungsaufgaben bzw. -aufgabenteile zu entwickeln. Nach einigen Wechseln in der Zusammensetzung der Gruppe, bei denen einige Länder diese aufgrund von wechselnden Mehrheiten in den jeweiligen Landesparlamenten verließen und andere Länder der A-Seite (d. h. Länder, in denen die SPD das Kultusministerium innehat) zur Gruppe hinzukamen, beteiligen sich seit etwa 2011 konstant die Länder Bayern, Hamburg, Mecklenburg-Vorpommern, Niedersachsen, Sachsen und Schleswig-Holstein an dem fortan als „länderübergreifendes Abitur" (LüA) bezeichneten Projekt (Wolf & Heinz, 2016). Seit dem Prüfungsjahr 2014 verwenden diese Länder die im Projekt LüA entwickelten Aufgaben bzw. Aufgabenteile in den schriftlichen

Abiturprüfungen der Fächer Deutsch, Mathematik und Englisch (StMUK, 2012; Wößmann, 2012). Dies umfasst in jedem Prüfungsjahr im Fach Deutsch eine gemeinsame Aufgabe, im Fach Englisch Aufgaben für die Kompetenzbereiche „Hörverstehen" und „Sprachmittlung", die jeweils von einem Teil der LüA-Länder eingesetzt werden, und im Fach Mathematik Aufgaben für den Prüfungsteil A, der ohne Hilfsmittel (d. h. ohne wissenschaftlichen Taschenrechner bzw. Computeralgebrasystem) zu bearbeiten ist (StMUK, 2012).

Im Jahr 2012 beauftragte die KMK in ihrer 337. Sitzung das IQB mit dem Aufbau der Gemeinsamen Abituraufgabenpools der Länder für die Fächer Deutsch, Mathematik, Englisch und Französisch. Als deren Hauptziele benannte die KMK die Sicherung der Vergleichbarkeit, Standardbasierung und Qualität von Abiturprüfungsaufgaben (KMK, 2015b). Ähnlich wie bei der in Abschnitt 3 erläuterten Einführung des Zentralabiturs verband man mit den Abituraufgabenpools nicht zuletzt die Annahme, dass dies positive Washback-Effekte zur Folge haben würde. Die mit den Pools zur Verfügung gestellten Aufgaben sollten eine normierende Wirkung entfalten, indem sie sowohl den Abituraufgabenkommissionen der Länder als auch den Lehrkräften an den Schulen als Orientierung bei der Entwicklung eigener Prüfungs- oder Klausuraufgaben dienen (KMK, 2017). Die konzeptionellen Grundlagen der Abituraufgabenpools, die vom IQB in enger Zusammenarbeit mit der damaligen Vorsitzenden des Schulausschusses sowie in Abstimmung mit dem Kuratorium des IQB und der Amtschefskommission „Qualitätssicherung in Schulen" entwickelt worden waren (Stanat & Pant, 2013), wurde im Jahr 2013 im Rahmen der 342. Sitzung der KMK beschlossen. Diese Grundlagen waren vom IQB in enger Zusammenarbeit mit der damaligen Vorsitzenden des Schulausschusses sowie in Abstimmung mit dem Kuratorium des IQB und der Amtschefskommission „Qualitätssicherung in Schulen" erarbeitet worden (ebd.). Berücksichtigung fanden dabei auch die zuvor von der LüA-Gruppe gewonnenen Erfahrungen.

Seit dem Prüfungsjahr 2017 können die Länder Prüfungsaufgaben aus den Abituraufgabenpools entnehmen und in ihren Abiturprüfungen einsetzen (Disdorn-Liesen, 2016; KMK, 2015b; Stanat, 2014; Stanat & Pant, 2013). LüA ist mittlerweile eng mit den Abituraufgabenpools verflochten. So umfassen die Pools in den Fächern Deutsch, Mathematik und Englisch auch die Prüfungsaufgaben, die im Rahmen des LüA-Projekts erarbeitet wurden. Somit haben auch Länder, die nicht zur LüA-Gruppe gehören, die Möglichkeit, deren Aufgaben aus dem Pool zu entnehmen und in ihren Prüfungen einzusetzen.

5.2 Aufgabenentwicklung

Wie bereits erwähnt, werden gegenwärtig in den Fächern Deutsch, Mathematik, Englisch und Französisch Prüfungsaufgaben für die Gemeinsamen Abiturprüfungspools der Länder entwickelt. Dabei wurde in jedem der vier Fächer ein Gremium eingerichtet, das für die Erstellung der Prüfungsaufgaben zuständig ist. Diese als „AGs Aufgaben" bezeichneten Gremien sind so zusammengesetzt, dass sowohl die Perspektiven der einzelnen Länder und der beruflichen Gymnasien Berücksichtigung finden als

auch aktuelle Erkenntnisse der Fachdidaktiken in den Arbeitsprozess einfließen. Dementsprechend ist in den einzelnen AGs Aufgaben in der Regel pro Land jeweils eine Expertin bzw. ein Experte vertreten, die oftmals auch selbst an Schulen unterrichten und umfangreiche Erfahrungen in der Erstellung von Abiturprüfungsaufgaben aufweisen. Zudem umfasst der Mitgliederkreis der AGs Aufgaben Expertinnen und Experten aus dem Bereich der beruflichen Gymnasien. Für das IQB arbeiten in den Arbeitsgruppen jeweils eine Fachkoordinatorin bzw. ein Fachkoordinator sowie zwei Wissenschaftlerinnen und Wissenschaftler der betreffenden Fachdidaktiken mit (KMK, 2017; Stanat, 2014; Stanat & Pant, 2013). Den Fachdidaktikerinnen und Fachdidaktikern kommt dabei vor allem eine beratende Funktion zu. Sie selbst entwickeln keine Prüfungsaufgaben, geben aber aus der Perspektive der jeweiligen Fachdidaktik Rückmeldungen zu den in den Arbeitsgruppen erarbeiteten Aufgabenvorschlägen. Zudem beteiligen sie sich an der konzeptionellen Weiterentwicklung einzelner Aufgabenarten und von Materialien zur Bewertung von Prüfungsleistungen.

Nachdem sich die AGs Aufgaben im Januar des Jahres 2014 im Rahmen eines Auftakttreffens in Berlin konstituiert hatten (Stanat, 2014), erstellten diese zunächst eine Sammlung von Abiturprüfungsaufgaben, die zur 350. Sitzung der KMK im Jahr 2015 veröffentlicht wurde und auf den Internetseiten des IQB zum Download zur Verfügung steht. Diese Aufgabensammlung soll exemplarisch veranschaulichen, wie die in den Bildungsstandards beschriebenen Kompetenzen und Vorgaben für die Abiturprüfung in Aufgaben und Erwartungshorizonte übersetzt werden können. Mit den Aufgaben der Sammlung wird nicht das Ziel verfolgt, alle einzelnen Kompetenzen oder das gesamte Spektrum der Möglichkeiten zur Gestaltung von Abiturprüfungen abzubilden. Vielmehr soll damit Lehrkräften sowie Schülerinnen und Schülern Orientierung gegeben werden hinsichtlich der Gestaltung und der zu erwartenden Anforderungen der Aufgaben, die in den Gemeinsamen Abituraufgabenpools der Länder bereitgestellt werden (Stanat et al., 2016).

Parallel zur Erstellung der Aufgabensammlung begannen die AGs Aufgaben außerdem mit der Entwicklung von Prüfungsaufgaben für die gemeinsamen Abituraufgabenpools, die den Ländern, wie bereits erwähnt, seit dem Jahr 2017 in jedem Prüfungsjahr zur Nutzung zur Verfügung gestellt werden. Der Prozess der Aufgabenerstellung für die Pools eines bestimmten Prüfungsjahres umfasst dabei einen Zeitraum von insgesamt etwa drei Jahren. Die Aufgabenentwicklung beginnt jeweils im Mai eines Jahres mit der Erarbeitung von Aufgabenvorschlägen durch die Abiturkommissionen der Länder. Diese sind aufgefordert, auf der Grundlage der Bildungsstandards bis zum März des darauffolgenden Jahres Aufgabenvorschläge an das IQB zu übermitteln. Im Anschluss daran erfolgt eine Bewertung der eingereichten Aufgabenvorschläge durch die AGs Aufgaben, wobei nur die positiv eingeschätzten Vorschläge für den weiteren Prozess ausgewählt werden. Innerhalb eines Jahres werden diese Vorschläge dann von den Mitgliedern der AGs Aufgaben weiterentwickelt. Als Ergebnis der Arbeit der AGs Aufgaben wird den Ländern pro Fach ein Pool mit Abiturprüfungsaufgaben zur Verfügung gestellt, aus denen sich die Länder für die Prüfungen des darauffolgenden Jahres bedienen können. Im Fach Deutsch umfasst dieser Pool

etwa 25 Aufgaben, in den beiden Fremdsprachen ca. jeweils 20 Aufgaben. Im Fach Mathematik werden für den ohne Hilfsmittel zu bearbeitenden Prüfungsteil A etwa 35 Aufgaben und für den mit Hilfsmitteln zu bearbeitenden Prüfungsteil B ca. 30 Aufgaben bereitgestellt. Nach den Abiturprüfungen werden die von den Ländern eingesetzten Poolaufgaben mitsamt den zugehörigen Erwartungshorizonten und Bewertungshinweisen auf den Internetseiten des IQB veröffentlicht (KMK, 2017), sofern entsprechende Nutzungsrechte für die ggf. in den Aufgaben verwendeten urheberrechtlich geschützten Materialien erteilt wurden.

5.3 Aufgabenentnahme und -einsatz in den Prüfungen

In den Ländern erfolgt der Einsatz der Poolaufgaben bis heute in der Regel so, dass das Aufgabenset[5] einer Prüfung zum Teil aus Aufgaben besteht, die aus dem Pool entnommen wurden, und zum Teil aus Aufgaben, die „landeseigen" sind, also ausschließlich vom jeweiligen Land entwickelt wurden und nur dort zum Einsatz kommen. Allerdings gibt es auch einige wenige Länder, die zumindest in einzelnen Prüfungsfächern bereits ausschließlich Aufgaben der Pools genutzt haben.

In den meisten Ländern und Fächern sehen die Vorgaben für die schriftlichen Abiturprüfungen Wahlmöglichkeiten vor. Zumeist können die Prüflinge nach bestimmten Vorgaben auswählen, welche Aufgaben des Aufgabensets sie bearbeiten möchten. Bisweilen entscheiden auch die Lehrkräfte an den Schulen, welche Teile des Aufgabensets von den Prüflingen zu bearbeiten sind. Aufgrund der in vielen Ländern und Fächern bestehenden Wahlmöglichkeiten kann es also vorkommen, dass ein Prüfling ausschließlich landeseigene und keine aus dem Pool entnommenen Aufgaben bearbeitet. Sofern die Annahme zutrifft, dass die Abituraufgabenpools eine normierende Wirkung haben, sollten solche Fälle aber unproblematisch sein, da sich in diesem Fall die landeseigenen Aufgaben in ihren Anforderungen und ihrem Schwierigkeitsniveau an den Aufgaben der Pools orientieren. Zudem ist weder für die Lehrkräfte noch für die Prüflinge erkennbar, welche Aufgaben aus dem Pool stammen und bei welchen es sich um landeseigene Aufgaben handelt (KMK, 2017). Eine Ausnahme hiervon bildet die Abiturprüfung in Rheinland-Pfalz, das, wie oben bereits erwähnt, als einziges Land kein Zentralabitur eingeführt hat. Hier bestehen die Prüfungen zum einen aus Aufgaben, die von der jeweiligen Schule entwickelt wurden, und zum anderen aus einem zentralen Element, das aus dem Pool stammt und entsprechend identifiziert werden kann.

Über die Jahre hinweg haben sich in den einzelnen Ländern „unterhalb" der in den Bildungsstandards für die AHR bzw. zuvor in den EPA festgehaltenen Vorgaben zu Abiturprüfungsaufgaben zum Teil recht unterschiedliche Prüfungs- und Aufgabentraditionen entwickelt. Dabei lassen sich, wie in Beitrag 1 des vorliegenden Bandes aufgezeigt, mitunter Entwicklungslinien nachzeichnen, die bis ins 19. Jahrhundert zurückreichen. Die spezifischen Aufgabentraditionen der einzelnen Länder manifes-

5 In den Fächern Deutsch, Mathematik, Englisch und Französisch werden den Prüflingen in den schriftlichen Abiturprüfungen jeweils mehrere Aufgaben zur Bearbeitung vorgelegt, wobei je nach Fach und Land ggf. Wahlmöglichkeiten bestehen. Die Gesamtheit aller Aufgaben einer Prüfung wird an dieser Stelle als Aufgabenset bezeichnet.

tieren sich in unterschiedlichen Vorgaben und Regelungen für die Abiturprüfungen. Diese Unterschiede betreffen etwa die Wahlmöglichkeiten in der Prüfung, die zur Verfügung stehende Bearbeitungszeit, die Bevorzugung bestimmter Aufgabenformate und -arten, Vorgaben für die Binnenstruktur von Aufgabenstellungen, die Vorgabe von Lektürelisten oder Themenkorridoren bzw. den Verzicht auf diese, die Art der Bewertung der in den Prüfungen verfassten Schreibprodukte (z. B. holistischer vs. analytischer Ansatz), den Grad der Offenheit von Aufgabenstellungen, die Art der in der Prüfung erlaubten Hilfsmittel (z. B. ein- oder zweisprachige Wörterbücher in den Fremdsprachen oder die Art des zulässigen Taschenrechners im Fach Mathematik) sowie bestimmte inhaltliche Schwerpunktsetzungen, wie etwa im Fach Mathematik die Bevorzugung innermathematischer Aufgabenstellungen gegenüber Modellbildungsaufgaben oder vice versa (Stanat & Pant, 2013).

Einige der genannten Länderdifferenzen, wie etwa die unterschiedlichen Regelungen zu Wahlmöglichkeiten, betreffen insbesondere die Rahmenbedingungen und die Durchführung der Prüfung; sie haben für die Entwicklung und Entnahme von Aufgaben der Pools kaum Relevanz. Demgegenüber stellen Länderunterschiede in Vorgaben und Regelungen für die Struktur von Aufgabenstellungen, Regelungen zu Aufgabenmaterialien und Hilfsmitteln, curricularen Vorgaben oder Bestimmungen für die Formulierung von Erwartungshorizonten und Bewertungshinweisen eine große Herausforderung für den Poolprozess dar. So bestand vor allem in den Anfangsjahren der Abituraufgabenpools das Problem, dass die bereitgestellten Aufgaben in einigen Fällen nicht den in den Ländern geltenden Prüfungsvorgaben und -regelungen entsprachen. Ein unveränderter Einsatz dieser Aufgaben in den Abiturprüfungen der Länder hätte die Prüflinge benachteiligt, da diese mit einer Aufgabenkultur konfrontiert worden wären, die in ihrem Land nicht implementiert war und auf die sie demgemäß nicht vorbereitet gewesen wären. Um den Ländern dennoch den Einsatz von Aufgaben aus den Pools zu ermöglichen, konnten sie Anpassungen an den entnommenen Aufgaben vornehmen. Dabei war ausdrücklich vereinbart worden, die Aufgaben „nur so viel wie nötig und so wenig wie möglich“ zu modifizieren. Der Kern der Aufgaben sollte erhalten und das Anspruchsniveau unverändert bleiben (KMK, 2017).

Solche Modifikationen an den aus dem Pool für das Fach Deutsch entnommenen Abituraufgaben waren etwa notwendig, wenn die Struktur der bereitgestellten Poolaufgaben, die zum Beispiel jeweils zwei Teilaufgaben umfassen, nicht den Vorgaben der betreffenden Länder zur Aufgabenstruktur entsprach. Modifikationen waren beispielsweise auch dann erforderlich, wenn die Abiturregelungen einzelner Länder spezifische, von den Poolaufgaben abweichende Anforderungen an bestimmte Aufgabenarten vorsahen. Solche Abweichungen betrafen etwa die Frage, inwiefern bei Gedichtvergleichen das jeweils zweite Gedicht, wie bei den Poolaufgaben vorgesehen, nur zum Vergleich herangezogen und aspektorientiert interpretiert werden soll oder ob es, wie in einigen Ländern tradiert, ebenfalls vollständig zu analysieren ist. Darüber hinaus machten zum Beispiel auch die in Beitrag 8 (Schröter et al.) des vorliegenden Bandes erläuterten Unterschiede in den Bewertungstraditionen der Länder in

einigen Fällen Modifikationen an den Erwartungshorizonten und Bewertungshinweisen der aus dem Pool des Faches Deutsch entnommenen Abituraufgaben notwendig.

In den Fächern Englisch und Französisch waren Modifikationen an den aus den Pools entnommenen Abituraufgaben zum Beispiel dann notwendig, wenn die im Kompetenzbereich „Schreiben" bereitgestellten Poolaufgaben nicht den Vorgaben der betreffenden Länder zum maximalen Umfang von Ausgangs- und Zieltexten entsprachen. Darüber hinaus machten beispielsweise diskrepante Ländervorgaben zur Bedeutung von Operatoren[6], zu den im Kompetenzbereich „Hörverstehen" zulässigen Itemformaten oder zur Art des in der Prüfung erlaubten Wörterbuchs Modifikationen an den aus den Pools entnommen Abituraufgaben erforderlich. Im Fach Mathematik waren Modifikationen zum Beispiel dann notwendig, wenn die Lösung einer Teilaufgabe der entnommenen Poolaufgabe die Kenntnis spezifischer Inhalte (wie etwa der Ableitungsregeln für eine ganz bestimmte Funktionsart im Rahmen der Differenzialrechnung) voraussetzte, die nicht im Lehrplan des betreffenden Landes verankert waren. Modifikationen an den aus dem Pool des Faches Mathematik entnommenen Abituraufgaben waren beispielsweise auch bei unterschiedlichen Ländervorgaben zur Bedeutung von Operatoren, zur Struktur von Aufgabenstellungen oder zur Formulierung von Erwartungshorizonten erforderlich.

Eine besondere, sowohl das Fach Deutsch als auch die Fächer Englisch und Französisch betreffende Herausforderung stellt zudem bis heute die Entwicklung und Bereitstellung ländergemeinsamer Poolaufgaben dar, welche die Kenntnis eines bestimmten literarischen Werkes voraussetzen. Damit solche Poolaufgaben überhaupt von einer größeren Anzahl von Ländern in der Abiturprüfung eingesetzt werden können, müssen diese bereits drei Jahre vor dem geplanten Einsatzzeitpunkt verabreden, das jeweilige Werk in den für den betreffenden Prüfungsjahrgang gültigen Lehrplan aufzunehmen.

5.4 Annäherung der Länder

Aufgrund der skizzierten Unterschiede in den Aufgabentraditionen der Länder und wegen der diskrepanten Länderregelungen und -vorgaben zu den Abiturprüfungen, wurde gleich zu Beginn der Arbeiten an den Gemeinsamen Abituraufgabenpools deutlich, dass allgemeine Vereinbarungen zur Gestaltung der Aufgaben in den Pools getroffen werden mussten. Entsprechend wurde die Struktur der zu entwickelnden Aufgaben sowie ein Grundstock von Operatoren für die Formulierung von Aufgabenstellungen abgestimmt. Zudem erfolgte eine Festlegung von Anforderungen, denen die Aufgaben der Pools sowie die zugehörigen Erwartungshorizonte und Bewertungshinweise genügen sollten. Diese allgemeinen Vereinbarungen bilden bis heute eine wichtige Arbeitsgrundlage für die ländergemeinsame Entwicklung von Aufgaben und sind auf den Internetseiten des IQB unter der Bezeichnung „begleitende Dokumente" veröffentlicht.

6 Als „Operatoren" bezeichnet man Handlungsanweisungen in Verbform, die in Aufgabenstellungen angeben, was ein Prüfling tun muss, wie z. B. „Beschreibe", Analysiere", „Beurteile".

Die Vereinbarungen galten allerdings zunächst nur für die Poolaufgaben und viele Länder legten Wert darauf, dass sie diese nicht würden übernehmen müssen. Dennoch wurde mit Beginn der Arbeiten an den Gemeinsamen Abituraufgabenpools ein umfangreicher Standardisierungs- und Annäherungsprozess der Länder zu den Regelungen und Vorgaben für Abiturprüfungen und Abiturprüfungsaufgaben ausgelöst, der bis heute andauert. Eine Triebfeder dieses Prozesses bildete zum einen die vonseiten der Amtschefskommission „Qualitätssicherung in Schulen" formulierte Zielvorgabe, dass die aus dem Pool entnommenen Aufgaben ab dem Jahr 2021 nicht mehr modifiziert werden sollten. Zum anderen wurde im Jahr 2020 im Rahmen einer auf der 371. Sitzung der KMK verabschiedeten Ländervereinbarung beschlossen, dass ab dem Jahr 2023 die Abiturprüfungen der Länder in den Fächern Deutsch, Mathematik, Englisch und Französisch zu mindestens 50 Prozent aus Poolaufgaben bestehen sollen. Die Ländervereinbarung sieht zudem vor, die Abituraufgabenpools so weiterzuentwickeln, dass die Abiturprüfungen der Länder auch vollumfänglich aus Aufgaben bestehen können, die den Pools entnommen wurden. Vor diesem Hintergrund wurden die AGs Aufgaben beauftragt, diejenigen Aspekte in den landesspezifischen Vorgaben und Regelungen zu identifizieren, die einem unveränderten Einsatz der Aufgaben der Pools im Wege stehen, und Empfehlungen für eine Vereinheitlichung dieser Aspekte zu formulieren. Dabei handelt es sich um einen sehr aufwändigen, langwierigen und teilweise auch schwierigen Prozess, bei dem sich die Ländervertreterinnen und -vertreter von den eigenen Traditionen lösen und auf gemeinsame Lösungen einigen mussten.

Bei der Planung der anschließenden Umsetzung der Empfehlungen in landesspezifische Vorgaben und Regelungen war zu beachten, dass diese eines gewissen zeitlichen Vorlaufs bedarf, da substanzielle Anpassungen frühestens für den Jahrgang gelten können, der im betreffenden Jahr in die gymnasiale Oberstufe eintritt. Dementsprechend war es erforderlich, realistische Übergangsphasen im Arbeitsprozess vorzusehen und Übergangsfristen zu vereinbaren, innerhalb derer es gestattet ist, die aus dem Pool entnommenen Aufgaben im Hinblick auf die noch nicht im jeweiligen Land angepassten Aspekte zu modifizieren.

Mittlerweile haben alle Länder den von den AGs Aufgaben erarbeiteten Empfehlungen zugestimmt. Die meisten Länder haben die notwendigen Anpassungen bereits vorgenommen oder zumindest angekündigt, entsprechende Änderungen innerhalb der nächsten Prüfungsjahre umsetzen zu wollen. Dementsprechend dürften die Länder inzwischen nur noch in wenigen Fällen gezwungen sein, die von ihnen aus den Pools entnommenen Aufgaben vor dem Einsatz in der Abiturprüfung zu modifizieren. Inwiefern dies tatsächlich zutrifft, wird sich in den zukünftigen Evaluationen der Nutzung der Abituraufgabenpools zeigen.

Die gemeinsame Nutzung der Abituraufgabenpools macht darüber hinaus weitere Vereinbarungen zwischen den Ländern erforderlich. So wurde etwa ein Sicherheitskonzept abgestimmt, das gewährleisten soll, dass die von den Ländern aus den Pools entnommenen Aufgaben nicht vor dem jeweiligen Prüfungstermin öffentlich bekannt werden. Da die Pools in ihrem Umfang begrenzt sind und die bereitgestell-

ten Aufgaben nach Möglichkeit jeweils von mehreren Ländern parallel eingesetzt werden sollen, ist es ferner erforderlich, dass die Länder gemeinsame Prüfungstermine vereinbaren. Diese Abstimmung ist nicht trivial: Sie bedarf zum einen eines entsprechenden zeitlichen Vorlaufs und zum anderen ist zu berücksichtigen, dass die Ferientermine und -regelungen der Länder voneinander abweichen und die Länder jeweils unterschiedlich zeitaufwendige Verfahren zur Bewertung von Abiturprüfungsarbeiten (z. B. Zweit- und Drittkorrekturen durch Lehrkräfte der jeweiligen Schule oder externe Lehrkräfte) haben. Trotz dieser Rahmenbedingungen konnten sich die Länder für die Prüfungsjahre 2017 bis 2021 in den Fächern Deutsch und Mathematik auf einen gemeinsamen Prüfungstermin und in den Fächern Englisch und Französisch auf zwei gemeinsame Prüfungstermine einigen. Ab dem Prüfungsjahr 2022 werden die Abiturprüfungen der Länder dann in jedem der vier beteiligten Fächer (von vereinzelten Ausnahmen abgesehen) zum selben Termin durchgeführt.

Die für die gemeinsame Nutzung der Abituraufgabenpools notwendigen Abstimmungen zwischen den Ländern erfolgen maßgeblich im Rahmen der für die Pools zuständigen Steuergruppe, der sogenannten „AG Abiturkommission“. Diese setzt sich aus Vertreterinnen und Vertretern der Länder zusammen, die jeweils in ihren Ländern für die Steuerung der Abiturprüfungen zuständig sind (Stanat, 2014). Zum Mitgliederkreis der Arbeitsgruppe gehören darüber hinaus eine Vertreterin bzw. ein Vertreter der beruflichen Gymnasien sowie die wissenschaftliche Leitung des IQB, die, gemeinsam mit der Vertreterin bzw. dem Vertreter eines Landes, den Vorsitz der AG Abiturkommission innehat (Stanat & Pant, 2013).

5.5 Ausblick

Ähnlich wie die übrigen Maßnahmen der Länder zur Erhöhung der Vergleichbarkeit des Abiturs werden auch die Gemeinsamen Abituraufgabenpools der Länder von Teilen der wissenschaftlichen Community und der medialen Öffentlichkeit kritisiert. Insbesondere wird argumentiert, die Strategie würde zu kurz greifen, da jeweils nur ein sehr kleiner Anteil an den Abiturgesamtnoten auf den Leistungen in ländergemeinsamen Aufgaben basiere (Brodkorb & Koch, 2020; Klein, 2016). Dabei wird allerdings übersehen, dass das Projekt nicht nur dazu dient, Aufgaben für die Abiturprüfungen der Länder bereitzustellen, sondern über diesen unmittelbaren Zweck hinaus weitere wichtige Prozesse angestoßen hat. Insbesondere wurde ein Standardisierungs- und Annäherungsprozess initiiert, der bis heute andauert und eine zunehmende Konvergenz der in den Ländern geltenden Vorgaben und Regelungen zur inhaltlichen und formalen Gestaltung von Abiturprüfungsaufgaben bewirkt. Gleichzeitig hat die ländergemeinsame Entwicklung von Abituraufgaben intensive Austauschprozesse dazu angeregt, welchen konkreten Anforderungen gute Prüfungsaufgaben in den vier Fächern der Abituraufgabenpools genügen sollten. Dies ist nicht zuletzt deshalb wichtig, weil sich die fachdidaktische Forschung bislang nur selten mit Abiturprüfungen beschäftigt und somit nur wenige wissenschaftlich fundierte Anhaltspunkte zur Gestaltung von Abiturprüfungsaufgaben vorgelegt hat. Ebensolche Anhaltspunkte sollen im Rahmen einer Evaluation des IQB gewonnen werden, die den Einsatz von Aufga-

ben der Pools in den Abiturprüfungen der Länder seit dem Prüfungsjahr 2017 kontinuierlich begleitet und dabei auch spezifische Studien zu Aspekten der Gestaltung von Aufgabenstellungen, Erwartungshorizonten und Bewertungshinweisen durchführt.

Die Gemeinsamen Abituraufgabenpools der Länder sollen also dazu beitragen, die Vergleichbarkeit des Abiturs und die Qualität von Abiturprüfungsaufgaben zu erhöhen. Inwieweit dies tatsächlich gelingt, dürfte maßgeblich davon abhängen, wie konsequent die Länder die Vereinbarungen zu den Abiturprüfungsaufgaben umsetzen, ob sie ihre eigenen Aufgaben an den Pools orientieren und inwieweit die erwarteten Effekte tatsächlich eintreten. Dies sollte in zukünftigen Evaluationsstudien zum Pool geprüft werden.

Ab dem Prüfungsjahr 2025 sollen auch für die naturwissenschaftlichen Fächer Biologie, Chemie und Physik ländergemeinsame Abituraufgabenpools bereitgestellt werden. Die Vorbereitungen für die Entwicklung dieser Pools wurden im Jahr 2021 begonnen.

Durch die Pools allein wird die angestrebte Vergleichbarkeit des Abiturs zwischen den Ländern allerdings nicht zu erreichen sein, in dieser Hinsicht haben die kritischen Stimmen recht. Entsprechend sieht die im Jahr 2021 verabschiedete Ländervereinbarung vor, dass auch die strukturellen Rahmenbedingungen der gymnasialen Oberstufe anzugleichen sind, wozu „Festlegungen zu den Voraussetzungen für den Besuch der gymnasialen Oberstufe, zu Unterrichtsfächern, zum Stundenumfang, für die Zulassung zur Abiturprüfung, zur Gestaltung der Abiturprüfung sowie zu Modalitäten für die Berechnung der Abiturdurchschnittsnote“ gehören (KMK, 2020, S. 20). Welche Länderunterschiede hierbei gegenwärtig bestehen, wird in Beitrag 3 von Schmid-Kühn und Groß analysiert.

Literatur

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

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Tillmann, K.-J. & Weiß, M. (Hrsg.). (2002). *PISA 2000. Die Länder der Bundesrepublik Deutschland im Vergleich*. Leske + Budrich.

Baumert, J., Bos, W. & Lehmann, R. (Hrsg.). (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*. Leske + Budrich.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J. & Weiß, M. (Hrsg.). (2001). *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Leske + Budrich.

Baumert, J. & Köller, O. (2000). Motivation, Fachwahlen, selbstreguliertes Lernen und Fachleistungen im Mathematik- und Physikunterricht 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. 181–213). Leske + Budrich.

Becker-Mrotzek, M., Gippner, G., Kämper-van den Boogaart, M., Köster, J. & Stanat, P. (2015). *Bildungsstandards aktuell: Deutsch in der Sekundarstufe II*. Schroedel.

Blossfeld, H.-P., Bos, W., Daniel, H.-D., Hannover, B., Lenzen, D., Prenzel, M., Roßbach, H.-G., Tippelt, R. & Wößmann, L. (2011). *Gemeinsames Kernabitur. Zur Sicherung von nationalen Bildungsstandards und fairem Hochschulzugang; Gutachten*. Waxmann.

Blum, W., Drüke-Noe, C., Vogel, S. & Roppelt, A. (2015). *Bildungsstandards aktuell: Mathematik in der Sekundarstufe II*. Schroedel.

Boggasch, M. (2011). *Wissenschaftspropädeutik in der Schule – „Musik und Literatur" als wissenschaftspropädeutisches Seminar in der gymnasialen Oberstufe* [Dissertation]. Universität der Künste, Berlin.

Böhme, K. & Hoffmann, L. (2016). Mittelwerte und Streuungen der im Fach Deutsch erreichten Kompetenzen. In P. Stanat, K. Böhme, S. Schipolowski & N. Haag (Hrsg.), *Iqb-bildungstrend 2015. Sprachliche kompetenzen am ende der 9. Jahrgangsstufe im zweiten ländervergleich* (S. 335–358). Waxmann.

Bölling, R. (2010). *Kleine Geschichte des Abiturs*. Schöningh.

Brodkorb, M. & Koch, K. (2020). *Der Abiturbetrug. Vom Scheitern des deutschen Bildungsföderalismus. Eine Streitschrift*. Dietrich zu Klampen.

BVerfG = Bundesverfassungsgericht. (2017). *Urteil des Ersten Senats vom 19. Dezember 2017*. https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/DE/2017/12/ls20171219_1bvl000314.html

Cheng, L., Watanabe, Y. & Curtis, A. (Hrsg.). (2004). *Washback in Language Testing: Research Contexts and Methods*. Lawrence Erlbaum.

Disdorn-Liesen, V. (2016). *Vergleichbarkeit in der Vielfalt*. Springer. https://doi.org/10.1007/978-3-658-12165-5

Grünkorn, J., Klieme, E. & Stanat, P. (2019). Bildungsmonitoring und Qualitätssicherung. In O. Köller, M. Hasselhorn, F. W. Hesse, K. Maaz, J. Schrader, H. Solga, C. K. Spieß & K. Zimmer (Hrsg.), *utb-studi-e-book: Vol. 4785. Das Bildungswesen in Deutschland: Bestand und Potenziale* (S. 263–298). Julius Klinkhardt.

Holmeier, M. (2012). Bezugsnormorientierung im Unterricht im Kontext zentraler Abiturprüfungen. In K. Maag Merki (Hrsg.), *Zentralabitur* (S. 237–261). VS Verlag für Sozialwissenschaften.

Holtmann, M., Becker, B. & Weirich, S. (2019). Mittelwerte und Streuungen der in den naturwissenschaftlichen Fächern erreichten Kompetenzen. In P. Stanat, S. Schipolowski, N. Mahler, S. Weirich & S. Henschel (Hrsg.), *IQB-Bildungstrend 2018. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich* (S. 213–236). Waxmann.

Jäger, D. J. (2012). Herausforderung Zentralabitur: Unterrichtsinhalte variieren und an Prüfungsthemen anpassen. In K. Maag Merki (Hrsg.), *Zentralabitur* (S. 175–201). VS Verlag für Sozialwissenschaften.

Jonkmann, K., Trautwein, U., Nagy, G. & Köller, O. (2010). Fremdsprachenkenntnisse in Englisch vor und nach der Neuordnung der gymnasialen Oberstufe in Baden-Württemberg. In U. Trautwein, M. Neumann, G. Nagy, O. Lüdtke & K. Maaz (Hrsg.), *Schulleistungen von Abiturienten: Die neugeordnete Gymnasiale Oberstufe auf dem Prüfstand* (S. 181–213). VS Verlag für Sozialwissenschaften.

Kampa, N., Leucht, M. & Köller, O. (2016). Mathematische Kompetenzen in unterschiedlichen Profilen der gymnasialen Oberstufe. In J. Kramer, M. Neumann & U. Trautwein (Hrsg.), *Abitur und Matura im Wandel* (S. 161–188). Springer.

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.

Klein, H. P. (2016). *Vom Streifenhörnchen zum Nadelstreifen: Das deutsche Bildungswesen im Kompetenztaumel*. Dietrich zu Klampen.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K. & Riquarts, K. (2007). *Bildungsforschung Band 1. Zur Entwicklung nationaler Bildungsstandards – Expertise*. Bundesministerium für Bildung und Forschung (BMBF).

KMK-Expertenkommission (1995). *Weiterentwicklung der Prinzipien der gymnasialen Oberstufe und des Abiturs: Abschlußbericht der von der Kultusministerkonferenz eingesetzten Expertenkommission*. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (1972). *Vereinbarung zur Neugestaltung der gymnasialen Oberstufe in der Sekundarstufe II: Beschluss der Kultusministerkonferenz vom 7. Juli 1972.*

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (1995). *Richtungsentscheidung zur Weiterentwicklung der Prinzipien der gymnasialen Oberstufe und des Abiturs (Beschluss der Kultusministerkonferenz vom 1.12.1995)*.

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (1999). *Vereinbarung zur Gestaltung der gymnasialen Oberstufe in der Sekundarstufe II: Beschluss der Kultusministerkonferenz vom 22.10.1999*.

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (2004). *Bildungsstandards im Fach Deutsch für den Mittleren Schulabschluss Beschluss vom 4.12.2003*. Luchterhand.

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (2005a). *Bildungsstandards im Fach Deutsch für den Primarbereich. Beschluss vom 15.10.2004.* Luchterhand.

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (2005b). *Bildungsstandards im Fach Deutsch für den Hauptschulabschluss Beschluss vom 15.10.2004.* Luchterhand.

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (2012). *Bildungsstandards im Fach Deutsch für die Allgemeine Hochschulreife (Beschluss der Kultusministerkonferenz vom 18.10.2012)*. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Deutsch-Abi.pdf

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (2015a). *Bericht über den Verfahrensstand bei der Implementation der Bildungsstandards für die Allgemeine Hochschulreife (Beschluss der Kultusministerkonferenz vom 06.11.2014 i. d. F. vom 12.11.2015)*. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2015/2015_11_12-Bericht-Implementation-Bildungsstandards-AHR.pdf

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (2015b). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2015/2015_06_11-Gesamtstrategie-Bildungsmonitoring.pdf

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (2017). *FAQs – Gemeinsamer Abituraufgabenpool der Länder*. https://www.kmk.org/fileadmin/Dateien/pdf/Bildung/AllgBildung/FAQs-Abiturpool.pdf

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik (2020). *Ländervereinbarung über die gemeinsame Grundstruktur des Schulwesens und die gesamtstaatliche Verantwortung der Länder in zentralen bildungspolitischen Fragen (Beschluss der Kultusministerkonferenz vom 15.10.2020)*. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2020/2020_10_15-Laendervereinbarung.pdf

Köller, O. (2007). Bildungsstandards, einheitliche Prüfungsanforderungen und Qualitätssicherung in der Sekundarstufe II. In D. Benner (Hrsg.), *Schöningh and Fink Social Sciences E-Books Online, Collection 2007–2017, ISBN. Bildungsstandards: Instrumente zur Qualitätssicherung im Bildungswesen : Chancen und Grenzen-Beispiele und Perspektiven* (S. 13–28). Schöningh.

Kühn, S. M. (2011). Exploring the use of statewide exit exams to spread innovation—The example of Context in science tasks from an international comparative perspective. *Studies in Educational Evaluation, 37*(4), 189–195. https://doi.org/10.1016/j.stueduc.2012.01.003

Kühn, S. M. (2012). Zentrale Abiturprüfungen im nationalen und internationalen Vergleich mit besonderer Perspektive auf Bremen und Hessen. In K. Maag Merki (Hrsg.), *Zentralabitur* (S. 27–44). VS Verlag für Sozialwissenschaften.

Maag Merki, K. (2012a). Die Leistungen der Gymnasiastinnen und Gymnasiasten in Mathematik und Englisch. In K. Maag Merki (Hrsg.), *Zentralabitur* (S. 263–292). VS Verlag für Sozialwissenschaften.

Maag Merki, K. (Hrsg.). (2012b). *Zentralabitur*. VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-94023-6

Maag Merki, K. (2016). Die Einführung zentraler Abiturprüfungen in Bremen und Hessen. Eine Bilanz nach fünf Jahren. In J. Kramer, M. Neumann & U. Trautwein (Hrsg.), *Abitur und matura im wandel* (S. 129–159). Springer.

Mahler, N. & Kölm, J. (2019). Mittelwerte und Streuungen der im Fach Mathematik erreichten Kompetenzen. In P. Stanat, S. Schipolowski, N. Mahler, S. Weirich & S. Henschel (Hrsg.), *IQB-Bildungstrend 2018. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich* (S. 201–212). Waxmann.

Maué, E. & Maag Merki, K. (2019). Zentrale Prüfungen – Wirkungspotenziale aus theoretischer und empirischer Perspektive. In N. Berkemeyer, W. Bos & B. Hermstein (Hrsg.), *Schulreform. Zugänge, gegenstände, trends* (S. 345–357). Beltz.

Nagy, G., Neumann, M., Becker, M., Watermann, R., Köller, O., Lüdtke, O. & Trautwein, U. (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). Waxmann.

Nagy, G., Neumann, M., Trautwein, U. & Lüdtke, O. (2010). Voruniversitäre Mathematikleistungen vor und nach der Neuordnung der gymnasialen Oberstufe in Baden-Württemberg. In U. Trautwein, M. Neumann, G. Nagy, O. Lüdtke & K. Maaz (Hrsg.), *Schulleistungen von Abiturienten: Die neugeordnete gymnasiale Oberstufe auf dem Prüfstand* (S. 147–180). VS Verlag für Sozialwissenschaften.

Neumann, M. (2010). Innovation oder Restauration – Die (Rück-?)Reform der gymnasialen Oberstufe in Baden-Württemberg. In U. Trautwein, M. Neumann, G. Nagy, O. Lüdtke & K. Maaz (Hrsg.), *Schulleistungen von Abiturienten: Die neugeordnete gymnasiale Oberstufe auf dem Prüfstand*(S. 37–90). VS Verlag für Sozialwissenschaften.

Neumann, M., Trautwein, U. & Baumert, J. (2012). Die Neuordnung der gymnasialen Oberstufe aus empirischer Perspektive: Hintergründe, Befunde und steuerungsrelevante Implikationen der TOSCA-Repeat-Studie. In A. Wacker, U. Maier & J. Wissinger (Hrsg.), *Schul- und Unterrichtsreform durch ergebnisorientierte Steuerung* (S. 277–301). VS Verlag für Sozialwissenschaften.

Oerke, B., Maag Merki, K., Maué, E. & Jäger, D. J. (2013). Zentralabitur und Themenvarianz im Unterricht: Lohnt sich Teaching-to-the-Test? In D. Bosse, F. Eberle & B. Schneider-Taylor (Hrsg.), *Standardisierung in der gymnasialen Oberstufe* (S. 27–49). Springer Fachmedien Wiesbaden.

Roeder, P. M. & Gruehn, S. (1996). Kurswahlen in der Gymnasialen Oberstufe. *Zeitschrift für Pädagogik, 42*, 497–518.

Schipolowski, S. & Sachse, K. (2016). Mittelwerte und Streuungen der im Fach Englisch erreichten Kompetenzen. In P. Stanat, K. Böhme, S. Schipolowski & N. Haag (Hrsg.), *IQB-Bildungstrend 2015. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich* (S. 359–375). Waxmann.

Schnabel, K. & Gruehn, S. (2000). Studienfachwünsche und Berufsorientierungen 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. 405–453). Leske + Budrich.

Stanat, P. (2014, January 17). *Aktuelle Entwicklungen der Überprüfung von Bildungsstandards in Deutschland*. DACH-Seminar 2014, Potsdam. https://bildungsserver.berlin-brandenburg.de/fileadmin/bbb/schule/schulentwicklung/bildungsdiskussion_berichte_und_studien/D-A-CH_Seminar_2014/Dokumentation/2014-02-17-9-40_Praesentation_Bildungsstandards_Prof._Stanat.pdf

Stanat, P., Becker-Mrotzek, M., Blum, W. & Tesch, B. (2016). Vergleichbarkeit in der Vielfalt. Bildungsstandards der Kultusministerkonferenz für die Allgemeine Hochschulreife. In J. Kramer, M. Neumann & U. Trautwein (Hrsg.), *Abitur und Matura im Wandel* (S. 29–58). Springer.

Stanat, P. & Pant, H. A. (2013). *Konzeption für die Entwicklung und Nutzung eines Pools von Abiturprüfungsaufgaben (Arbeitsfassung)*. http://docplayer.org/49069143-Konzeption-fuer-die-entwicklung-und-nutzung-eines-pools-von-abiturpruefungsaufgaben-arbeitsfassung.html

Statistisches Bundesamt (2020). *Anteil der Schulabsolventen/-innen mit allgemeiner Hochschulreife* an der gleichaltrigen Bevölkerung in Deutschland nach Bundesländern im Abgangsjahr 2019 [Graph]*. https://de.statista.com/statistik/daten/studie/255393/umfrage/anteil-der-schulabsolventen-innen-mit-abitur-in-deutschland-nach-bundeslaendern/

StMUK (= Bayerisches Staatsministerium für Unterricht und Kultus) (2012). *Länderübergreifendes Abitur; Gemeinsame Aufgaben bzw. Aufgabenteile in der schriftlichen Abiturprüfung (Schreiben vom 27.02.2012)*. https://www.isb.bayern.de/download/22378/laenderuebergreifendes_abitur_gemeinsame_aufgaben_in_der_schriftlichen_abiturpruefung.pdf

Tenorth, H.-E. (1996). Reform der gymnasialen Oberstufe - Praxis ihrer Arbeit. Zur Einleitung in den Themenschwerpunkt. *Zeitschrift Für Pädagogik, 42*, 493–496.

Tesch, B., Hammerstein, X. von, Stanat, P. & Rossa, H. (2017). *Bildungsstandards aktuell: Englisch/Französisch in der Sekundarstufe II*. Schroedel.

Trautwein, U., Neumann, M., Nagy, G., Lüdtke, O. & Maaz, K. (2010). Institutionelle Reform und individuelle Entwicklung: Hintergrund und Fragestellungen der Studie TOSCA-Repeat. In U. Trautwein, M. Neumann, G. Nagy, O. Lüdtke & K. Maaz (Hrsg.), *Schulleistungen von Abiturienten: Die neugeordnete gymnasiale Oberstufe auf dem Prüfstand* (S. 15–36). VS Verlag für Sozialwissenschaften.

Wolf, F. & Heinz, D. (2016). Schulpolitik: neue Koordination und neue Unterschiede. In A. Hildebrandt & F. Wolf (Hrsg.), *Die Politik der Bundesländer: Zwischen Föderalismusreform und Schuldenbremse* (2. Aufl., S. 11–34). Springer.

Wößmann, L. (2003). Zentrale Prüfungen als „Währung" des Bildungssystems: Zur Komplementarität von Schulautonomie und Zentralprüfungen. *Vierteljahrshefte zur Wirtschaftsforschung, 72*, 220–237.

Wößmann, L. (2008). Zentrale Abschlussprüfungen und Schülerleistungen. Individualanalysen anhand von vier internationalen Tests. *Zeitschrift für Pädagogik, 54*, 810–826.

Wößmann, L. (2012). Ein Gemeinsames Kernabitur für Deutschland: Der Vorschlag des Aktionsrats Bildung. *Ifo Schnelldienst, 65*(02), 12–21.

3 Struktur der gymnasialen Oberstufe und Rahmenbedingungen für die Abiturprüfung im Ländervergleich

Svenja Mareike Schmid-Kühn & Alexander Groß

Zusammenfassung

Die Ausgestaltung der Struktur der gymnasialen Oberstufe und der formal-organisatorischen Rahmenbedingungen für die Abiturprüfung obliegen aufgrund des föderalen Systems in Deutschland den einzelnen Bundesländern, wobei die ländergemeinsam verabschiedete *Vereinbarung zur Gestaltung der gymnasialen Oberstufe und der Abiturprüfung* der Kultusministerkonferenz einen Referenzrahmen vorgibt, der sowohl bundesweit einheitliche Setzungen enthält als auch länderspezifische Gestaltungsspielräume gewährt. In diesem Beitrag werden die Ergebnisse einer Dokumentenanalyse vorgestellt: Die für die gymnasiale Oberstufe und die Abiturprüfung relevanten, institutionell verantworteten Vorgaben (z. B. Abiturprüfungsordnungen) aus allen 16 Bundesländern wurden kategoriengeleitet analysiert, um die strukturelle Ausgestaltung der gymnasialen Oberstufe und die formal-organisatorische Ausgestaltung der Abiturprüfungsverfahren in allen Ländern zu erfassen und einer vergleichenden Betrachtung zu unterziehen. Der Beitrag schließt mit der Frage, wie vergleichbar die länderspezifischen Strukturen und Rahmenbedingungen sind, wobei der Begriff der Vergleichbarkeit aus unterschiedlichen Perspektiven betrachtet wird. Die Darstellung der Limitationen des Ländervergleichs sowie ein Ausblick auf weiterführende Forschungsperspektiven runden den Beitrag ab.

1 Einleitung

Im Zuge der wiederkehrenden Diskussion in (Bildungs-)Politik, Wissenschaft und Öffentlichkeit über die (mangelnde) Vergleichbarkeit des Abiturs in Deutschland hat die Kultusministerkonferenz (KMK) in den vergangenen Jahren verschiedene Maßnahmen auf den Weg gebracht, die eine stärkere Vergleichbarkeit der Leistungsanforderungen (vgl. Beitrag 2 von Hoffmann, Schröter & Stanat in diesem Band) sowie der organisatorischen Ausgestaltung der gymnasialen Oberstufe und der Abiturprüfung gewährleisten sollen. Der vorliegende Beitrag thematisiert die Struktur der gymnasialen Oberstufe sowie die Rahmenbedingungen für die Abiturprüfung in allen 16 Bundesländern. Die differenzierte Betrachtung dieser organisatorischen Aspekte im Ländervergleich ist mit Blick auf die Forderung nach mehr Vergleichbarkeit insofern

relevant, als dass sich die Abiturdurchschnittsnote zu zwei Dritteln aus den Leistungen der vier Schulhalbjahre der Qualifikationsphase der gymnasialen Oberstufe (Block I der Gesamtqualifikation) und zu einem Drittel aus den in der Abiturprüfung gezeigten Leistungen (Block II der Gesamtqualifikation) zusammensetzt. Folglich tangieren länderspezifische Unterschiede in der Ausgestaltung der gymnasialen Oberstufe (z. B. Pflichtfächer, Anzahl der Unterrichtsstunden) und der Abiturprüfung (z. B. Anzahl abzulegender Prüfungen, zentral oder dezentral gestellte Prüfungsaufgaben) die intendierte bundesweite Vergleichbarkeit des Abiturs. Ziel des Beitrags ist es daher, die strukturelle Ausgestaltung der gymnasialen Oberstufe und die formal-organisatorische Ausgestaltung der Abiturprüfungsverfahren in allen Ländern zu erfassen, zu vergleichen und diese im Kontext der fortwährenden Vergleichbarkeitsdebatte zu verorten.

Die Grundstruktur der gymnasialen Oberstufe sowie die grundsätzlichen Rahmenbedingungen für die Abiturprüfung und den Erwerb der Allgemeinen Hochschulreife wurden erstmals in der Bonner Vereinbarung der KMK vom 07.07.1972 festgelegt, die seitdem mehrfach fortgeschrieben wurde, seit 2018 unter dem Titel *Vereinbarung zur Gestaltung der gymnasialen Oberstufe und der Abiturprüfung*. Diese Rahmenvereinbarung umfasst seit jeher bundesweit gültige Vorgaben (z. B. die Gliederung in eine einjährige Einführungsphase und eine zweijährige Qualifikationsphase, das Punktesystem zur Ermittlung der Gesamtqualifikation), gewährt den Ländern gleichwohl auch Spielräume bei der Ausgestaltung. Überblicksarbeiten (Klein et al., 2009; Kühn, 2010; Aktionsrat Bildung, 2011; Neumann & Trautwein, 2014) haben deutliche Länderunterschiede in der Ausgestaltung der gymnasialen Oberstufe und der Abiturprüfungsverfahren aufgezeigt. Die vorgenannten Arbeiten spiegeln jedoch in weiten Teilen nicht mehr die aktuelle Situation wider, da im Zuge der jüngeren Entwicklungen im Sekundarbereich II (vgl. Beitrag 2 von Hoffmann, Schröter & Stanat in diesem Band) Annäherungsprozesse der Länder hinsichtlich der Struktur der gymnasialen Oberstufe sowie der Rahmenbedingungen für die Abiturprüfung erfolgt sind. Der vorliegende Beitrag stellt insofern einen aktualisierten Gesamtüberblick dar.

2 Methodisches Vorgehen

Die nachfolgenden Ausführungen zur Struktur der gymnasialen Oberstufe und zur formal-organisatorischen Ausgestaltung der Abiturprüfungsverfahren beziehen sich auf die Vorgaben für den Abiturjahrgang 2021[1] an allgemeinbildenden Schulformen[2] mit gymnasialer Oberstufe aus allen 16 Bundesländern. Dabei werden überwiegend

1 Im Kontext der Corona-Pandemie haben mehrere Bundesländer für die Abiturjahrgänge 2020 und/oder 2021 die Bedingungen für die Abiturprüfung im Vergleich zu den üblicherweise geltenden Regelungen verändert (z. B. Bereitstellung zusätzlicher Auswahlaufgaben, Gewährung zusätzlicher Auswahl-/Bearbeitungszeit, Ausschluss potenziell möglicher Prüfungsthemen u. Ä.). Diese pandemiebedingten Ausnahmeregelungen sind *nicht* Gegenstand dieses Beitrags; die Darstellungen beziehen sich auf die üblicherweise geltenden Vorgaben.

2 Analysiert wurden die Vorgaben für die Regelform der gymnasialen Oberstufe an allgemeinbildenden Gymnasien und Gesamtschulen (bzw. – je nach länderspezifischer Bezeichnung – andere Schularten mit mehreren Bildungsgängen, die zur allgemeinen Hochschulreife führen). Etwaige abweichende Regelungen (z. B. Waldorfschulen, Schulen des zweiten Bildungsweges u. Ä.) werden in diesem Beitrag nicht thematisiert.

grundsätzliche Strukturen und allgemeine Rahmenbedingungen erfasst – lediglich mit Blick auf die schriftlichen Abiturprüfungen werden auch fachspezifische Aspekte in den Blick genommen, wobei der Fokus auf den Fächern Deutsch und Mathematik sowie den fortgeführten Fremdsprachen (Englisch/Französisch) liegt, da für diese Fächer Bildungsstandards der KMK vorliegen (KMK, 2012a, b, c), die in allen Ländern bereits rechtlich verankert sind (z. B. standardbasierte Lehrpläne und Abiturprüfungen), und in allen Bundesländern mindestens eine Prüfungsaufgabe im schriftlichen Abitur in diesen Fächern aus den ländergemeinsamen Abituraufgabenpools (vgl. Beitrag 2 von Hoffmann, Schröter & Stanat in diesem Band) gestellt wird. Die fachspezifischen Analysen sind weitgehend auf formal-strukturelle Rahmenbedingungen konzentriert; inhaltsbezogene bzw. fachdidaktische Aspekte (z. B. geprüfte Kompetenzen, Aufgabenarten) sind nicht Gegenstand dieses Beitrags.

Die Darstellung stützt sich auf eine systematische Analyse von für die gymnasiale Oberstufe und die Abiturprüfung relevanten, institutionell verantworteten Vorgaben; dazu gehören insbesondere die länderspezifischen Abiturprüfungsordnungen sowie ergänzende Verwaltungsvorschriften, Verfahrensordnungen, Erlasse und/oder sonstige öffentlich zugängliche Dokumente (z. B. Kontaktbriefe, Rundschreiben, Leitfäden u. Ä.), die von den zuständigen Ministerien und/oder den Landesinstituten bzw. Qualitätseinrichtungen der Länder bereitgestellt werden. In Fällen, in denen die verfügbaren Informationen unklar oder unvollständig erschienen, wurden die über die Homepages der zuständigen Institutionen ermittelten fachlichen Ansprechpersonen in den Ländern schriftlich kontaktiert, um eine Klärung herbeizuführen.

Die Dokumentenanalyse erfolgte kategoriengeleitet für die Bereiche *Struktur der gymnasialen Oberstufe* und *Rahmenbedingungen der Abiturprüfung*, wobei nur solche Aspekte analysiert wurden, die in der Debatte über die (mangelnde) Vergleichbarkeit des Abiturs wiederholt diskutiert werden. Die Kategorienbildung erfolgte entlang eines deduktiv-induktiven Vorgehens: Die Hauptkategorien wurden auf Basis der KMK-Vorgaben – der bereits erwähnten *Vereinbarung zur Gestaltung der gymnasialen Oberstufe und der Abiturprüfung* (i. d. F. vom 15.02.2018) und den *Bildungsstandards für die Allgemeine Hochschulreife* (Deutsch, Mathematik, fortgeführte Fremdsprache) – erstellt, wobei diese im Laufe der Analyse der länderspezifischen Dokumente induktiv ergänzt sowie spezifiziert wurden. Folgende Analysekategorien strukturieren den bundesweiten Vergleich:

- Struktur der gymnasialen Oberstufe (siehe Abschnitt 3)
  - *Kanon unterrichteter Fächer in der Oberstufe* (ländergemeinsam/länderspezifisch)
  - *Differenzierung von Anforderungsniveaus* (grundlegend und erhöht, fächer-/kurspezifische Umsetzung in den Ländern)
  - *Belegpflichten in der Qualifikationsphase* (Mindestanzahl zu belegender Fächer, Mindestanzahl der Unterrichtsstunden pro Woche)
  - *Einbringpflichten in der Qualifikationsphase* (Gesamtzahl der Halbjahresergebnisse als Teil der Gesamtqualifikation)

- Rahmenbedingungen der Abiturprüfung (siehe Abschnitt 4)
  - *Gesamtanzahl abzulegender Prüfungen* (Anzahl verpflichtender schriftlicher und mündlicher Prüfungen, Anforderungsniveau der Prüfungen)
  - *allgemeine Rahmenbedingungen der schriftlichen Abiturprüfung* (zentrale und/oder dezentrale Aufgabenstellung, Vorgaben für die Korrektur und Bewertung der Prüfungsleistungen, Korrekturverfahren)
  - *fachspezifische Rahmenbedingungen der schriftlichen Abiturprüfung* (zeitlicher Umfang des der Prüfung vorgelagerten Unterrichts, Vorgabe von Schwerpunktthemen (o. Ä.), Pflichtaufgaben und Möglichkeiten der Aufgabenauswahl, zeitlicher Umfang der Prüfung, zugelassene Hilfsmittel)
  - *mündliche Prüfungen* (Varianten der verpflichtenden mündlichen Abiturprüfung, Einzel- vs. Gruppenprüfungen, Prüfungsdauer)

Zunächst erfolgte eine Beschreibung der landesspezifischen Strukturen und Verfahren in Form einzelner Länderberichte. Diese können aufgrund ihres Umfangs in diesem Beitrag jedoch nicht dargestellt werden. Auf der Grundlage der o. g. Analysekategorien werden die Länderbefunde nach einheitlichen Kriterien nebeneinandergestellt und analysiert sowie einer vergleichenden Betrachtung unterzogen.

3 Struktur der gymnasialen Oberstufe

3.1 Fächerkanon und Anforderungsniveaus

Die KMK gibt mit der Benennung von Aufgabenfeldern einen Rahmen für den Fachunterricht in der gymnasialen Oberstufe vor (KMK 2018, Ziff. 4); der KMK-Beschluss beinhaltet eine allgemeine Untergliederung in drei Aufgabenfelder – das *sprachlich-literarisch-künstlerische Aufgabenfeld*, das *gesellschaftswissenschaftliche Aufgabenfeld* und das *mathematisch-naturwissenschaftlich-technische Aufgabenfeld* (KMK, 2018, Ziff. 4.1) – ergänzt um Religionslehre und Sport.[3] Innerhalb dieser Felder werden entweder zugehörige Unterrichtsfächer explizit benannt oder inhaltlich umschrieben. Die Bundesländer ordnen den Aufgabenfeldern selbstständig spezifische Fächer zu; zudem besteht die Möglichkeit, bereits festgeschriebene Fächer innerhalb eines Feldes frei um weitere zu ergänzen (KMK, 2018, Ziff. 4.3). Einen Überblick über die bundesweit in der Oberstufe unterrichteten Fächer gibt Tabelle 1.

Hier wird deutlich, dass auf der einen Seite ein gemeinsamer Kanon von in allen Bundesländern unterrichteten Fächern existiert (13 Fächer zzgl. Sport), welcher z. T. umfassend länderseitig ergänzt wird. Konkret erweitern insgesamt sieben Länder ihr Fächerangebot um 17 zusätzliche Fächer oder mehr, neun Länder um 13 oder weniger;[4] das insgesamt umfassendste Angebot bietet Bayern mit einem Additum von über 20, die wenigsten Ergänzungen des Grundkanons nimmt Schleswig-Holstein

3 Das Fach Sport wird keinem dieser Felder zugewiesen. Das Fach Religionslehre kann durch die Länder einem der Felder frei zugeordnet werden – wird aber zumeist dem gesellschaftswissenschaftlichen Aufgabenfeld zugeordnet (vgl. Tabelle 1).

4 Hierbei wurden nur die in den Verordnungen explizit aufgelisteten Fächer berücksichtigt.

vor mit insgesamt acht zusätzlichen Fächern. Auf dieser Grundlage ergibt sich in Bezug auf die belegbaren Fächer in der gymnasialen Oberstufe im Ländervergleich ein sehr heterogenes Bild. Dies ist auch der Tatsache geschuldet, dass in dem am wenigsten vorstrukturierten gesellschaftswissenschaftlichen Aufgabenfeld die Länder unterschiedlichste Zugänge zur Vermittlung politischer, sozialer, wirtschaftlicher und rechtlicher Kompetenzen wählen; aufgrund jeweils unterschiedlicher länderinterner Schwerpunktsetzungen sind allein in diesem Bereich bundesweit über ein Dutzend unterschiedliche Fächer mit ähnlichem thematischen Bezug entstanden.[5]

Tabelle 1: Gesamtkanon unterrichteter Fächer in der gymnasialen Oberstufe

	ländergemeinsamer Grundkanon*	**darüber hinaus in den Ländern unterrichtete Fächer (Pflicht- oder Wahlbereich, mit Lehrplan, ohne Länderzuordnung)**
sprachlich-literarisch-künstlerisches Aufgabenfeld	Deutsch, Englisch, Französisch, Latein, (Bildende) Kunst, Musik	Arabisch, Chinesisch, Dänisch, Darstellen und Gestalten, Darstellendes Spiel, Farsi, Alt- u. Neugriechisch, Hebräisch, Italienisch, Japanisch, Literatur, Literatur und Theater, Niederdeutsch, Niederländisch, Polnisch, Portugiesisch, Russisch, Sorbisch (Wendisch), Spanisch, Schwedisch, Theater, Tschechisch, Türkisch, Vokal-/Instrumentalensemble, Theater und Film, Vertiefungskurs Sprache
mathematisch-naturwissenschaftlich-technisches Aufgabenfeld	Mathematik, Biologie, Chemie, Physik, Informatik	Astronomie, Astrophysik, Bautechnik, biologisch-chemisches Praktikum, Biophysik, Chemietechnik, Darstellende Geometrie, Elektrotechnik, Ernährungslehre (mit Chemie), Geologie, Gestaltungs- und Medientechnik, Maschinentechnik, Problemlösen mit einem Computer-Algebra-System, Technik, Vertiefungskurs Mathematik, Wirtschaftsinformatik
gesellschaftswissenschaftliches Aufgabenfeld	Geschichte[a], Erdkunde/Geografie[b]	Archäologie, Erziehungs- wissenschaft, Ethik, Gemeinschaftskunde, Gemeinschaftskunde/Rechtserziehung/Wirtschaft, Geologie, Pädagogik, Politische Bildung, Politik, Politik/Gesellschaft/Wirtschaft, Politik-Wirtschaft, Philosophie, Psychologie, Rechnungswesen, Recht, Rechtskunde, Religion bzw. Religionslehre, Sozialkunde, Sozialwissenschaften, Soziologie, Werte und Normen, Wirtschaft, Wirtschaftslehre, Wirtschaft/Politik, Wirtschaft und Recht, Wirtschaftswissenschaft

Anmerkungen: * Das Fach Sport ist keinem dieser Felder zugewiesen und entsprechend hier nicht aufgeführt, wird jedoch in allen Ländern unterrichtet. [a] In Mecklenburg-Vorpommern wird das integrierende Fach „Geschichte und Politische Bildung" unterrichtet, in Bayern das Kombinationsfach „Geschichte und Sozialkunde". [b] In Rheinland-Pfalz wird auf grundlegendem Anforderungsniveau das Kombinationsfach Sozialkunde/Erdkunde unterrichtet.

Bezogen auf die dem Fachunterricht zugrunde liegenden Anforderungen sieht die KMK-Vereinbarung eine Unterteilung in die beiden Anspruchsebenen *grundlegendes* (gA) sowie *erhöhtes Anforderungsniveau* (eA) (KMK, 2018, Ziff. 3.2) vor. Hierbei ist mit dem Unterricht auf grundlegendem Niveau konkret das Ziel einer wissenschaftspropädeutischen Ausbildung der Schülerinnen und Schüler verbunden, welche auf

5 Die Frage der Integration bzw. Separation im gesellschaftswissenschaftlichen Aufgabenfeld wird auch aus fachdidaktischer Perspektive diskutiert (vgl. z. B. Weber, 2019).

erhöhtem Anforderungsniveau weiter vertieft wird (KMK 2018, Ziff. 3.2). In den Ländern werden entsprechende fächer-/kursbezogene Ausdifferenzierungen dieser Niveaus vorgenommen, wie Tabelle 2 zu entnehmen ist.

Tabelle 2: Anforderungsniveaus und Stundenumfang in der Qualifikationsphase*

	grundlegendes Anforderungsniveau (gA)		erhöhtes Anforderungsniveau (eA)	
	Bezeichnung	**Stundenumfang pro Fach und Woche**	**Bezeichnung**	**Stundenumfang pro Fach und Woche**
KMK	gA	2–4	eA	4–5
BW	Basisfächer	2–3	Leistungsfächer	5
BY[a]	alle Fächer außer Grundlagenfächer	2–3	Grundlagenfächer	4
BE	Grundkurse	2–4	Leistungskurse	5
BB	Grundkurse	3–4	Leistungskurse	5
HB	Grundkurse	2–3	Leistungskurse	5
HH	Kernfach + sonstige Fächer	2–4	Profilgebendes Fach, Kernfächer[b]	4
HE	Grundkurse	2–4	Leistungskurse	5
MV	Grundkurse	2–3	Leistungskurse	5
NI	Kernfächer/Ergänzungsfächer/ Wahlfächer[c]	2–4	Schwerpunktfächer	5
NW	Grundkurse	3–4	Leistungskurse	5
RP	Grundkurse	2–3	Leistungskurse	4–5
SL	Grundkurse	2–4	Leistungskurse	5
SN	Grundkurse	2–4	Leistungskurse	5
ST	Kernfächer/Profil-/[d] Wahlpflicht-fächer	2–3	Kernfächer/Profil-fächer	5
SH	Profil ergänzende Fächer, alle sonstigen Fächer	2–4	Kernfächer, Profil gebende Fächer	4
TH[e]	Fächer mit grundlegendem Anforderungsniveau	2–3	Fächer mit erhöhtem Anforderungsniveau	4–5

Anmerkungen: * ohne ggf. noch zusätzlich vorhandene Projekt-/Seminar-/Vertiefungskurse. [a] Bayern sieht keine *explizite* Unterscheidung verschiedener Anforderungsniveaus vor; gleichwohl werden drei Grundlagenfächer (Deutsch, Mathematik, fortgeführte Fremdsprache) als einzige vierstündig und auf erhöhtem Anforderungsniveau unterrichtet (siehe hierzu auch Neumann & Trautwein, 2019, S. 546). [b] Insgesamt werden drei Kernfächer belegt, wovon mindestens zwei auf erhöhtem Anforderungsniveau unterrichtet werden. [c] Eines dieser Fächer wird je nach Vorgabe zum dritten Prüfungsfach und dann auch ab der Qualifikationsphase auf erhöhtem Anforderungsniveau unterrichtet. [d] Aus dem Kern- und Profilbereich (D, M, FS, NaWi) sind zu Beginn der Qualifikationsphase drei Fächer als schriftliche Prüfungsfächer auszuwählen; diese werden dann auf erhöhtem Anforderungsniveau unterrichtet. [e] Deutsch und Mathematik werden explizit als Kernfächer bezeichnet; eines der beiden Fächer muss auf erhöhtem, das andere auf grundlegendem Anforderungsniveau belegt werden.

Die Länder greifen die vorgegebenen Anspruchsebenen auf und bilden diese z. T. auch unmittelbar in ihrer spezifischen Fächer-/Kursstruktur ab. Hierbei setzen insgesamt 10 Bundesländer[6] die tradierte dichotomische Unterscheidung von Grund- und Leistungskursen um; in vier Bundesländern sieht man zwar ebenso inhaltliche Niveauabstufungen vor, gleichzeitig ist die Bezeichnung (und damit Zuordnung) entsprechender Fächer bzw. Kurse (z. B. Kern- oder Profilfächer) weniger eindeutig, da man dort auf eine solche grundsätzliche Differenzierung verzichtet. Bayern bildet eine Ausnahme in diesem Kontext, da eine distinkte Differenzierung von Anforderungsniveaus nicht (explizit) vorgenommen wird – gleichwohl werden vier Grundlagenfächer benannt, welche auf erhöhtem Niveau unterrichtet werden.

Der Unterricht auf grundlegendem Anforderungsniveau soll gemäß KMK-Vereinbarung zwei- bis vierstündig pro Woche erteilt werden, wobei für die Fächer Deutsch und Mathematik sowie die Fremdsprachen im Sinne einer qualitativen Abstufung eine wöchentliche Mindeststundenzahl von drei Stunden festgelegt ist (vgl. KMK, 2018, Ziff. 7.2). Alle Länder bewegen sich innerhalb dieses Rahmens; zwei Länder verzichten gleichwohl auf zweistündig unterrichtete Fächer. Der Unterrichtsumfang in Kursen mit erhöhtem Anforderungsniveau beträgt vier oder fünf Wochenstunden – mit Ausnahme von Bayern, Hamburg und Schleswig-Holstein sehen alle Länder fünf Wochenstunden vor.[7]

Über den sich aus den Aufgabenfeldern ergebenen Fachunterricht auf unterschiedlichen Niveaustufen hinaus ist in 11 Ländern noch Unterricht in einem weiteren Fach-/Kursformat mit zumeist fächerübergreifenden oder überfachlichen Perspektiven vorgesehen (z. B. Seminar-/Projekt-/Vertiefungskurse; die Bezeichnungen variieren länderspezifisch), wobei keine Differenzierung von Anforderungsniveaus stattfindet.[8]

3.2 Beleg- und Einbringpflichten in der Qualifikationsphase

Wenn es um die Frage nach dem Zustandekommen der Abiturdurchschnittsnote geht, spielen wie einleitend erwähnt neben der Prüfung selbst (vgl. Abschnitt 4) die vier Halbjahre der Qualifikationsphase eine zentrale Rolle, da die in diesem Kontext erbrachten Leistungen zwei Drittel der Abiturdurchschnittsnote ausmachen. Insofern sind die in den Ländern implementierten Vorgaben dazu, wie viele Fächer, in welchem Stundenumfang und auf welchem Anforderungsniveau belegt werden müssen, von besonderer Bedeutung. Auf Basis jener *Belegpflichten* wird von den Ländern des Weiteren festgelegt, wie viele der Halbjahresergebnisse letztlich in die Gesamtqualifikation einfließen (*Einbringpflichten, Block I*). Die länderspezifischen Beleg- und Einbringpflichten in der Qualifikationsphase sind Tabelle 3 zu entnehmen.

6 Baden-Württemberg verwendet eine leicht abweichende Terminologie.

7 In Rheinland-Pfalz werden die Leistungsfächer Geschichte, Erdkunde und Sozialkunde nur vierstündig unterrichtet; alle anderen Leistungsfächer fünfstündig.

8 Die länderspezifischen Regelungen zu diesen Kursformaten (z. B. inhaltliche Schwerpunkte, als Vertiefung zu anderen Fächern oder als Additum, Belegpflicht, Verknüpfung mit der Besonderen Lernleistung etc.) können in diesem Beitrag nicht ausführlich dargestellt werden. Eine erste Durchsicht der entsprechenden Vorgaben weist jedoch auf eine große Bandbreite an Konzepten in den Ländern und an Möglichkeiten zur Ausgestaltung auf Einzelschulebene hin.

Laut KMK-Vereinbarung müssen mindestens acht Fächer in der Qualifikationsphase belegt werden, und zwar Deutsch, eine fortgeführte Fremdsprache, ein Fach aus dem literarisch-künstlerischen Fächerspektrum, Geschichte (oder ein anderes gesellschaftswissenschaftliches Fach, in dem Geschichte mit festen Anteilen unterrichtet wird)[9], Mathematik, eine Naturwissenschaft (Biologie, Chemie oder Physik), Sport und Religionslehre bzw. ein dementsprechendes Ersatzfach (KMK 2018, Ziff. 7.1); mit Ausnahme des literarischen/künstlerischen Faches müssen alle Fächer durchgängig während aller vier Schulhalbjahre der Qualifikationsphase belegt werden.

Tabelle 3: Beleg- und Einbringpflichten in der Qualifikationsphase

	Mindestanzahl zu belegender Fächer[a]			Unterrichtsstunden pro Woche (Mindestanzahl)	Anzahl der einzubringenden Halbjahresergebnisse
	insgesamt	davon auf gA zu belegen	davon auf eA zu belegen		
KMK	8	–	2–4	–	32–40
BW	11	8	3	32	40
BY	10	7	3	33	40
BE	9	7	2	33	32
BB	9	7	2	34	38
HB	9	7	2	34	40
HH	9	5	4	34	32–40
HE	10	8	2	33	32
MV	9	7	2	35	36
NI	10	7	3	32	32–36
NW	9	7	2	34	35–40
RP	10	7	3	32	35
SL	10	8	2	34	40
SN	11	9	2	32	40
ST	11	8	3	34	36–40
SH	11	7	4	32	36
TH	11	7	4	34,5	40

Anmerkung: [a] Die z. T. in einigen Ländern verpflichtend zu belegenden Seminarfächer/-kurse werden an dieser Stelle aus Gründen der Vergleichbarkeit nicht berücksichtigt.

Alle Bundesländer überschreiten die KMK-Empfehlung von acht belegpflichtigen Fächern: in sechs Ländern müssen mindestens neun, in fünf Ländern mindestens 10 und in weiteren fünf Ländern mindestens 11 Fächer belegt werden; teilweise ist zudem zusätzlich noch ein Seminar-/Projekt-/Vertiefungskurs (o. Ä.) belegungspflichtig (siehe Fußnote 8). Laut KMK-Beschluss müssen darunter zwei bis vier Fächer auf erhöhtem Anforderungsniveau sein (ebd., Ziff. 7.2); darunter (nach Wahl der Schülerin-

9 Sofern ein gesellschaftswissenschaftliches Fach gewählt wird, in dem Geschichte nicht mit festen Anteilen unterrichtet wird, sind zusätzlich mindestens zwei Schulhalbjahre Geschichte zu belegen (KMK, 2018, Ziff. 7.1, Spiegelpunkt 2); die Mindestanzahl der zu belegenden Fächer erhöht sich demzufolge auf neun.

nen und Schüler) Deutsch, Mathematik, eine Fremdsprache oder eine Naturwissenschaft (vgl. dazu Abschnitt 4.2). Die Anzahl belegpflichtiger Fächer auf erhöhtem Anforderungsniveau variiert länderspezifisch: Acht Länder sehen zwei Fächer, fünf Länder drei Fächer und weitere drei Länder vier Fächer auf erhöhtem Anforderungsniveau vor. Bezogen auf das grundlegende Anforderungsniveau sind in der Mehrheit der Länder sieben Fächer belegpflichtig (10 Länder), vier Länder sehen acht und Sachsen sogar neun Fächer auf grundlegendem Anforderungsniveau vor – in Hamburg sind es mindestens fünf.[10] Insgesamt manifestiert sich so die bereits dargelegte Erweiterung des Fächerangebots auf Länderebene in Form umfassenderer Belegpflichten auf grundlegendem Niveau. Aus der vorgegebenen Anzahl (mindestens) zu belegender Fächer und dem (je nach Anspruchsniveau unterschiedlichen) vorgesehenen Umfang an Unterrichtsstunden pro Woche ergibt sich eine Mindestanzahl an Unterrichtsstunden pro Woche. Diese liegt zwischen 32 und 35 Stunden, wobei acht Bundesländer 34 Stunden oder mehr vorsehen, sieben Länder 33 Stunden oder weniger.

Die von den Schülerinnen und Schülern in den Fächern erbrachten Leistungen in der Qualifikationsphase werden als Kursabschlussnoten pro Schulhalbjahr mittels eines Punktsystems festgehalten und gehen als sogenannte Schulhalbjahresergebnisse in die Berechnung der Gesamtqualifikation und damit der Abiturdurchschnittsnote ein (vgl. KMK, 2018, Ziff. 9)[11]. Laut KMK-Beschluss müssen 32 bis 40 Halbjahresergebnisse eingebracht werden (sogenannte Einbringpflichten) (KMK, 2018, Ziff. 9.3.3). Innerhalb dieses breiten Möglichkeitsraums verorten sich dann entsprechend auch alle Bundesländer[12] – zwei Länder sehen 32 einzubringende Halbjahresergebnisse vor, sechs Länder sehen 40 Ergebnisse vor und die übrigen Länder bewegen sich zwischen diesen beiden Polen, wobei sich lediglich vier Länder tatsächlich auf eine Zahl festlegen, nämlich ein Land auf 35 Halbjahresergebnisse, zwei Länder auf 36 und eines auf 38 Ergebnisse. Die restlichen geben ihrerseits selbst wiederum eine mögliche Spanne an, was darin begründet liegt, dass es gleichermaßen innerhalb dieser Länder unterschiedlichste Möglichkeiten zur Fächerwahl und -kombination gibt, die wiederum zu divergierenden Einbringpflichten führen. Die größte Spannbreite an dieser Stelle sieht Hamburg vor, welches den vonseiten der KMK festgelegten Rahmen (32–40) übernimmt, also hier keine Konkretisierung vornimmt, was, wie bereits angedeutet, auf die Strukturbedingungen der dort eingeführten Profiloberstufe zurückzuführen ist (vgl. Fußnote 10).

10 Die vergleichsweise geringe Anzahl belegpflichtiger Kurse in Hamburg ist der Struktur der Hamburger Profiloberstufe mit z. T. sehr komplexen Wahlmöglichkeiten geschuldet, wobei eine geringere Anzahl gewählter Fächer auch immer mit einer höheren Wochenstundenzahl für die jeweils belegten einhergeht. Ergänzt um die Tatsache, dass man in Hamburg immer vier Fächer auf eA mit entsprechend höherem Stundenvolumen belegt, ergibt sich dort mit 34 eine Wochenstundenzahl, welche im Ländervergleich sogar im oberen Bereich anzusiedeln ist.

11 Einzelne Fächer können dabei unterschiedlich gewichtet werden (vgl. KMK, 2018, Ziff. 9); auf die unterschiedlichen Umsetzungsvarianten in den Ländern kann aufgrund der Komplexität nicht weiter eingegangen werden.

12 Je nach Bundesland gehen entweder alle Halbjahresergebnisse in einfacher Gewichtung in die Gesamtqualifikation ein oder – wie in der Mehrheit der Länder – sind unterschiedliche Gewichtungen vorgesehen (z. B. doppelte Gewichtung von Kursen auf erhöhtem Anforderungsniveau oder dem Seminar-/Projektkurs (o. ä.) oder der besonderen Lernleistung). Aufgrund der Komplexität und Vielschichtigkeit können die länderspezifischen Regelungen in diesem Beitrag nicht dargestellt werden.

4 Rahmenbedingungen der Abiturprüfung

Laut KMK-Beschluss soll die Abiturprüfung insgesamt vier oder fünf Prüfungsfächer umfassen, wobei mindestens drei schriftliche Prüfungsfächer – darunter mindestens zwei Fächer mit erhöhtem Anforderungsniveau – und mindestens ein mündliches Prüfungsfach verpflichtend sind (vgl. KMK 2018, Ziff. 8.1.1/4)[13]; die länderspezifischen Regelungen sind Tabelle 4 zu entnehmen.

Die Mehrheit der Länder sieht eine Gesamtanzahl von fünf Prüfungen vor, wobei in nahezu allen Ländern die Möglichkeit besteht, eine der fünf Prüfungsleistungen – entweder eine mündliche Prüfung (sechs Länder) oder eine schriftliche Prüfung in einem Fach auf grundlegendem Anforderungsniveau (zwei Länder) – durch eine besondere Lernleistung zu ersetzen.[14] In fünf Ländern sind vier Prüfungsfächer vorgesehen; die Möglichkeit der Einbringung einer besonderen Lernleistung als zusätzliche fünfte Prüfungskomponente ist gegeben.[15] Schleswig-Holstein sieht vier oder fünf Prüfungsfächer nach Wahl der Schülerinnen und Schüler vor (vgl. § 8.1 OAPVO), wobei die optionale fünfte Prüfungskomponente entweder eine zusätzliche mündliche Prüfung oder eine besondere Lernleistung sein kann.

In den meisten Ländern sind drei schriftliche Prüfungen abzulegen; in drei Ländern sind vier schriftliche Prüfungen vorgesehen. Die schriftlichen Prüfungsfächer werden entweder ausschließlich auf erhöhtem Anforderungsniveau geprüft (sechs Länder) oder es sind schriftliche Prüfungsleistungen auf erhöhtem und grundlegendem Anforderungsniveau abzulegen (zehn Länder), wobei i. d. R. zwei Prüfungen auf erhöhtem und eine Prüfung auf grundlegendem Anforderungsniveau erfolgen; in zwei Ländern sind jeweils zwei schriftliche Prüfungsleistungen auf erhöhtem und grundlegendem Anforderungsniveau zu erbringen. Ferner sind entweder eine (neun Länder) oder zwei (sieben Länder) reguläre mündliche Prüfungen (siehe Abschnitt 4.2) abzulegen.

In der Gesamtschau sind drei schriftliche und zwei mündliche Prüfungsleistungen die häufigste Kombination in der Abiturprüfung (sieben Länder), gefolgt von drei schriftlichen und einer mündlichen Prüfungsleistung (sechs Länder). Lediglich drei Länder kombinieren vier schriftliche mit einer mündlichen Prüfung.

13 In den Fächern Bildende Kunst, Musik und weiteren Fächern des künstlerischen Spektrums sowie im Fach Sport sind auch fachpraktische Prüfungen möglich (z. B. sportpraktische Aufgaben, praktisches Musizieren, gestaltungspraktische Aufgaben im künstlerischen Bereich, spielpraktische Gestaltungsaufgaben, o. Ä.), die durch einen schriftlichen oder mündlichen Prüfungsteil ergänzt werden (vgl. KMK 2018, Ziff. 8.1.6). Regelungen zu den fachpraktischen Prüfungsanteilen sind jedoch nicht Gegenstand dieses Beitrags.

14 Lediglich in Bayern und im Saarland besteht die Möglichkeit der Einbringung einer besonderen Lernleistung im Rahmen der Abiturprüfung nicht; hier ist jedoch eine Einbringung im Rahmen der Qualifikationsphase möglich.

15 Eine besondere Lernleistung kann z. B. ein umfassender Beitrag aus einem durch die Länder geförderten Schülerwettbewerb, eine Jahres- oder Seminararbeit oder das Ergebnis eines fachlichen oder fachübergreifenden Projekts oder Praktikums sein (vgl. KMK 2018, Ziff. 7.7). Die länderspezifischen Regelungen zur besonderen Lernleistung (z. B. Möglichkeiten der Realisierung, Voraussetzungen für die Einbringung, Einbringmöglichkeiten im Zuge der Ermittlung der Gesamtqualifikation etc.) können in diesem Beitrag nicht ausführlich dargestellt werden. Eine erste Durchsicht der entsprechenden Vorgaben weist jedoch auf eine große Bandbreite an Ausgestaltungsvarianten in den Ländern hin.

Tabelle 4: Gliederung der Abiturprüfung im bundesweiten Vergleich

	Anzahl der Abiturprüfungen insgesamt	schriftliche Abiturprüfungen			mündliche Abiturprüfungen	besondere Lernleistung
		insgesamt	davon: eA	davon: gA		
KMK	4–5	mind. 3	mind. 2	–	mind. 1	
BW	5	3	3	0	2	statt 1 (von 2) mündlichen Prüfung
BY	5	3	3	0	2	–[a]
BE	5	3	2	1	2	statt 1 (von 2) mündlichen Prüfung
BB	4	3	3	0	1	als zusätzliche 5. Prüfungskomponente
HB	4	3	2	1	1	als zusätzliche 5. Prüfungskomponente
HH	4	3	3 oder 2	0 oder 1	1	als zusätzliche 5. Prüfungskomponente
HE	5	3	2	1	2	statt 1 (von 2) mündlichen Prüfung
MV	5	3	2	1	2	statt 1 (von 2) mündlichen Prüfung
NI	5	4	3	1	1	statt schriftlicher Prüfung auf gA
NW	4	3	2	1	1	als zusätzliche 5. Prüfungskomponente
RP	4	3	3	0	1	als zusätzliche 5. Prüfungskomponente
SL	5	4	2	2	1	–[a]
SN	5	3	2	1	2	statt 1 (von 2) mündlichen Prüfung
ST	5	4	2	2	1	statt 1 (von 2) schriftlichen Prüfung auf gA
SH	4–5	3	3	0	1–2	als zusätzliche 5. Prüfungskomponente
TH	5	3	3	0	2	statt 1 (von 2) mündlichen Prüfung[b]

Anmerkungen: [a] Besondere Lernleistung kann im Rahmen der Qualifikationsphase eingebracht werden. [b] Eine mündliche Prüfung kann durch eine Seminarfachleistung (eine mögliche Variante der besonderen Lernleistung) ersetzt werden.

4.1 Allgemeine Rahmenbedingungen für die schriftliche Abiturprüfung

Aufgaben für die schriftliche Abiturprüfung werden den Vorgaben der KMK entsprechend (vgl. KMK 2018, Ziff. 8.3.1) von der Schulaufsichtsbehörde gestellt (zentrales Verfahren) oder genehmigt (dezentrales Verfahren). Mit Ausnahme von Rheinland-Pfalz werden in allen Bundesländern die schriftlichen Abiturprüfungsaufgaben (zumindest teilweise) landeszentral gestellt: In acht Bundesländern gibt es in *allen* Prüfungsfächern zentral gestellte schriftliche Abiturprüfungsaufgaben; in sieben Ländern erfolgt die Aufgabenstellung im schriftlichen Abitur lediglich in bestimmten Fächern bzw. Fächergruppen zentral, während die übrigen schriftlichen Prüfungen dezentral organisiert sind. In welchen Fächern die Aufgabenstellung dezentral erfolgt, variiert länderspezifisch – der Anteil an dezentralen schriftlichen Prüfungen ist in Schleswig-Holstein am größten (zentral gestellte Abiturprüfungsaufgaben nur in Deutsch, Mathematik und den fortgeführten Fremdsprachen, in allen anderen Fächern dezentral[16]) und in Niedersachsen am geringsten (dezentrale Aufgabenstellung in bilingual unterrichteten Sachfächern, in allen anderen Fächern zentral). Die Aufgabenstellung im schriftlichen Abitur erfolgt in den Fächern Deutsch, Mathematik, Englisch und Französisch in allen Ländern (mit Ausnahme von Rheinland-Pfalz) zentral; die landeszentralen Abiturprüfungen umfassen – je nach Bundesland und Fach in unterschiedlichem Umfang – ländergemeinsame/-übergreifende und landeseigene Prüfungsaufgaben (vgl. Beitrag 2 von Hoffmann, Schröter & Stanat in diesem Band). In Rheinland-Pfalz ist nach wie vor die gesamte Abiturprüfung dezentral organisiert – in den Fächern Deutsch, Mathematik, Englisch und Französisch werden jedoch die von den einzelnen Lehrkräften für ihren Kurs individuell gestellten schriftlichen Abiturprüfungsaufgaben um eine landesweit einheitliche Prüfungskomponente aus den ländergemeinsamen Abituraufgabenpools ergänzt (vgl. ebd.).[17]

Die von der Schulaufsichtsbehörde gestellten oder genehmigten schriftlichen Abiturprüfungsaufgaben werden um vorgegebene oder genehmigte Korrektur- und Bewertungsvorgaben[18] ergänzt (vgl. KMK 2018, Ziff. 8.3.2 und 8.4), die festlegen sollen, welche Leistungen von den Schülerinnen und Schülern in der schriftlichen Prüfung erwartet werden, welche Leistung wie zu bewerten ist und welche Leistung zu welcher Note führt. Die schriftlichen Prüfungsleistungen werden auf dieser Grundlage von mindestens zwei Instanzen eigenverantwortlich korrigiert, beurteilt und bewertet (vgl. KMK 2018, Ziff. 8.4.1): Die Erstkorrektur erfolgt immer durch die das Prüfungsfach unterrichtende Lehrkraft. Die Zweitkorrektur wird zumeist von einer weiteren Fachlehrkraft derselben Schule (13 Länder) vorgenommen. Davon sehen fünf Länder grundsätzlich auch die Möglichkeit einer Zweitkorrektur durch eine Fachlehrkraft einer anderen Schule vor (i. d. R. begründet und nach Abstimmung zwischen Schule und Schulaufsichtsbehörde); Hessen weist als einziges Bundesland zu-

16 Für einzelne Teile der Prüfung kann auch eine zentrale Aufgabenstellung vorgesehen werden – so werden z. B. in den Fremdsprachen (Profilfächer) bestimmte Prüfungsteile zentral und andere dezentral gestellt.

17 Auffällig ist, dass die Angleichprozesse im Abiturbereich durch ländergemeinsame/-übergreifende Prüfungsaufgaben nur selten erwähnt werden – nur wenige Länder weisen explizit darauf hin.

18 Der Terminus *Korrektur- und Bewertungsvorgaben* ist als Sammelbegriff für die Vielzahl an länderspezifischen Bezeichnungen dieser Vorgaben (z. B. Erwartungshorizont, Lösungserwartungen/-vorschläge, Korrekturanweisungen/-hinweise, Bewertungsvorgaben/-maßstäbe, Richtlinien für die Bewertung u. Ä.) zu verstehen.

dem explizit darauf hin, dass „das Kultusministerium zur Entwicklung und Sicherung einheitlicher Bewertungsmaßstäbe anordnen [kann], dass für alle oder einzelne Fächer landesweit oder für bestimmte Regionen die Zweitkorrektur der schriftlichen Arbeit von Lehrkräften anderer Schulen vorgenommen wird“ (OAVO § 33, 3). In Baden-Württemberg und Sachsen erfolgt die Zweitkorrektur grundsätzlich durch eine Fachlehrkraft einer anderen Schule.

In den meisten Ländern ist dem/der Zweitgutachter/in die Bewertung der/des Erstgutachters/der Erstgutachterin bekannt – in Mecklenburg-Vorpommern und im Saarland erfolgt die Zweitkorrektur ohne Kenntnis des Ergebnisses der Erstkorrektur. Im Saarland erfolgt die Zweitkorrektur zudem in anonymisierter Form, d.h. der Zweitgutachter bzw. die Zweitgutachterin kennt den Namen der zu prüfenden Schülerin bzw. des zu prüfenden Schülers nicht.

Laut KMK-Beschluss wird die endgültige Bewertung der schriftlichen Prüfungsleistungen von dem vorsitzenden Mitglied der Prüfungskommission (i. d. R. die Schulleitung) oder von der Schulaufsichtsbehörde vorgenommen, wobei optional eine weitere, dritte Fachlehrkraft zur Bewertung hinzugezogen werden kann (vgl. KMK 2018, Ziff. 8.4.1). In drei Ländern ist im Fall einer Abweichung zwischen Erst- und Zweitkorrektur um mehr als drei Notenpunkte eine Drittkorrektur durch eine Lehrkraft derselben oder einer anderen Schule obligatorisch. In den meisten Ländern entscheidet im Falle einer Abweichung zwischen Erst- und Zweitkorrektur jedoch eine dritte Instanz – je nach Land der/die Vorsitzende der Prüfungskommission, die gesamte Prüfungskommission, der/die Vorsitzende des Fachausschusses oder eine von der Schulaufsichtsbehörde beauftragte, nicht näher bestimmte Person – über die endgültige Benotung der schriftlichen Prüfungsleistung, die bei Bedarf die Einschätzung einer weiteren Fachlehrkraft einholen kann. Die Voraussetzungen für die Einbindung dieser dritten Instanz (z. B. notwendige Notenpunktedifferenz, Beurteilung der Prüfungsleistung mit 0 Punkten), der Prozess (z. B. Entscheidung nach Aktenlage, Anhörung von Erst- und Zweitgutachtern bzw. Erst- und Zweitgutachterinnen, Alleinentscheidung) sowie die Modalitäten für die Festsetzung der endgültigen Bewertung (z. B. innerhalb der Bandbreite der vorliegenden Bewertungen oder von den vorliegenden Bewertungen abweichende Bewertung) variieren zwischen den Ländern.

4.2 Fachspezifische Rahmenbedingungen für die schriftliche Abiturprüfung

Die KMK-Vereinbarung sieht vor, dass mindestens eine schriftliche Abiturprüfung in Deutsch, Mathematik, einer Fremdsprache oder einem naturwissenschaftlichen Fach auf erhöhtem Anforderungsniveau erfolgt (vgl. KMK 2018, Ziff. 8.1.4). Die Vorgaben in 11 Bundesländern sehen demzufolge mindestens eine schriftliche Abiturprüfung in den vorgenannten Fächern bzw. Fächergruppen auf erhöhtem Anforderungsniveau vor; in fünf Ländern sind sogar zwei schriftliche Prüfungsleistungen in diesen Fächern bzw. Fächergruppen auf erhöhtem Anforderungsniveau obligatorisch.

In allen Bundesländern erfolgt zumindest die Aufgabenstellung im schriftlichen Abitur in den Fächern Deutsch, Mathematik, Englisch und Französisch zentral – mit Ausnahme von Rheinland-Pfalz, wo die dezentral gestellten Prüfungsaufgaben in diesen Fächern um eine zentrale Prüfungskomponente aus den ländergemeinsamen

Abituraufgabenpools ergänzt werden – (s. o.), sodass für diese vier Fächer nachfolgend die Rahmenbedingungen für die schriftliche Abiturprüfung vertiefend in den Blick genommen werden. Zu erwähnen ist für diese Fächer, dass in allen Ländern eine schriftliche Prüfung auf erhöhtem Anforderungsniveau abgelegt werden kann; schriftliche Prüfungen auf grundlegendem Anforderungsniveau sind in den Fächern Deutsch und Mathematik nur in elf Ländern, in Englisch und Französisch nur in zehn Ländern möglich.

4.2.1 Schriftliche Abiturprüfung im Fach Deutsch

Das Fach Deutsch wird im der Abiturprüfung vorgelagerten Unterricht gemäß KMK-Vereinbarung auf erhöhtem Anforderungsniveau vier- oder fünfstündig und auf grundlegendem Anforderungsniveau drei- oder vierstündig unterrichtet (vgl. KMK 2018, Ziff. 7.2); in Bayern und Schleswig-Holstein wird im Fach Deutsch nur Unterricht auf erhöhtem Anforderungsniveau erteilt. In 13 Ländern umfasst der Deutschunterricht auf erhöhtem Anforderungsniveau fünf Unterrichtsstunden pro Woche, in drei Bundesländern vier Unterrichtsstunden pro Woche. Das Fach Deutsch wird auf grundlegendem Anforderungsniveau in den meisten Ländern dreistündig unterrichtet, in vier Bundesländern hingegen vierstündig. Unter der Annahme von etwa 38 Unterrichtswochen pro Schuljahr[19] sind die unterrichtlichen Voraussetzungen für die Abiturprüfung in Abhängigkeit der vorgesehenen Unterrichtsstunden pro Woche unterschiedlich – eine Unterrichtsstunde pro Woche mehr ergibt in der zweijährigen Qualifikationsphase rein rechnerisch eine Differenz von etwa 76 Unterrichtsstunden bis zur Abiturprüfung.

Die Mehrheit der Länder benennt im Vorfeld Schwerpunktthemen (elf Länder), also bestimmte Themen/Bereiche des Lehrplans, und/oder gibt Pflichtlektüren vor (zehn Länder), die potenziell Gegenstand der Abiturprüfung sein können.

In der schriftlichen Abiturprüfung stehen den Schülerinnen und Schülern im Fach Deutsch in allen Ländern mehrere Aufgabenvorschläge zur Wahl; Pflichtaufgaben, die von allen Prüflingen in der Abiturprüfung zu bearbeiten sind, gibt es keine. Die Spannweite der zur Auswahl stehenden Aufgaben liegt zwischen drei und vier (grundlegendes Anforderungsniveau) bzw. fünf Aufgaben (erhöhtes Anforderungsniveau), wobei die Mehrheit der Länder für beide Niveaustufen drei Aufgaben zur Auswahl stellt, von denen dann eine Aufgabe bearbeitet werden muss. Für beide Niveaustufen sieht die überwiegende Mehrheit der Länder eine alleinige Auswahl für die Schülerinnen und Schüler vor; lediglich zwei Bundesländer sehen eine Aufgabenauswahl zunächst durch die Lehrkräfte und anschließend durch die Prüflinge vor.

Die KMK-Vereinbarung sieht für das Prüfungsfach Deutsch eine bundesweit einheitliche Bearbeitungszeit von 270 Minuten (erhöhtes Anforderungsniveau) bzw. 210 Minuten (grundlegendes Anforderungsniveau) vor (vgl. KMK 2018, Ziff. 8.3.3). Zusätzlich können die Länder eine Auswahlzeit bis zu 45 Minuten vorsehen (ebd.) –

19 Die Angabe von 38 Unterrichtswochen pro Schuljahr ist lediglich als Näherungswert zu sehen: In dem regelmäßig erscheinenden Berichtsband *Education at a Glance* (zuletzt OECD, 2020) werden für Deutschland durchschnittlich 37,6 Unterrichtswochen pro Schuljahr angegeben; je nach Bundesland dürfte die Zahl der Unterrichtsstunden jedoch variieren (z. B. aufgrund gesetzlicher Feiertage in den Ländern, unterschiedlicher Abiturprüfungstermine).

die Mehrheit der Länder gewährt auch diese maximal vorgesehene Auswahlzeit, vereinzelt sind nur 30 Minuten vorgesehen.

Zugelassene Hilfsmittel sind anzugeben (vgl. KMK 2012a, S. 23); weitergehende bundesweite Vorgaben zu Hilfsmitteln gibt es keine. Nahezu alle Bundesländer gestatten die Verwendung eines Wörterbuches der deutschen Sprache. Sofern eine Prüfungsaufgabe eine Auseinandersetzung mit (umfangreichen) literarischen Texten umfasst (z. B. Aufgabenstellungen im Zusammenhang mit vorgegebenen Pflichtlektüren), sind die vollständigen Textausgaben als Hilfsmittel zugelassen, und zwar entweder eine unkommentierte Textausgabe (sieben Länder), eine Textausgabe mit Anmerkungen und Markierungen der Schülerinnen und Schüler (zwei Länder) oder eine Textausgabe mit Worterläuterungen des Verlages (ein Land).

4.2.2 Schriftliche Abiturprüfung im Fach Mathematik

In der Qualifikationsphase wird das Fach Mathematik gemäß KMK-Vereinbarung auf erhöhtem Anforderungsniveau vier- oder fünfstündig und auf grundlegendem Anforderungsniveau drei- oder vierstündig unterrichtet (vgl. KMK 2018, Ziff. 7.2); in Bayern und Schleswig-Holstein wird auch in Mathematik nur Unterricht auf erhöhtem Anforderungsniveau erteilt. In 13 Bundesländern umfasst der Mathematikunterricht auf erhöhtem Anforderungsniveau fünf Unterrichtsstunden pro Woche, in drei Ländern vier Unterrichtsstunden pro Woche. Mathematik wird auf grundlegendem Anforderungsniveau in neun Ländern dreistündig unterrichtet, in fünf Bundesländern hingegen vierstündig. Der unter Abschnitt 4.2.1 bereits erläuterten Annahme von etwa 38 Unterrichtswochen pro Schuljahr[20] folgend, ergibt sich auch für das Fach Mathematik eine entsprechende zeitliche Diskrepanz zwischen den Ländern bezüglich der unterrichtlichen Voraussetzungen für die Abiturprüfung (siehe Abschnitt 4.2.1).

In zwölf Ländern werden im Vorfeld mögliche Schwerpunkte benannt, die potenziell Gegenstand der Abiturprüfung sein können, bzw. es werden explizit bestimmte Bereiche ausgeschlossen.

Knapp die Hälfte der Länder stellt in der schriftlichen Abiturprüfung auf erhöhtem Anforderungsniveau ausschließlich Aufgaben, die verpflichtend von allen Schülerinnen und Schülern zu bearbeiten sind (sieben Länder), d. h. es bestehen keine Aufgabenauswahlmöglichkeiten. Drei dieser Bundesländer legen jedoch den Lehrkräften mehrere Aufgabenvorschläge vor, aus denen diese dann für ihre Schülerinnen und Schüler einen zur Bearbeitung auswählen. In neun Ländern gibt es eine Kombination aus Pflicht- und Wahlaufgaben, d. h. bestimmte Aufgabenteile sind von allen Schülerinnen und Schülern zu bearbeiten und zusätzlich werden zwei (acht Länder) oder drei (ein Land) Aufgaben zur Auswahl gestellt, von denen dann eine bearbeitet werden muss. In den Ländern, in denen die schriftliche Abiturprüfung auch auf grundlegendem Anforderungsniveau erfolgt, sind entweder ausschließlich Pflichtaufgaben (ohne Wahlmöglichkeit) zu bearbeiten (vier Länder) oder es ist eine Kombina-

20 Die Angabe von 38 Unterrichtswochen pro Schuljahr ist lediglich als Näherungswert zu sehen: In dem regelmäßig erscheinenden Berichtsband *Education at a Glance* (zuletzt OECD, 2020) werden für Deutschland durchschnittlich 37,6 Unterrichtswochen pro Schuljahr angegeben; je nach Bundesland dürfte die Zahl der Unterrichtsstunden jedoch variieren (z. B. aufgrund gesetzlicher Feiertage in den Ländern, unterschiedlicher Abiturprüfungstermine).

tion aus Pflicht- und Wahlaufgaben vorgesehen (sieben Länder), wobei Schülerinnen und Schüler im Bereich der Wahlaufgaben aus zwei (fünf Länder) oder drei (zwei Länder) Aufgaben auswählen können.

Die KMK-Vereinbarung sieht für das Prüfungsfach Mathematik eine bundesweit einheitliche Bearbeitungszeit von 270 Minuten (erhöhtes Anforderungsniveau) bzw. 225 Minuten (grundlegendes Anforderungsniveau) vor (vgl. KMK 2018, Ziff. 8.3.3). Zusätzlich können die Länder eine Auswahlzeit bis zu 30 Minuten vorsehen (ebd.).

Zugelassene Hilfsmittel sind anzugeben (vgl. KMK 2012b, S. 24); weitergehende bundesweite Vorgaben zu Hilfsmitteln gibt es keine. Nahezu alle Bundesländer gestatten die Verwendung eines Wörterbuches der deutschen Sprache. Als weitere Hilfsmittel sind in zwölf Ländern (Tabellenwerk und) Formelsammlungen zugelassen – dazu werden teils keine weiteren, teils sehr konkrete Vorgaben gemacht (z. B. Auflistung zugelassener Formelsammlungen); in sieben Ländern werden zudem explizit bestimmte Zeichengeräte (z. B. Zirkel, Kurvenschablonen) als Hilfsmittel ausgewiesen. Zugelassen sind je nach Land zudem wissenschaftliche Taschenrechner (WTR), Taschenrechner mit Computer-Algebra-System (CAS) und/oder grafikfähige Taschenrechner (GTR), mit denen jeweils unterschiedliche Aufgabenstellungen in der schriftlichen Prüfung einhergehen.[21]

4.2.3 Schriftliche Abiturprüfung in den fortgeführten Fremdsprachen (Englisch/Französisch)

Englisch und Französisch werden als fortgeführte Fremdsprachen im der Abiturprüfung vorgelagerten Unterricht gemäß KMK-Vereinbarung auf erhöhtem Anforderungsniveau vier- oder fünfstündig und auf grundlegendem Anforderungsniveau drei- oder vierstündig unterrichtet (vgl. KMK 2018, Ziff. 7.2); in Bayern wird Unterricht in den fortgeführten Fremdsprachen ausschließlich auf erhöhtem Anforderungsniveau erteilt. In zwölf Ländern umfasst der Unterricht in den fortgeführten Fremdsprachen Englisch und Französisch auf erhöhtem Anforderungsniveau fünf Unterrichtsstunden pro Woche, in vier Bundesländern vier Unterrichtsstunden pro Woche. Die Fächer werden auf grundlegendem Anforderungsniveau in den meisten Ländern dreistündig unterrichtet, in zwei Bundesländern auch vierstündig. Folglich besteht zwischen den Ländern auch im Bereich der fortgeführten Fremdsprachen Englisch und Französisch eine zeitliche Diskrepanz bezüglich der unterrichtlichen Voraussetzungen für die Abiturprüfung (siehe Abschnitt 4.2.1).

Neun Länder benennen im Vorfeld Schwerpunktthemen und/oder geben Pflichtlektüren vor (fünf Länder), die potenziell Gegenstand der Abiturprüfung sein können.

Die schriftliche Abiturprüfung in den fortgeführten Fremdsprachen Englisch und Französisch besteht in nahezu allen Ländern aus einer Kombination von Pflicht- und Wahlaufgaben, d. h. bestimmte Aufgabenteile sind von allen Schülerinnen und Schülern zu bearbeiten und zusätzlich werden den Prüflingen zwei Aufgaben zur Auswahl gestellt, von denen dann eine bearbeitet werden muss (grundlegendes und

21 Die Vorgaben und Modalitäten zur Verwendung der unterschiedlichen Taschenrechnerarten sind so unterschiedlich, dass sie in diesem Beitrag nicht differenziert dargestellt werden können.

erhöhtes Anforderungsniveau). In Rheinland-Pfalz und Sachsen gibt es keine Auswahlmöglichkeiten, sondern alle Aufgaben sind verpflichtend zu bearbeiten.

In den fortgeführten Fremdsprachen gibt es einen verpflichtenden Prüfungsteil *Schreiben* (vgl. Bildungsstandards für die fortgeführte Fremdsprache (Englisch/Französisch) für die Allgemeine Hochschulreife, S. 25), für den gemäß KMK-Vereinbarung eine bundesweit einheitliche Bearbeitungszeit von 210 Minuten (erhöhtes Anforderungsniveau) bzw. 180 Minuten (grundlegendes Anforderungsniveau) vorgesehen ist (vgl. KMK 2018, Ziff. 8.3.3). Daneben gibt es einen weiteren, ebenfalls verpflichtenden Prüfungsteil, der grundsätzlich aus zwei Aufgaben zu unterschiedlichen Kompetenzbereichen besteht (vgl. KMK 2012c, Ziff. 3.2.1.1) und für die unterschiedliche Bearbeitungszeiten festgelegt sind (vgl. KMK 2018, Ziff. 8.3.3). Insofern unterscheidet sich die in den Ländern vorgesehene Bearbeitungszeit in Abhängigkeit von den ausgewählten Kompetenzbereichen für den zweiten Prüfungsteil neben der Schreibaufgabe. Zusätzlich können die Länder eine Auswahlzeit bis zu 30 Minuten vorsehen (ebd.), die im Falle bestehender Wahlmöglichkeiten auch gewährt werden.

Als Hilfsmittel sind den bundesweiten Vorgaben entsprechend (vgl. KMK 2012c, S. 25) ein- und zweisprachige Wörterbücher zugelassen. Zudem gestatten fast alle Bundesländer die Verwendung eines Wörterbuches der deutschen Sprache. Sofern eine Prüfungsaufgabe eine Auseinandersetzung mit den Pflichtlektüren umfasst, sind die vollständigen Textausgaben als Hilfsmittel zugelassen, und zwar entweder als Textausgabe mit Markierungen der Schülerinnen und Schüler (ein Land) oder eine unkommentierte Textausgabe bzw. eine Textausgabe mit Worterläuterungen des Verlages (ein Land).

4.3 Mündliche Abiturprüfungen

Die KMK-Vereinbarung sieht mindestens eine verpflichtende mündliche Prüfung in einem Fach vor, das nicht schon schriftlich geprüft wurde (vgl. KMK 2018, Ziff. 8.1.1 und 8.1.5) – wie einleitend dargestellt, ist in neun Ländern eine mündliche Prüfung für alle Schülerinnen und Schüler obligatorisch, in sieben Ländern sind zwei mündliche Prüfungen der Regelfall (siehe Abschnitt 4). Die Rahmenbedingungen für diese für alle Prüflinge verpflichtenden mündlichen Prüfungen werden nachfolgend dargestellt;[22] die länderspezifischen Regelungen sind Tabelle 5 zu entnehmen.

Gemäß KMK-Vereinbarung wird eine mündliche Abiturprüfung in der Regel als Einzelprüfung durchgeführt (ebd., Ziff. 8.5.1), die üblicherweise wie folgt abläuft: Die Prüflinge erhalten eine Aufgabe zur Vorbereitung und stellen – nach einer Vorbereitungszeit – der Prüfungskommission ihre Überlegungen zur vorgelegten Aufgabe in

22 Neben den regulären mündlichen Abiturprüfungen können zusätzlich auch in schriftlich geprüften Fächern mündliche Prüfungen angesetzt werden (vgl. KMK 2018, Ziff. 8.1.4). Zu unterscheiden sind – je nach Bundesland – zusätzliche mündliche Prüfungen auf Antrag der Schülerinnen und Schüler (i. d. R. verbunden mit dem Ziel einer Notenverbesserung), verpflichtende mündliche Zusatzprüfungen (z. B. wenn eine schriftliche Prüfungsleistung mit 0 Punkten bewertet wurde oder bei erheblichen Abweichungen der Prüfungsnote von der Vornote) sowie mündliche Prüfungen zum Erreichen der Gesamtqualifikation (erforderliche Mindestpunktzahl) und damit zum Bestehen der Abiturprüfung. Für diese mündlichen Prüfungsformen sind von den für die regulären mündlichen Prüfungen geltenden Vorgaben abweichende Regelungen möglich (vgl. KMK 2018, Ziff. 8.5.3). Regelungen zu diesen besonderen mündlichen Prüfungen sind nicht Gegenstand dieses Beitrags – die Durchsicht der entsprechenden Vorgaben weist jedoch auf eine große Bandbreite an Gestaltungsvarianten hin.

einem Kurzvortrag vor, woran sich ein Prüfungsgespräch zum Vortrag und weiteren Lerninhalten anschließt. Vorbereitungszeit und Prüfung sollen entsprechend der KMK-Vereinbarung in der Regel jeweils 20 Minuten umfassen (ebd., Ziff. 8.5.1 und 8.5.2). Neun Bundesländer sehen jeweils 20 Minuten Vorbereitungszeit vor, sechs Länder hingegen 30 Minuten und in Hessen wird ein Zeitkorridor von 20–30 Minuten angegeben. Auch die Prüfungsdauer variiert: In sieben Ländern dauern die Prüfungen 20 Minuten, in drei Ländern 30 Minuten, fünf Länder geben einen zeitlichen Korridor von 20–30 Minuten und ein Land von 20–25 Minuten vor. Gruppenprüfungen sind laut KMK-Vereinbarung grundsätzlich zulässig (ebd., Ziff. 8.5.1); dabei ist die Bewertung der individuellen Prüfungsleistung sicherzustellen. Diese Prüfungsform – mit bis zu drei Prüflingen – ist jedoch nur in 4 Bundesländern möglich; die Prüfungsdauer erhöht sich in der Folge (je nach Bundesland) auf bis zu 70 Minuten.

Ferner können die Länder „besondere mündliche Prüfungsformen" (ebd., Ziff. 8.5.3) durchführen. In fünf Bundesländern gibt es sogenannte Präsentationsprüfungen als Variante der mündlichen Prüfung. Schülerinnen und Schüler erhalten dazu – je nach Länderregelung – einige Zeit vor dem Prüfungstermin eine Aufgabenstellung bzw. ein Thema (ggf. können sie selbst Vorschläge machen) und fertigen dazu eine schriftliche Dokumentation an, die i. d. R. im Vorfeld bei der prüfenden Lehrkraft einzureichen ist. Die Prüfung selbst umfasst zumeist einen mediengestützten Vortrag, an den sich ein Prüfungsgespräch anschließt. In der Ausgestaltung dieser Variante der mündlichen Prüfung zeigen sich Länderunterschiede etwa hinsichtlich der Vorbereitungszeit, der Ausgestaltung der Dokumentation sowie der zeitlichen Gliederung der Prüfung (z. B. Länge des mediengestützten Vortrags). In vier Ländern ist eine Prüfungsdauer von 30 Minuten vorgesehen, Niedersachsen gibt einen zeitlichen Korridor von 30–45 Minuten vor. Die Präsentationsprüfung ist i. d. R. eine Einzelprüfung, in drei Ländern ist auch eine Gruppenpräsentationsprüfung möglich.

In Hamburg und Niedersachsen ist eine mündliche Prüfung im Abitur vorgesehen – die Prüflinge können das Format dieser Prüfung (klassische mündliche Prüfung oder Präsentationsprüfung) wählen. In Berlin und Hessen findet im vierten Abiturprüfungsfach eine klassische mündliche Prüfung statt – als fünfte Prüfungskomponente ist in Berlin eine Präsentationsprüfung vorgesehen, die durch eine besondere Lernleistung (s. o.) ersetzt werden kann, und in Hessen nach Wahl der Schülerinnen und Schüler entweder eine klassische mündliche Prüfung oder eine Präsentationsprüfung oder eine besondere Lernleistung. In Schleswig-Holstein ist eine mündliche Prüfung (klassische mündliche Prüfung oder Präsentationsprüfung) vorgesehen; das Einbringen einer fünften Prüfungskomponente ist freiwillig, wobei diese nach Wahl der Schülerinnen und Schüler entweder eine zusätzliche klassische mündliche Prüfung oder eine besondere Lernleistung sein kann.

Tabelle 5: Mündliche Abiturprüfungen im bundesweiten Vergleich

	Anzahl	mündliche Prüfung			Präsentationsprüfung[a]		Prüfungsformat nach Wahl der SuS
		Vorbereitungszeit	Dauer der Prüfung	Gruppenprüfung möglich	Dauer der Prüfung	Gruppenprüfung möglich	
KMK	mind. 1	20	20	x			
BW	2[b]	20	20	x			
BY	2	30	30	–			
BE	2[b]	20	20	–	30	x	–
BB	1	30	20	–			
HB	1	20	20–25	–			
HH	1	30	30	x	30	x	x
HE	2[b]	20–30	20	x	30	x	x
MV	2[b]	20	20	–			
NI	1	20	20–30	x	30–45	–	x
NW	1	30	20–30	–			
RP	1	20	20	–			
SL	1	30	20–30	–			
SN	2[b]	20	30	–			
ST	1	20	20–30	–			
SH	1–2	30	20	–	30	–	x
TH	2[b]	20	20–30	–			

Anmerkungen: [a] Die KMK spricht von „besonderen mündlichen Prüfungsformen". [b] Eine der beiden verpflichtenden Prüfungen kann durch eine besondere Lernleistung ersetzt werden.

5 Fazit

Vor dem Hintergrund der in jüngerer Zeit getroffenen Maßnahmen zur Herstellung von mehr Vergleichbarkeit in der gymnasialen Oberstufe und der Abiturprüfung zwischen den Bundesländern wurden in dem vorliegenden Beitrag die Strukturen der gymnasialen Oberstufen und die formal-organisatorische Ausgestaltung der Abiturprüfungsverfahren in allen deutschen Ländern kategoriengeleitet analysiert. Der Beitrag schließt nun mit der Frage, wie vergleichbar diese Strukturen und Rahmenbedingungen sind, wobei der Begriff der Vergleichbarkeit aus unterschiedlichen Perspektiven betrachtet wird. Die Darstellung der Limitationen des Ländervergleichs sowie ein weiterführender Ausblick runden den Beitrag ab.

5.1 Gleichartig oder gleichwertig? – Wie vergleichbar sind die Strukturen der gymnasialen Oberstufen und die Rahmenbedingungen der Abiturprüfungsverfahren in Deutschland?

Die Beantwortung der Frage, wie vergleichbar die Strukturen und Rahmenbedingungen zwischen den Ländern sind, ist nicht trivial – insbesondere je nachdem, welches Verständnis von *vergleichbar* zugrunde gelegt wird. Die Betrachtung häufig genannter Synonyme – *ähnlich, entsprechend, gleichartig, identisch, ohne Unterschied, übereinstimmend* (Dudenredaktion, o.J.) – legt eine Lesart von *vergleichbar* im Sinne von *gleich* nahe, was „in allen oder wesentlichen Merkmalen übereinstimmend" bedeutet (Etymologisches Wörterbuch des Deutschen, o.J.). Vor diesem Hintergrund wird nachfolgend auf Basis der vorangegangenen Analysen zusammenfassend dargelegt, in welchen Merkmalen die gymnasialen Oberstufen sowie die Abiturprüfungsverfahren der 16 Länder übereinstimmen – und in welchen nicht. Bestehende Unterschiede lassen folglich eine weitere Lesart von *vergleichbar* im Sinne von *gleichwertig* zu, wonach bestimmte Merkmale der gymnasialen Oberstufe und der Abiturprüfung einzelner Bundesländer zwar andersartig, aber vor dem Hintergrund der bestehenden KMK-Vereinbarung zur gegenseitigen Anerkennung der Zeugnisse der Allgemeinen Hochschulreife (vgl. z. B. aktuell KMK, 2021, Ziff. 13.1) dennoch offenbar „von gleichem Wert" (Etymologisches Wörterbuch des Deutschen, o. J.) sind. Insofern wird neben der Frage nach der *Gleichartigkeit* auch der Aspekt der *Gleichwertigkeit* aufgegriffen, um Anstöße für zukünftige Diskussionen in (Bildungs-)Politik, Wissenschaft und Öffentlichkeit zu geben.

Die *Struktur der gymnasialen Oberstufe* umfasst in allen Ländern eine zweijährige Qualifikationsphase, in der Unterricht auf zwei verschiedenen Anspruchsebenen – auf grundlegendem und erhöhtem Anforderungsniveau – erteilt wird. Die damit verbundene Anzahl an wöchentlichen Unterrichtsstunden variiert jedoch länderspezifisch, d. h. hinter formal gleichen Kursniveaus verbergen sich unterschiedliche Unterrichtsvolumen, die sich im Verlauf der zweijährigen Qualifikationsphase entsprechend aufsummieren. Insgesamt erscheinen die Unterschiede in der fachspezifischen Wochenstundenzahl auf beiden Anspruchsebenen größer als die Unterschiede in der Gesamtheit der Wochenpflichtstunden über alle Fächer hinweg.

Die unterrichteten Fächer sind überwiegend drei Aufgabenfeldern zugeordnet, innerhalb derer ein Grundkanon von 13 Fächern identifiziert werden konnte, die in allen Bundesländern unterrichtet werden – dieser wird gleichwohl in allen Ländern um weitere, teilweise mehr als 20 zusätzliche Unterrichtsfächer erweitert, sodass im Sinne der vereinbarten Möglichkeit einer individuellen Schwerpunktsetzung für Schülerinnen und Schüler eine große Bandbreite an Fächerwahlmöglichkeiten besteht, die jedoch länderspezifisch variiert.

In der Qualifikationsphase werden ferner Pflicht- und Wahlfächer unterschieden, wobei in *allen* Ländern *alle* Schülerinnen und Schüler in der Qualifikationsphase die Fächer Deutsch und Mathematik – sowie das Fach Sport, was jedoch keinem Aufgabenfeld zugeordnet ist – belegen *müssen*; daneben gibt es zwar weitere, für alle belegpflichtige Fächergruppen, jedoch immer verbunden mit Wahlmöglichkeiten für

die Schülerinnen und Schüler (z. B. *eine* Naturwissenschaft oder *eine* Fremdsprache) – allerdings variieren die Mindestanzahl belegpflichtiger Fächer sowie die damit verbundenen Anspruchsebenen länderspezifisch. Wenngleich Deutsch und Mathematik Pflichtfächer sind, unterscheiden sich die Ländervorgaben dahin gehend, ob der Fachunterricht auf grundlegendem und (oder ausschließlich) auf erhöhtem Anforderungsniveau erteilt wird und welcher Umfang an Wochenstunden mit den beiden unterschiedlichen Anspruchniveaus einhergeht. Ein Fach, das in allen Ländern von allen Schülerinnen und Schülern auf gleichem Anspruchsniveau belegpflichtig ist und mit gleicher Wochenstundenzahl unterrichtet wird, gibt es nicht.

Auch mit Blick auf die *Rahmenbedingungen der Abiturprüfung* weist der Ländervergleich mehr Unterschiede als Gemeinsamkeiten auf: Die Abiturprüfungsvorgaben unterscheiden sich hinsichtlich der Anzahl verpflichtend abzulegender schriftlicher und mündlicher Prüfungen und den damit verbundenen formalen Anforderungsniveaus der Prüfungen. Schriftliche Abiturprüfungsaufgaben werden je nach Bundesland entweder vollständig zentral, vollständig dezentral oder (fachabhängig) entweder zentral oder dezentral gestellt; folglich variiert auch der Rahmen für die Korrektur und Bewertung der Prüfungsleistungen, ebenso die im schriftlichen Abitur vorgesehenen Korrekturinstanzen/-verfahren – Konsens besteht lediglich darin, dass die Erstkorrektur der in der schriftlichen Abiturprüfung gezeigten Leistungen durch die zuständige Kurslehrkraft erfolgt. Auch mit Blick auf die mündlichen Abiturprüfungen finden sich vielfältige, länderspezifische Ausgestaltungsvarianten (klassische mündliche Prüfungen vs. Präsentationsprüfungen, Einzel- vs. Gruppenprüfungen) und formale Vorgaben (z. B. Wahlmöglichkeiten hinsichtlich des Prüfungsformats, Prüfungsdauer).

Alle Länder sehen vor, dass unter den Abiturprüfungsfächern zwei der drei Fächer Deutsch, Fremdsprache oder Mathematik sein müssen, d. h. von den beiden von allen Schülerinnen und Schülern in der Qualifikationsphase verpflichtend zu belegenden Fächern Deutsch und Mathematik (s. o.) wird in jedem Fall eines im Rahmen der Abiturprüfung verpflichtend geprüft, wobei die Abiturprüfung auf unterschiedlichen Anforderungsniveaus und in unterschiedlichen Prüfungsformaten (schriftlich/mündlich) abgelegt werden kann. Gerade mit Blick auf diese zentralen Fächer unterscheiden sich die Ländervorgaben erheblich – während einige Länder maximale Wahlmöglichkeiten für die Schülerinnen und Schüler zulassen, machen andere Länder enge Vorgaben wie etwa für alle verpflichtende schriftliche Abiturprüfungen in (mindestens) Deutsch und Mathematik auf grundsätzlich erhöhtem Anforderungsniveau.

Bezogen auf die in diesem Beitrag besonders fokussierten Fächer Deutsch, Mathematik, Englisch und Französisch können hinsichtlich zentraler Merkmale der schriftlichen Abiturprüfung bundesweite Gemeinsamkeiten festgestellt werden: Die Aufgabenstellung erfolgt in allen Ländern (mit Ausnahme von Rheinland-Pfalz) landeszentral und umfasst mindestens eine Prüfungsaufgabe aus dem ländergemeinsamen Abituraufgabenpool (vgl. Beitrag 2 von Hoffmann, Schröter & Stanat in diesem Band); landeseigene sowie länderübergreifend entwickelte Prüfungsaufgaben sind je-

doch ebenfalls zulässig. Die vorgesehene Bearbeitungszeit ist bundesweit einheitlich festgelegt – zumindest in diesem Punkt ist im Vergleich mit früheren Ländervorgaben (vgl. Kühn, 2010, S. 79 ff.) eine sichtbare Vereinheitlichung erfolgt, was vermutlich auf die Annäherungsprozesse der Länder im Kontext der Einführung der Abituraufgabenpools zurückgeführt werden kann. Trotz dieser Gemeinsamkeiten bestehen auch in diesen Fächern weiterhin Unterschiede, insbesondere hinsichtlich des zeitlichen Umfangs des der Prüfung vorgelagerten Unterrichts (s. o.), der Eingrenzung möglicher Prüfungsgegenstände, vorgesehener Pflichtaufgaben und Möglichkeiten der Aufgabenauswahl sowie zugelassener Hilfsmittel.

In der Gesamtschau ist festzuhalten, dass die einzelnen Strukturentscheidungen zur Organisation der Oberstufe und zu den Abiturprüfungsverfahren in den Ländern zwar durchaus Gemeinsamkeiten und Ähnlichkeiten aufweisen, tatsächlich identisch sind die Vorgaben zwischen den Ländern jedoch nur an sehr wenigen Stellen – auch lassen sich keine klaren Ländertypen identifizieren, in denen sich über mehrere Struktur- und Prüfungselemente hinweg einheitliche Gestaltungsvarianten finden. Das bedeutet, dass die Abiturdurchschnittsnote insgesamt immer noch auf Basis heterogener Strukturvoraussetzungen zustande kommt, gleichwohl Homogenisierungsbemühungen seitens der Länder nicht von der Hand zu weisen sind.

Legt man zur Beantwortung der Frage nach der Vergleichbarkeit der organisatorischen Ausgestaltung der gymnasialen Oberstufe und der Abiturprüfung in Deutschland ein Verständnis von *vergleichbar* im Sinne von *gleich* nahe, so ist diese kaum gegeben. Die weitgehende Andersartigkeit der Strukturen und Rahmenbedingungen führt vielmehr zu der Frage, inwieweit diese über alle Länder hinweg *gleichwertig*, d. h. „von gleichem Wert" sind. Für nahezu alle in diesem Beitrag betrachteten Analysebereiche lässt sich diese Frage stellen: Betrachtet man beispielsweise die Breite des Fächerangebots in der Qualifikationsphase (vgl. Tabelle 1), so ist die Frage nach der Gleichwertigkeit der belegbaren Fächer mit Blick auf die Zuerkennung der Allgemeinen Hochschulreife am Ende der Qualifikationsphase – sowohl innerhalb der Bundesländer als auch zwischen ihnen – nicht von der Hand zu weisen, insbesondere auch mit Blick auf die Vermittlung vertiefter *Allgemein*bildung als wesentliches Ziel des Unterrichts in der gymnasialen Oberstufe.

Die KMK-Vereinbarung lässt jedenfalls den Schluss zu, dass die länderspezifischen Strukturen und Rahmenbedingungen als gleichwertig angesehen werden, da die Zeugnisse der Allgemeinen Hochschulreife gegenseitig anerkannt werden. In der Vereinbarung gibt es erkennbare Ansätze einer bundesweiten Standardisierung der Strukturen und Rahmenbedingungen – gleichzeitig wird den Ländern jedoch nahezu durchgängig ein nicht unerheblicher Spielraum, sei es direkt (etwa „Die nähere Ausgestaltung obliegt den Ländern.") oder indirekt (durch Formulierungen wie „in der Regel", „oder", „können", „bis zu", etc.), zugestanden und damit eine Heterogenität der Gestaltungsvarianten zugelassen. Insofern ist die KMK-Vereinbarung lediglich als Rahmen anzusehen, der aufgrund des kulturell stark verankerten (Bildungs-)Föderalismus von den Ländern ganz unterschiedlich ausgestaltet werden kann und auch ausgestaltet wird (zum Spannungsfeld zwischen dem Bestreben nach Standardisie-

rung und der Kulturhoheit der Länder vgl. den Beitrag 6 von Groß & Schmid-Kühn in diesem Band). In der öffentlich geführten Vergleichbarkeitsdebatte wird diese Gleichwertigkeit jedoch fortwährend hinterfragt, nicht zuletzt nach einem Urteil des Bundesverfassungsgerichts (vgl. BVerfG, 1BvL 3/14 vom 19.12.2017), das nicht nur „Vergleichbarkeitsmängel" (ebd., Absatz 184) sondern gar ein „Vergleichbarkeitsdefizit" (ebd.) attestiert und von einer „unzureichenden bundesweiten Vergleichbarkeit der Abiturnoten" (ebd., Absatz 239) spricht, die aus „länderspezifisch unterschiedlichen Bildungs- und insbesondere auch Bewertungssystemen" (ebd., Absatz 182) resultiert. Wie Vergleichbarkeit im Sinne von Gleichwertigkeit jenseits von Gleichartigkeit hergestellt werden kann, ist insofern noch eine weitgehend offene Frage. Daraus ergibt sich nicht zuletzt die Notwendigkeit der begrifflichen Präzisierung von *Vergleichbarkeit* – das Begriffsverständnis scheint vage; der Begriff wird in der Vergleichbarkeitsdebatte offenkundig weder einheitlich noch trennscharf verwendet, was die Kommunikation zwischen den Akteuren erheblich erschwert. Die Herausarbeitung der genauen Bedeutung des Begriffs bzw. des Verständnisses (bzw. variierender Verständnisse) von Vergleichbarkeit und relevanter Dimensionen dieses Konzepts im Kontext der gymnasialen Oberstufe und der Abiturprüfung stellt jedoch bislang noch ein Forschungsdesiderat dar.

5.2 Limitationen des Ländervergleichs und Ausblick

Der vorliegende Beitrag bietet einen aktualisierten Gesamtüberblick über die Strukturen der gymnasialen Oberstufen und die Rahmenbedingungen der Abiturprüfungsverfahren in Deutschland und stellt darüber hinaus auch eine Erweiterung zu den einschlägigen themenbezogenen Veröffentlichungen dar, da bestimmte Aspekte (z. B. Fächerkanon der gymnasialen Oberstufe, mündliche Abiturprüfungen) bisher nicht thematisiert wurden. Die Befunde basieren auf der Analyse institutionell verantworteter Dokumente (siehe Abschnitt 2); die länderspezifischen Vorgaben sind sehr komplex und teilweise nur schwer verständlich[23], sodass die Analysen aufwändig und sehr ressourcenintensiv waren. Mit der durchgeführten Dokumentenanalyse gehen verschiedene Limitationen einher: Die Analysen beziehen sich auf die Vorgaben für den Abiturjahrgang 2021 – pandemiebedingte Ausnahmeregelungen wurden im Rahmen dieses Beitrags nicht thematisiert. Zum Zeitpunkt der Veröffentlichung des vorliegenden Herausgeberbandes sind vermutlich einige Angaben schon nicht mehr aktuell; so gibt es mittlerweile eine neue KMK-Vereinbarung zur Gestaltung der gymnasialen Oberstufe und der Abiturprüfung (KMK, 2021), in der zwar wesentliche Elemente unverändert sind, aber weitere Annäherungsprozesse sichtbar werden (z. B. zeitliche Vorgaben für die schriftlichen Abiturprüfungen, u. a. gibt es nun auch einheitlich festgeschriebene Bearbeitungszeiten in den Naturwissenschaften, vgl. ebd., Ziff. 8.3.3). Auch in einzelnen Ländern sind Änderungen bereits umgesetzt oder zumindest beschlossen (z. B. in Bayern). Limitationen ergeben sich folglich dadurch, dass die Thematik durch eine Dynamik in allen Bereichen gekennzeichnet ist. Ferner wurden nur

23 Hierzu sei weiterführend auf eine Publikation von Lambert (2005) mit dem treffenden Titel *Warum die Abiturordnung Gymnasium (NGVO) so kompliziert ist* verwiesen.

die Vorgaben für allgemeinbildende Schulformen mit gymnasialer Oberstufe analysiert – angesichts der Pluralität der Bildungswege zum Abitur könnten die Analysen noch deutlich erweitert werden, womit eine weitere Komplexitätssteigerung einhergehen würde. Im Fokus des Beitrags standen weitgehend allgemeine Strukturen und Rahmenbedingungen; sofern fachspezifische Betrachtungen erfolgten, waren diese auf die Fächer mit Bildungsstandards für die Allgemeine Hochschulreife konzentriert; die Regelungen für weitere Fächer – insbesondere auch „kleine Fächer" – waren kein Thema. Analysiert wurden ferner nur solche Aspekte, die in der Debatte über die (mangelnde) Vergleichbarkeit des Abiturs wiederholt diskutiert werden – angesichts der Komplexität und Vielschichtigkeit der Vorgaben konnten einige durchaus interessante und relevante Aspekte (z. B. besondere Lernleistung, Projekt-/Seminarkurse o. Ä.) nicht einbezogen werden. Limitationen ergeben sich auch durch die Dokumentenanalyse selbst – analysiert wurde das, was in den jeweiligen Dokumenten geschrieben steht. Aus anderen Projektkontexten ist den Autoren des Beitrags gleichwohl bekannt, dass es weitergehende Strukturen und Regelungen gibt, die jedoch aus öffentlich zugänglichen Dokumenten nicht ersichtlich werden (z. B. Kontrollmechanismen im Rahmen der Beurteilung von Abiturprüfungsleistungen und damit einhergehende Konsequenzen).

Letztlich wird mit der Auswertung der Dokumente eine Formalstruktur skizziert – nicht beantwortet werden kann folglich, welche Folgen sich aus bestimmten Strukturvorgaben, die auf der formalen Ebene implementiert sind, auf der Aktivitätsebene ergeben (für den hinter diesen Überlegungen stehenden theoretischen Ansatz vgl. Groß & Schmid-Kühn in diesem Band, Abschnitt 3). So haben beispielsweise alle Länder formal festgeschrieben, dass die Beurteilung von schriftlichen Prüfungsleistungen im Abitur auf der Basis festgelegter Korrektur- und Bewertungsvorgaben erfolgen soll (vgl. Abschnitt 4.1) – Aussagen darüber, wie diese Vorgaben ausgestaltet sind, wie die beurteilenden Lehrkräfte diese wahrnehmen und letztlich nutzen, sind nicht möglich, sondern müssten Gegenstand weiterführender Forschungsarbeiten sein (vgl. Beitrag 8 von Schröter et al. in diesem Band). Wiederholt wurde in diesem Beitrag auch auf die unterschiedliche Anzahl an Wochenstunden in Kursen mit formal gleichem Anspruchniveau (z. B. vier- oder fünfstündige Kurse mit erhöhtem Anforderungsniveau) hingewiesen – auch hier ist unklar, welche Folgen aus dem formalen Mehr an Lernzeit auf der unterrichtlichen Aktivitätsebene resultieren und welche Wirkungen auf Schülerebene sich dadurch ergeben. So muss mehr Unterrichtszeit nicht notwendigerweise mit mehr Kompetenzerwerb einhergehen – entscheidend ist vielmehr die aktive Nutzung der nominellen Lernzeit in einem qualitativ hochwertigen Unterricht (vgl. Schroeders et al., 2013, S. 340 ff.). Auch dazu sind auf Basis der vorliegenden Dokumentenanalyse keine Aussagen möglich; empirische Studien zum Unterricht in der gymnasialen Oberstufe stellen ohnehin weitgehend ein Desiderat dar.

Trotz der genannten Limitationen kann der Beitrag dazu beitragen, die fortwährende Vergleichbarkeitsdebatte zu fundieren und zu versachlichen; zudem wurden sich aus dieser Analyse ergebene Forschungsdesiderata im Bereich Abitur/gymnasiale Oberstufe benannt, deren empirische Bearbeitung noch aussteht.

Literatur

Aktionsrat Bildung (2011). *Gemeinsames Kernabitur. Zur Sicherung von nationalen Bildungsstandards und fairem Hochschulzugang. Gutachten*. Münster: Waxmann.

Dudenredaktion (o. J.). Synonyme zu *vergleichbar*. https://www.duden.de/synonyme/vergleichbar

Etymologisches Wörterbuch des Deutschen (o. J.). gleich, … https://www.dwds.de/wb/etymwb/gleich

Klein, E. D., Kühn, S. M., van Ackeren, I. & Block, R. (2009). Wie zentral sind zentrale Prüfungen? Zentrale Abschlussprüfungen am Ende der Sekundarstufe II im nationalen und internationalen Vergleich. *Zeitschrift für Pädagogik, 55*(4), 596–621.

KMK – Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2012a). *Bildungsstandards im Fach Deutsch für die Allgemeine Hochschulreife. Beschluss der Kultusministerkonferenz vom 18.10.2012*. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Deutsch-Abi.pdf

KMK – Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2012b). *Bildungsstandards im Fach Mathematik für die Allgemeine Hochschulreife. Beschluss der Kultusministerkonferenz vom 18.10.2012*. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Mathe-Abi.pdf

KMK – Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2012c). *Bildungsstandards für die fortgeführte Fremdsprache (Englisch/Französisch) für die Allgemeine Hochschulreife. Beschluss der Kultusministerkonferenz vom 18.10.2012*. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Fortgef-FS-Abi.pdf

KMK – Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2018). *Vereinbarung zur Gestaltung der gymnasialen Oberstufe und der Abiturprüfung. Beschluss der Kultusministerkonferenz vom 07.07.1972 i. d. F. vom 15.02.2018*.

KMK – Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2021). *Vereinbarung zur Gestaltung der gymnasialen Oberstufe und der Abiturprüfung. Beschluss der Kultusministerkonferenz vom 07.07.1972 i. d. F. vom 18.02.2021*. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/1972/1972_07_07-VB-gymnasiale-Oberstufe-Abiturpruefung.pdf

Kühn, S. M. (2010). *Steuerung und Innovation durch Abschlussprüfungen?* Wiesbaden: Verlag für Sozialwissenschaften.

Lambert, J. (2005). Warum die Abiturordnung Gymnasien (NGVO) so kompliziert ist. *SchulVerwaltung, Ausgabe Baden-Württemberg, 14*(10), 215–216.

Neumann, M. & Trautwein, U. (2014). Die (Rück?)Reform der gymnasialen Oberstufe. Hintergründe, Entwicklungen in den Bundesländern und empirische Befunde aus der TOSCA-Repeat-Studie. In C. Ritzi (Hrsg.), *Gymnasium im strukturellen Wandel: Befunde und Perspektiven von den preußischen Reformen bis zur Reform der gymnasialen Oberstufe* (S. 247–276). Bad Heilbrunn: Klinkhardt.

Schroeders, U., Siegle, T., Weireich, S. & Pant, H. A. (2013). Der Einfluss von Kontext- und Schülermerkmalen auf die naturwissenschaftlichen Kompetenzen. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 331–344). Münster: Waxmann.

Weber, B. (2019). Die Didaktiken der Gesellschaftswissenschaften zwischen Zersplitterung, Dominanz und Interdependenz. *Zeitschrift für Didaktik der Gesellschaftswissenschaften*, *10*(2), 11–42.

Tabellenverzeichnis

4 Ein Blick in andere Länder: Abschlussprüfungen im Spannungsfeld zwischen Standardisierung und Autonomie

Pauline Schröter, Lars Hoffmann & Svenja Mareike Schmid-Kühn

Zusammenfassung

Die Sicherung der Qualität und Vergleichbarkeit von Abschlussprüfungen geht mit Herausforderungen einher, denen unterschiedlich begegnet werden kann. Standardisierungsmaßnahmen und die Zentralisierung prüfungsrelevanter Prozesse stehen dabei in einem Spannungsverhältnis zur Autonomie und Flexibilität einzelner Akteure im Bildungssystem. Der folgende Beitrag wirft einen Blick auf die Prüfungssysteme anderer Länder und vermittelt einen Eindruck von der Spannbreite verschiedener Ansätze im Umgang mit bestehenden Herausforderungen. Mit Irland, Australien, den Niederlanden und der Schweiz werden vier Länder vorgestellt, die mit ihren Abschlussprüfungen ein breites Spektrum an Gestaltungsmöglichkeiten zwischen Standardisierung und Autonomie abdecken. Für jedes Land werden zunächst die Grundzüge des Prüfungssystems dargestellt sowie die Organisation und Durchführung der Prüfungen beschrieben. Anschließend folgen Auszüge aus Interviews mit Vertreter:innen der Bildungsadministration aus den jeweiligen Ländern, die zur Sicherung der Qualität, Vergleichbarkeit ihrer Abschlussprüfungen und zum Umgang mit daraus resultierenden Herausforderungen befragt wurden.

Einleitung

Mit der Diskussion über die Qualität und Vergleichbarkeit des Abiturs in Deutschland geht die Frage nach der „richtigen" Ausgestaltung des Prüfungsverfahrens einher. Vor diesem Hintergrund erfolgte in den letzten 15 Jahren eine zunehmende Standardisierung und Zentralisierung der Abiturprüfungsverfahren. Der Umstellung von dezentralen, schulintern organisierten Abiturprüfungen auf ein landeseigenes Zentralabitur folgten bald Initiativen mehrerer Bundesländer zur Entwicklung gemeinsamer Abituraufgaben und schließlich die Einrichtung eines ländergemeinsamen Pools von Abiturprüfungsaufgaben in den Fächern Deutsch, Mathematik, Englisch und Französisch, sowie zukünftig auch Biologie, Chemie und Physik (vgl. Beitrag 2). Weitergehende Zentralisierungsmaßnahmen wie ein gemeinsames Kernabitur (vgl. Aktionsrat Bildung, 2011) oder ein gesamtdeutsches Zentralabitur (z. B. Hoymann, 2005) werden zwar in der öffentlichen Diskussion thematisiert, sind derzeit aber nicht vor-

gesehen. Zugleich liegen bestimmte Prüfungen und Prüfungsteile (d. h. schriftliche Prüfungen in Fächern, die nicht zentral geprüft werden, mündliche Prüfungen und praktische Prüfungsteile) nach wie vor in der Verantwortung einzelner Schulen bzw. Lehrkräfte. Folglich ist die Diskussion über die Ausgestaltung von Prüfungsverfahren zur Sicherung von Qualität und Vergleichbarkeit zumindest aktuell immer vor dem Hintergrund eines Spannungsfeldes zwischen Standardisierung und Zentralisierung einerseits und der Autonomie und Flexibilität einzelner Bundesländer, Schulen und Lehrkräfte anderseits zu führen.

Auch international sind Abschlussprüfungen am Ende der Sekundarstufe II[1] ein etabliertes Instrument (vgl. Klein et al., 2009), wobei sich die Prüfungsverfahren erheblich voneinander unterscheiden (vgl. Klein & van Ackeren, 2012). Das vorliegende Kapitel soll einen Eindruck von der Spannbreite der Ansätze vermitteln, wie andere Länder mit dem Spannungsfeld zwischen Standardisierung und Autonomie umgehen. Ziel ist dabei nicht, die verschiedenen Prüfungssysteme vollständig und im Detail zu beschreiben, sondern einen Überblick über verschiedene Gestaltungsvarianten und Möglichkeiten des Umgangs mit bestehenden Herausforderungen zu geben. Vor dem Hintergrund der spezifischen Strukturmerkmale der Abiturprüfungen in Deutschland – *landeszentrale Prüfungsverfahren* aufgrund der föderalen Organisationsstruktur mit *ländergemeinsam* entwickelten Anteilen und zugleich auch *dezentralen Elementen* – wurden vier Länder ausgewählt, die ein breites Spektrum an Gestaltungsvarianten im Spannungsfeld von Standardisierung und Autonomie abdecken:

1. **Irland** als Zentralstaat mit zentraler Prüfungsorganisation
2. **Australien** als föderaler Staat mit landeszentraler Prüfungsorganisation in den einzelnen Bundesstaaten und Territorien
3. **Niederlande** als Zentralstaat mit teilweise zentraler sowie dezentraler Prüfungsorganisation
4. **Schweiz** als föderaler Staat mit dezentraler Prüfungsorganisation

Für alle Länder wurden Interviews mit Vertreter:innen der Bildungsadministration geführt, die über langjährige Erfahrungen mit den jeweiligen Prüfungsverfahren in den einzelnen Ländern verfügen. Die Interviews wurden im Sommer 2021 separat für jedes Land als Videokonferenz durchgeführt, aufgezeichnet und anschließend transkribiert. Die Darstellung der Kernaussagen erfolgt anhand einer Auswahl zentraler Interviewpassagen, die zur besseren Lesbarkeit aufbereitet wurden. Im Rahmen der Vorbereitung auf die Interviews wurden grundlegende Informationen zu den landesspezifischen Prüfungsverfahren mittels Dokumentenanalysen (z. B. Homepages, Prüfungsordnungen, Materialien zur Aufgabenerstellung und Bewertung für Lehrkräfte) und – sofern verfügbar – anhand wissenschaftlicher Fachveröffentlichungen erfasst und kategoriengeleitet in einer Übersichtstabelle für jedes Land dargestellt. Ebenso wie die Prüfungsverfahren in Deutschland (vgl. Beitrag 3) sind die jeweiligen

1 Stufe 3A der Internationalen Standardklassifikation des Bildungswesens (ISCED 3A, International Standard Classification of Education), die Schüler:innen auf ein Studium auf Universitätsniveau vorbereitet.

Prüfungssysteme und -verfahren der vier Länder sehr komplex, sodass sich die Darstellung auf wesentliche Kernelemente und den Regelfall konzentriert. Die Übersichtstabellen wurden den Expert:innen zur kommunikativen Validierung vorgelegt. Alle Beteiligten stimmten der Veröffentlichung der Informationen und der ausgewählten Interviewpassagen zu.

Die Abschnitte zu den spezifischen Prüfungsverfahren der vier Länder sind jeweils wie folgt gegliedert: Zunächst bietet eine Abbildung einen ersten Überblick über die Grundzüge des Prüfungssystems. Darauf folgt tabellarisch eine kategoriengeleitete Beschreibung der Prüfungsstruktur und einzelner Verfahrensschritte. Anschließend werden zentrale Interviewpassagen mit den Kernaussagen der Expert:innen dargelegt. Da die teilweise sehr komplexen Verfahren nur stark vereinfacht dargestellt werden können, sind am Ende jedes Abschnitts abschließende Empfehlungen zur vertiefenden Lektüre aufgeführt.

Insgesamt bietet der Beitrag einen Überblick über die Möglichkeiten der Gestaltung von Abschlussprüfungen am Ende der Sekundarstufe II und den zugrunde liegenden Steuerungsphilosophien in den einzelnen Ländern. Besonders im Hinblick auf die Entwicklung der Prüfungsaufgaben und die Bewertung der Prüfungsleistungen wird dabei deutlich, dass die Standardisierungs- bzw. Autonomiegrade im internationalen Vergleich sehr unterschiedlich ausfallen. Die Abiturprüfungen in Deutschland weisen, je nach Bundesland, einen vergleichsweise niedrigen bis mittleren Standardisierungsgrad auf (vgl. auch Klein et al., 2009, S. 615) und sind vor dem Hintergrund der in diesem Beitrag vorgestellten Prüfungssysteme eher in der Mitte des Spektrums zwischen Autonomie und Standardisierung zu verorten. Eine kritische Einordnung verschiedener Standardisierungsmaßnahmen in Bezug auf die Qualität und Vergleichbarkeit von Prüfungen erfolgt im Schlusswort dieses Bandes.

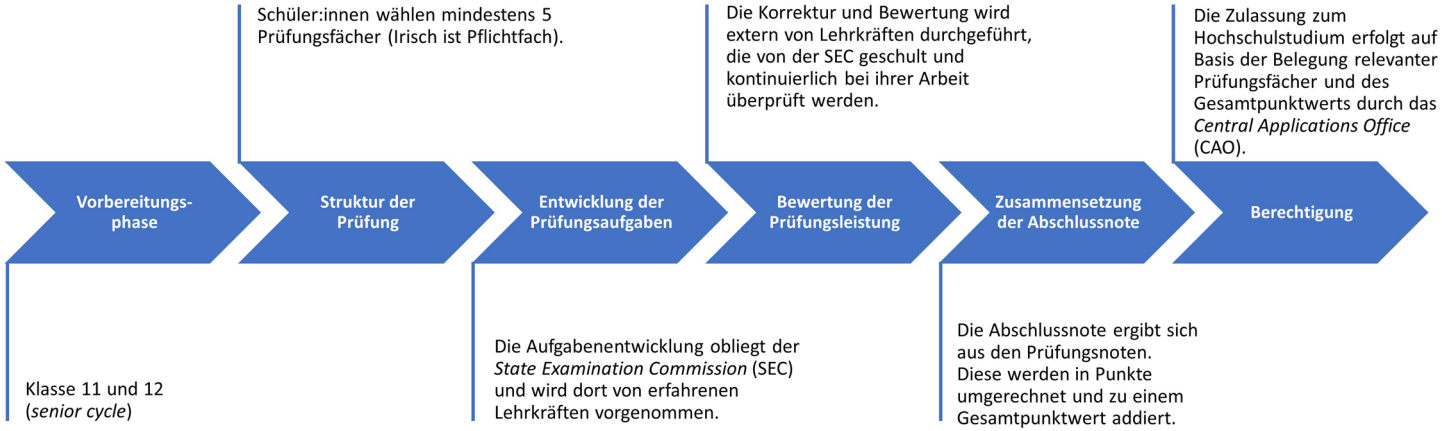
1 Das Leaving Certificate in Irland

"We treat everybody absolutely fairly, possibly maybe brutally fairly at times."

Tim Desmond

Irland verfügt über ein zentrales, landesweit einheitliches Prüfungssystem. Zuständig für die Entwicklung, Durchführung, Bewertung und Zertifizierung von Abschlussprüfungen in der Sekundarstufe II ist die *State Examination Commission* (SEC), eine unabhängige Einrichtung des Bildungsministeriums Irlands. Als Experte aus der Bildungsadministration wurde Tim Desmond befragt, promovierter Chemiedidaktiker, der bis zu seiner Pensionierung im Jahr 2021 als Leiter der Abteilung für *Examinations and Assessment* der SEC in Cork tätig war. Seit der Gründung der SEC im Jahr 2003 war er dort maßgeblich an der Entwicklung der Abschlussprüfungen am Ende der Sekundarstufe II beteiligt.

Leaving Certificate



1.1 Überblick über die Organisation und Durchführung

Rahmenbedingungen	
Organisation	Insgesamt drei verschiedene Abschlussprüfungen (das *established Leaving Certificate* (LC), das *Leaving Certificate Vocational Programme* (LCVP) und das *Leaving Certificate Applied* (LCA)) werden durch die *State Examination Commission* (SEC) organisiert und verwaltet. Die SEC ist eine unabhängige Einrichtung des Bildungsministeriums, die für die Durchführung und Zertifizierung der staatlichen Prüfungen verantwortlich ist.
Qualifikationsphase	Die Schüler:innen müssen in den letzten beiden Schuljahren der Sekundarstufe II den *Senior Cycle* (Qualifikationsphase: Klasse 11 und 12) durchlaufen. Optional besteht die Möglichkeit eines Vorbereitungsjahres (*Transition Year*), welches als „Brücke" zwischen dem *Junior Cycle* und dem *Senior Cycle* von den Schulen selbst gestaltet wird.
Vorbereitung	
Prüfungsfächer	Die Schüler:innen können aus 36 Fächern auswählen, welche in zwei Versionen (English oder Irisch) und auf drei verschiedenen Anforderungsniveaus angeboten werden (*Higher*, *Ordinary* und *Foundation* (nur Mathematik und Irisch)). Irisch ist das einzige Pflichtfach. Darüber hinaus gibt es keine Belegpflichten.
Aufgabenentwicklung	Verantwortlich für die Aufgabenentwicklung ist die SEC. Alle Prüfungen werden von erfahrenen Lehrkräften entwickelt.
Durchführung der Prüfung	
Art, Anzahl und Format der Prüfungen	In allen Fächern gibt es schriftliche Prüfungen (in den sprachlichen Fächern zusätzlich mündliche und hörbasierte, in manchen Fächern auch praktische Prüfungen). Die Schüler:innen müssen mindestens fünf Fächer für die Prüfungen auswählen.
Zeitlicher Umgang und Hilfsmittel	Die Dauer der Prüfungen variiert zwischen den Fächern und Anforderungsniveaus. In der Regel dauern die Prüfungen zwischen zwei und drei Stunden. Spezifische Hilfsmittel (Zirkel, Geodreiecke etc.) sind in einigen Fächern erlaubt. Zweisprachigen Schüler:innen ist die Verwendung von zweisprachigen Wörterbüchern erlaubt (außer im Fach Englisch).
Korrektur und Bewertung der Prüfungsleistung	
Erwartungshorizonte und Bewertungsmuster	Die Erwartungshorizonte werden zusammen mit den Prüfungsaufgaben von der SEC entwickelt. Nach den Prüfungen wird eine zufällige Auswahl von Arbeiten von einem Bewertungsteam vorläufig benotet, um die Erwartungshorizonte zu prüfen und ggf. nochmals zu überarbeiten.
Korrektur und Bewertungsverfahren	Lehrkräfte, die im Prüfungsjahr keinen eigenen Prüfungskurs unterrichten, können sich für die Durchführung und Bewertung der Prüfungen als *Superintendent* oder *Examiner* bewerben. Ausgewählt werden sie auf der Grundlage ihrer Qualifikation und Erfahrung und anschließend durch das SEC umfassend geschult. *Superintendents* erhalten Instruktionen, wie die Prüfungen in der Schule durchzuführen sind. *Examiners* nehmen an einer Bewertungskonferenz teil und erhalten eine Sonderzahlung für die Bewertungen von Arbeiten in Irisch. Mündliche und praktische Prüfungen werden von den *Examiners* vor Ort in der Schule bewertet. Schriftliche Prüfungen werden zu Hause bewertet – in den Hauptfächern durch ein von der SEC bereitgestelltes digitales Bewertungssystem (dafür werden die Prüfungsarbeiten gescannt). Die Noten von 1 (höchste) bis 8 (niedrigste) werden anhand von Bewertungstabellen vergeben. Leistungsbewertungen zwischen 100 % und 30 % werden in 7 Bereiche eingeteilt, die sich jeweils um 10 % voneinander unterscheiden.

Prüfungsergebnisse	
Zusammensetzung der Abschlussnote	Die Prüfungsergebnisse aus den besten Fächern werden genutzt, um das Gesamtergebnis zu berechnen. Für Mathematik auf erhöhtem Anforderungsniveau werden Bonuspunkte vergeben.
Berechtigung	Entsprechend ihres Rangplatzes auf der Punkteskala werden die Schüler:innen vom *Central Applications Office* (CAO) – einer Organisation, die für die Studienplatzvergabe zuständig ist – für die Hochschulen zugelassen. Für einige Studienfächer setzen die Universitäten die Belegung spezifischer Fächer im LC voraus.

1.2 Interview mit Tim Desmond

How would you describe the philosophy of your examination system?

It is predominantly a terminal examination. For a substantial number of subjects – particularly English, Irish, Mathematics and Sciences – there are simple straightforward terminal examinations at some time in June. A good number of subjects also have a coursework element or a project element. If you take the languages, candidates do an oral examination typically around Easter. For History, Geography, for instance, there is coursework which is carried out sometime during senior cycle and candidates produce a profile document with a report that is worth 20 % of the mark in that subject. If you take technical subjects like Metalwork, Woodwork, or Construction Studies, candidates do a project which is produced over the two years and the value there can be up to 50 % of the mark in that subject. In Music there is a performance test which is weighted to up to around that as well. But for the most part the Leaving Certificate is very much weighted in terms of one point in time during three weeks in June. And it is of absolute importance. The simple facts of the matter are that the Leaving Certificate examination is a gatekeeper in the extreme. There are requirements in some degree courses that you may have done certain subjects at school. Engineering would technically demand that you would have to do Chemistry and Physics. Medicine in most of the colleges would require Chemistry. The colleges also do aptitude tests which are part of their admission. But in terms of the transition process to third level education, the Leaving Certificate is of absolute importance – nothing less than that. It is the single examination which dictates opportunity and life chances with qualifications.

How fair do you think that is?

It is fair in that we treat everybody absolutely fairly, possibly maybe brutally fairly at times. Our system has almost exclusively external examination assessment for everything. If we start with the written examinations, you are assigned a center number and a candidate number. It is ensured that a teacher never marks a student from their own school. So, from the system's point of view, everybody is treated equally. You are just a number the teacher works through. When it comes to the oral and practical examinations, external teachers are appointed. There is no internal influence and a level of anonymity again. So, from that point of view this is very good. The reality is: there will

always be candidates who don't lend themselves to that kind of pressure and performance to do it on the day or do it for a week or 10 days in one shot. So, there is an issue there. How do you deal with that? Certainly, having more school-based continuous assessment or at least multiple points in time examinations is a possibility. At the very least, you could move to two points in time and give some assessment at the end of year 1 and year 2 in senior cycle as a way of maintaining the external examination but at the same time not having a "Today's the day – do it or die" situation. It is difficult to know what possibilities there are within the jurisdiction to get schools more involved in the examinations. It has been resisted very strongly by teacher unions.

Why do teachers resist to become more involved?

They largely hang it on two things. One, they see themselves as advocates for their own students. It is very hard to be an advocate for a student and be impartial after that. Second, there is the feeling that pressure will be brought to bear on teachers by school authorities and by parents. When you look at the experience of the last twelve months where we moved to calculated grades in the country due to Covid-19[2], the marks submitted by teachers were vastly inflated on what pattern of marks you would have got in any typical year there at school. And even after moderation the outcomes that were issued were very significantly inflated on any externally managed examinations in the previous x number of years. This suggests that we would have some challenges in removing the rose-tinted glasses from teachers if they were giving the marks. We are also engaged in a senior cycle reform at the moment, which will lead to a reform of the Leaving Certificate. It is very unclear what will happen, but I would certainly see that there are a lot of people pushing for more and more school-based assessments. This can be a double-edged sword from the point of view that it is shown here and elsewhere that such things tend not to be very discriminating. They also lend themselves to external interference and unwanted activities from helpers.

In Germany, people tend to believe that examination requirements and the value of results have declined. Is there a discussion like that in Ireland, too?

Much is said about declining results, the grading of results over time and the inflation of grades. There probably has been a slight inflation since the SEC was formed, but very slight by comparison. You can argue it away in terms of student improvement both from the point of view of more resource material available and more understanding of exams. They are just trained better at doing exams rather than necessarily becoming better students. Within the system, there is not much you can do about it. The shifting interest in students is a completely different problem though. There is research on this phenomenon going back to the mid-nineties. At the time Math and Physics in particular were probably a grade and a half harder than Home Economics and Geography, for example. If you take that into account regarding the race for points, students become mini economists when it comes to picking subjects. They

2 In the summer of 2020, leaving examinations in Ireland were cancelled due to Covid-19. Students were offered calculated grades based on their school achievement for each subject or the option to sit a written examination in winter of 2020.

want maximum return for minimum investment. So, if they perceive Geography to be easier than Chemistry, they do Geography, because the entry to tertiary education works just on points. Where you got them from does not actually matter. So, this can feed into subject choice. At the time when I was involved, the participation rates for Chemistry and Physics were in rapid decline and I argued for a restructure of the papers[3] and an easing of the profile to improve the results. I actually succeeded in winning the argument and moved what was the A rate, i. e. the return of As at the time, for Chemistry from 11 % to 20 %. The realities are that subjects are just different. It is comparing apples and oranges. A subject like Chemistry is going to be abstract, so for a set core of the population it is going to be very difficult. You are never going to make it as concrete as Home Economics. It is as simple as that.

What are other challenges?

Every examination paper we do in English we have to do in Irish. For the most part, those translators are retired teachers or retired inspectors from the department who have good Irish. But not all of them always have the subject-related expertise, so you have to be very wary that things such as technical terminology are what they should be. Sometimes a word cannot be translated into an equivalent word. I am thinking of a classic example where a plane for smoothing wood got translated into an aeroplane. But the translation is a legal requirement. It would be a political hot potato to even broach the subject of not offering it. In a subject like Chemistry there would be maybe 9000 candidates of which 250 are Irish students. There would be subjects where there might never be a candidate who would take the subject through Irish, but you would still have to translate it just in case. And you would still publish the Irish version of it because if students in the future wanted to take it through Irish, you could not disadvantage them by not having access to an Irish version of the paper. Again, this comes back to the school system. There are Gael schools, which are Irish schools and offer the subjects through the medium of Irish. Typically, all their students take all the subjects using Irish. But you may find that some of those would eventually sit an exam like Chemistry in English. There is also a bonus for answering questions in Irish, which is the result of a political decision. Candidates getting up to 75 % get 10 % of that mark added if they answered the questions in Irish. So, there is an advantage which was put in place to encourage Irish and supposedly to compensate Irish students for the lack of resources that might be there in terms of textbooks and resource material. I do not see that politicians have the nerve to suggest the abolition of that.

Is comparability in the exams an issue?

It pops up occasionally, but it is not perceived as an issue. It is something that those involved in the delivery of the system need to be conscious of, but it is not really that kind of all-consuming passion of everybody out there. When delivering the system, there are three things we do to ensure comparability. We have the paper, there is a marking scheme and we also do assessment grids. Assessment grids map the parts of

3 Here the term *paper* refers to a set of printed questions used for the examination.

the question against the specific educational standards, so that dreaded thing of "you shouldn't have asked that question – it was not on the course" does not arise. It focuses you on how the paper is balanced and pitched, so you have an examination which is of comparable standard and demand to previous years. The other thing tied to that is that we actually have absolutely rigid boundaries on scores. You have a cut line every 10 %. So, a paper marked out of 400 has the first line cut at 360 plus, the second line cut at 320 plus and so on. If you take into account that you have a significant cohort of candidates doing it, then the laws of large numbers start kicking in. As there is no reason why this year's group should be significantly superior or inferior in performance based to last year's, they should have a similar set of outcomes. So, within reason, this year's profile for a subject should be similar to last year's profile, somewhere like the mean profile for the last four or five years. You then take a sample of papers for starters, sort out the papers, see your advisor team. They look at those and see if the candidates are giving back what you expected and if it looks reasonable. You may make minor tweaks to the marking scheme at that stage and then, when marking starts properly, you mark a random sample of scripts of about 6 or 7 % of the cohort. And at the end of that sample, you have some indication of the profile. You also have a feel for how various questions are working or not working in the scheme and you will then maybe make some minor revisions to the marking scheme again. So, you may loosen the marks in certain cases or tighten them depending on whether the paper was panning out on the difficult or the easy side.

How do you ensure comparability of the results?

The marking takes about a month and during the first days we have a three-day conference with the senior advising examining team. The first day is an interrogation of the paper and scheme. The second day is applying that scheme to a number of selected scripts, which will be used as exemplars for the training of examiners. The third day is getting some training in and also applying any little refinements from our learning. We then go into marking properly and again into a three-day event. Here, the scheme is presented to the wider team that will be maybe 30 people in the case of Chemistry, maybe 100 people in the case of English. Again, there is a discussion and interrogation, and there may be some minor tweaks where somebody comes and says, "What if?''. Then they mark those exemplar scripts at the conference. Typically, there are 6 to 8 examiners to one advising examiner, who will watch them applying the scheme properly and making the right decisions. If they are marking on screen, they will do a practice screen on screen, which is standardized. And then x number of scripts are released to them. The advising examiners will monitor a number of those scripts while the examiners are marking them to see that their standard is right and the procedure is right and that both the application and mechanics are being applied properly. Typically, they monitor a minimum of 5 % of every examiner's work through the process.

What would you like to change about the system?

I would like to see "curriculum" defined a little looser. If you define curriculum a bit looser, then it allows you to ask more open questions. If you specify a core of knowl-

edge and then allow a level of freedom as to where you can go from there, then you can see whether students can jump to the next stage and do something with it. It will make comparability harder, but I think it will make education and the outcomes better. It is very hard to have an open exam with a closed curriculum. A closed curriculum tells you what to do, what you need to know, so at a particular level where you can go with that knowledge. I think you can develop better exams if you have the freedom within the system to bounce students into something a bit more unseen. I think if you could go there, then you will have teachers who will teach the subject and the understanding and the feel for the subject as opposed to turning the pages and just learning the lines.

Is there a development towards online exams?

From the delivery point of view, I think the movements in the last 12 months, where we have gone to remote learning due to the Covid-19 situation, have opened up an awareness, an engagement and a lack of fear. That has to feed into the examination situation. I think in due course you will see more online testing. It may be one of the ways to get to more multipoint assessments. I think the capabilities are not there at this stage to have testing with levels of artificial intelligence to actually select what question to ask next. CITO did quite a lot of work on that in The Netherlands, so I can see things like that happening.

How do you see Ireland's system in comparison to other countries?

I think that any assessment system has to be viewed within its own social, economic, and cultural context. There are historical issues which are always tied to acceptance. Whatever system people have grown up with and survived, it seems okay and therefore is accepted – no matter how bad it might be. Systems that work in one place will work for a variety of reasons. And they do not regularly translate from one jurisdiction to another because of the cultural side of things. "No system is perfect" is the first thing that has to be accepted. I believe you should be looking at systems to tweak and to change first of all. Because unless the system is completely screwed up, there is usually some redeeming feature in whatever is going on there. I mean we saw it in South Australia where they decided to more or less abolish central exams completely and within a couple of years ACER was asked to put something in place that made sense. What you need to look at is what features of your system are good. In our case, the fairness is good, the brutality is not. So, you have to see what you can do to alleviate the brutality. It is incremental refinement that is required. And if you can keep nudging in the right direction, two things will happen: one, you will improve, and two, if some bright idea was not that bright after all, you will go backward. Development bodies and their credibility is very much based on producing something new. So, they are quite often not really attracted towards tweaking something – they want something new. We lose the ability to be able to say, "right, it's not perfect, so let's try and move it". If I can move it forward a little bit in the right direction and the next person does the same thing – then we are going somewhere.

Empfehlungen zur vertiefenden Lektüre

Auf den Internetseiten der SEC finden sich unter https://www.examinations.ie/ weiterführende Informationen zum *Leaving Certificate*; u. a. sind dort auch die Prüfungsaufgaben (*Examination Papers*) sowie die Richtlinien zur Bewertung (*Marking Schemes*) frei zugänglich.

Eine wissenschaftliche Auseinandersetzung mit unterschiedlichen Aspekten zur Sekundarstufe II in Irland und dem Leaving Certificate findet sich u. a. in folgenden Publikationen:

Jeffers, G. (2011). The Transition Year programme in Ireland. Embracing and resisting a curriculum innovation. *The Curriculum Journal, 22*(1), 61–76.

Klein, E. D. (2013). *Statewide Exit Exams, Governance, and School Development: An International Comparison*. Münster: Waxmann.

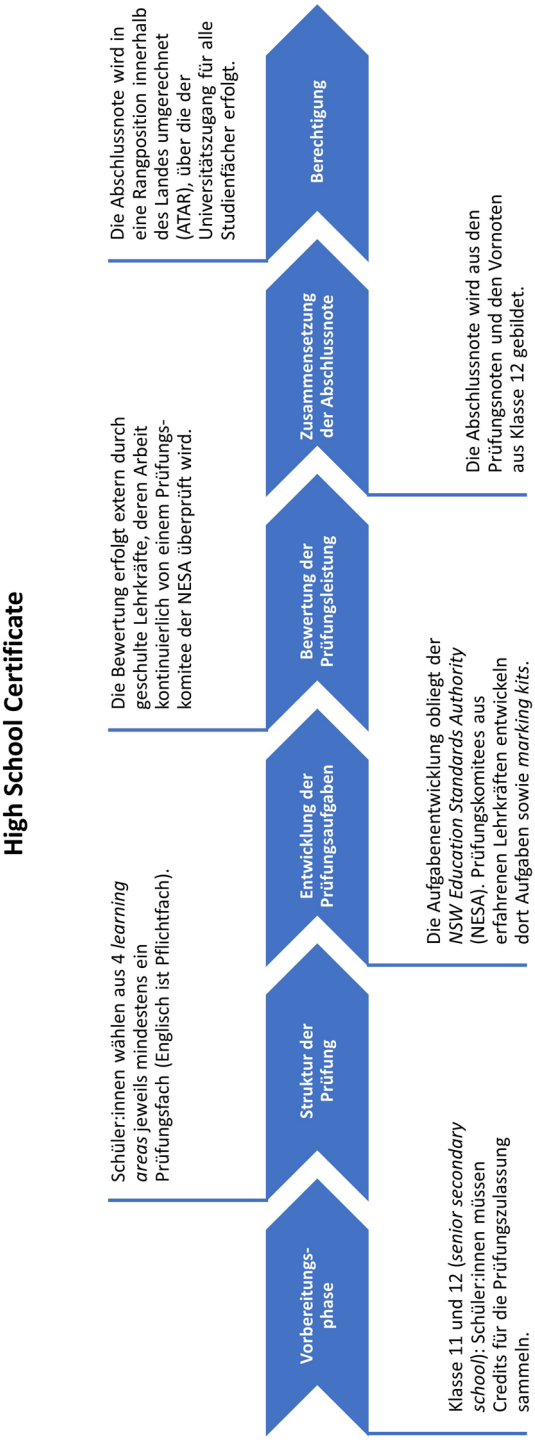
McGrath, P. (2021). The Leaving Certificate Examination – A Target for Unfair Criticism? *Irish Journal of Education, 44*(5), 1–23.

2 Das High School Certificate in New South Wales, Australien

"We are very rigorous on our procedures."

James Tognolini

Die Prüfungssysteme Australiens unterscheiden sich zum Teil erheblich zwischen den einzelnen Bundesstaaten und Territorien, innerhalb derer sie zentral organisiert sind. Mit dem *Australian Curriculum* gibt es eine ländergemeinsame Festlegung über Inhalte und Bildungsstandards in den Kernfächern, deren Implementation und Prüfung in der Verantwortung der jeweiligen Bundesstaaten und Territorien liegt. Die Ergebnisse der zentralen Abschlussprüfungen werden in den *Australian Tertiary Admission Rank* (ATAR) umgerechnet, der die Platzierung eines Prüflings innerhalb aller Prüflinge desselben Prüfungsjahres in einem Bundesstaat oder Territorium angibt. Dieser Prozentwert wird als länderübergreifend vergleichbar betrachtet und dient bundesweit zur Universitätszulassung. Das Prüfungssystem wird in diesem Abschnitt am Beispiel des Bundesstaats New South Wales (NSW) beschrieben. Als Experte für die Abschlussprüfungen in NSW wurde James Tognolini befragt, Professor für Bildungsforschung und Direktor des *Centre for Educational Measurement and Assessment* an der Universität Sydney. Als ehemaliger Direktor für Forschung des *Australian Council for Educational Research* (ACER) war er federführend bei der Entwicklung der australischen Abschlussprüfungen und berät seitdem nationale sowie internationale Regierungsorganisationen bei der Konzeption und Weiterentwicklung von Prüfungssystemen.


High School Certificate
Vorbereitungs-phase
Klasse 11 und 12 (*senior secondary school*): Schüler:innen müssen Credits für die Prüfungszulassung sammeln.
Struktur der Prüfung
Schüler:innen wählen aus 4 *learning areas* jeweils mindestens ein Prüfungsfach (Englisch ist Pflichtfach).
Entwicklung der Prüfungsaufgaben
Die Aufgabenentwicklung obliegt der *NSW Education Standards Authority* (NESA). Prüfungskomitees aus erfahrenen Lehrkräften entwickeln dort Aufgaben sowie *marking kits*.
Bewertung der Prüfungsleistung
Die Bewertung erfolgt extern durch geschulte Lehrkräfte, deren Arbeit kontinuierlich von einem Prüfungs-komitee der NESA überprüft wird.
Zusammensetzung der Abschlussnote
Die Abschlussnote wird aus den Prüfungsnoten und den Vornoten aus Klasse 12 gebildet.
Berechtigung
Die Abschlussnote wird in eine Rangposition innerhalb des Landes umgerechnet (ATAR), über die der Universitätszugang für alle Studienfächer erfolgt.

2.1 Überblick über die Organisation und Durchführung

Rahmenbedingungen	
Organisation	Alle Bundesstaaten und Territorien Australiens haben jeweils eine eigene Behörde und Bezeichnung für ihre Abschlussprüfungen. Das in New South Wales (NSW) verliehene Zertifikat für den Abschluss der Sekundarstufe II ist das *High School Certificate* (HSC). Zuständig für die Organisation und Durchführung der Prüfungen ist die *NSW Education Standards Authority* (NESA). Über die Summe der Noten aus dem HSC wird der sogenannte *Australian Tertiary Admission Rank* (ATAR) berechnet, der die Position eines Prüflings relativ zu allen anderen Prüflingen desselben Prüfungsjahres beschreibt und für die Zulassung zur Universität entscheidend ist.
Qualifikationsphase	In der *Senior Secondary School* (Klasse 11 und 12) sind alle Fächer vier verschiedenen Lernbereichen zugeordnet: Englisch, Mathematik, Naturwissenschaften, Geistes- und Sozialwissenschaften. Die meisten Fächer umfassen einen Basis- und einen Aufbaukurs, der jeweils mit 2 Credits vergütet wird. Um für die Prüfungen zugelassen zu werden, müssen die Schüler:innen mindestens 12 Credits durch Basiskurse (meist in Klasse 11) und 10 Credits durch Aufbaukurse (meist in Klasse 12) erworben haben In Klasse 12 müssen sie zusätzlich an einem *School Assessment* teilnehmen, das aus vier fachspezifischen Aufgaben besteht, die in der Schule gewichtet und bewertet werden. Die sich daraus ergebende Vornote geht in die spätere Berechnung der Gesamtnote ein. Um das HSC zu bekommen, müssen die Schüler:innen ebenfalls nachweisen, dass sie die Minimalstandards für Lese- und Schreibfähigkeiten sowie Rechenkenntnisse erreichen, indem sie im Laufe der Oberstufe jeweils einen standardisierten Lese-, Schreib- und Rechentest absolvieren.
Vorbereitung	
Prüfungsfächer	Aus jedem Lernbereich muss mindestens ein Prüfungsfach gewählt werden, wobei jedoch nicht alle Fächer für den ATAR angerechnet werden können.
Aufgabenentwicklung	Die Entwicklung der Prüfungsaufgaben wird von der NESA koordiniert. Lehrpersonen aus dem sekundären und tertiären Bildungsbereich mit angemessener Erfahrung und Qualifizierung können sich für die Mitarbeit in fächerspezifischen Komitees bewerben, die von *Chief Examiners* angeleitet werden. Die Komitees zur Aufgabenentwicklung werden jährlich für jedes Fach neu berufen und von der NESA vor Ort geschult. Die entwickelten Aufgaben und Materialien zur Bewertung werden von mehreren Experten geprüft und genehmigt.
Durchführung der Prüfung	
Art, Anzahl und Format der Prüfungen	Die Prüfungen können je nach Fach schriftlich, mündlich oder praktisch angelegt sein. Manche Prüfungen bestehen aus zwei Teilen, die an unterschiedlichen Tagen stattfinden. Aufgaben können im Multiple-Choice-Format, als halboffene Fragen mit Kurzantworten und als offene Fragen mit längeren Antworten (inkl. Aufsätzen) formuliert sein.
Zeitlicher Umgang und Hilfsmittel	Der zeitliche Umfang der Prüfungen variiert zwischen den Fächern und liegt bei ca. 2 bis 3 Stunden. Spezifische Hilfsmittel sind in einzelnen Fächern zugelassen (z. B. in Mathematik der wissenschaftliche Taschenrechner, in den Sprachen mono- sowie bilinguale Wörterbücher).

Korrektur und Bewertung der Prüfungsleistung	
Erwartungshorizonte und Bewertungsmuster	Die Bewertungsmaterialien werden von den *Senior Markers* und den *Supervisors of Marking* entwickelt und enthalten zusätzlich zum Erwartungshorizont auch Bewertungshinweise und Benchmarktexte.
Korrektur und Bewertungsverfahren	Die Prüfungen werden von Lehrkräften mit umfangreicher Bewertungserfahrung benotet, die sich jedes Jahr neu für das Bewertungsverfahren bewerben. Angeleitet von einem *Senior Marker* kümmert sich jedes Team, bestehend aus 6 bis 8 Mitgliedern, jeweils nur um eine einzelne Fragestellung bzw. einen Ausschnitt der Prüfung. Für die Zusammenführung der einzelnen Bewertungen zur Gesamtbewertung der Arbeit sind die *Supervisors of Marking* verantwortlich. Die Bewerter:innen durchlaufen eine Schulung und trainieren den Umgang mit den Bewertungsmaterialien. Sie diskutieren typische Antworten für verschiedene Notenstufen und bewerten ein Set von Antworten, bis die Bewertungen übereinstimmen. Schriftliche Prüfungen werden zu Hause durch ein von der NESA bereitgestelltes digitales Bewertungssystem bewertet (dafür werden die Prüfungsarbeiten gescannt). Aufgaben mit kurzen Antworten werden von einer Person bewertet, umfangreichere Aufgaben (z. B. Essays, Kreatives Schreiben, Projekte und Darbietungen) von mindestens zwei Personen. Untypische Antworten werden den *Senior Markers* oder *Supervisors of Marking* zugewiesen. Diese stellen außerdem sicher, dass die Teams die Bewertungsmuster korrekt anwenden. Wenn bei Bewerter:innen Unsicherheiten festgestellt werden, werden diese erneut geschult. Die *Supervisors of Marking* nutzen statistische Daten, um den Bewertungsvorgang zu überwachen und zu optimieren. Der Bewertungsprozess wird jedes Jahr von der Bildungsbehörde evaluiert.
Prüfungsergebnisse	
Zusammensetzung der Abschlussnote	In einem ersten Schritt werden die Vornoten aus dem *School Assessment* an die Prüfungsnoten angepasst (unter Beibehaltung der Rangreihenfolge der Schüler:innen und ihrer relativen Unterschiede in der Klasse), sodass diese über verschiedene Schulen miteinander verglichen werden können. Dieser Prozess wird als *Moderation* bezeichnet. In einem zweiten Schritt werden die Prüfungsnoten an Leistungsstandards angepasst. Diese können somit über verschiedene Jahre miteinander verglichen werden. Die Leistungsstandards haben sechs Levels, die als *Performance Band* bezeichnet werden. Ein Team aus Bewerter:innen legt die Mindeststandards fest, die in jeder Prüfung erforderlich sind, um den Leistungsstandard für jedes *Performance Band* zu erfüllen. Die Prüfungsnoten können somit auf einer standardbasierten Skala abgebildet werden. Dieser Prozess nennt sich *Standard Setting*. Die HSC-Note ist der Mittelwert aus der Vornote und der Prüfungsnote.
Berechtigung	Die Schüler:innen erhalten für jedes Fach die HSC-Note und das entsprechende *Performance Band*. Der ATAR wird aus der Summe der HSC-Noten gebildet und gibt die Position eines Prüflings in Relation zu allen Prüflingen für das entsprechende Prüfungsjahr in einem Bundesstaat an. Die Zulassung zur Universität basiert auf dem ATAR und ggf. zusätzlichen Kriterien (z. B. Interviews, Tests oder Portfolios), die von den Universitäten selbst bestimmt werden.

2.2 Interview mit James Tognolini

How has the Australian examination system developed into what it is today?

Most of our systems were developed to serve two purposes: secondary exit and tertiary entrance. So, if you track back to the 1900's, public examinations were mainly developed by universities to select the people for university. Given that in those subjects that

were examined we had information about students' performance, there also was a secondary exit function. As the number of people staying on the high school level increased, we broadened our course offerings and the way we thought about our exam system. So, certification became the main point there rather than tertiary entrance. Yet, we never really made a cut between them because it causes all sorts of trouble if we do. If we say we are only going to have exams in the subjects for university level, everybody will start to pick those subjects and get tutoring. And the other subjects will not be studied by students willing to go to university anymore. Since we do not want a two-tiered system – those who are going to university and those who are not – in NSW we now have something like 150 examinations and we bring it all back to one number for university entrance. We have set up structures to enable us to serve that dual purpose: secondary exit, i. e. what the students have done after 12 years, and tertiary entrance. But the challenge is always there for us.

What role does school-based assessment play?

We have our curriculum and we want to assess if the students are capable of the things it entails. We used to have just a three-hour examination on the curriculum, but you can only assess so much in a paper and pencil exam. So, there was a question on the validity of our examination process. What we tried to do – and most states grappled with that – was to assess more of the curriculum. We said, "Let's do a school-based assessment model where the schools are encouraged to assess the curriculum, in all their different subjects, what isn't already being assessed by the exam. That way we're getting more coverage and therefore a more valid result." What we found was – because the schools basically just replicated the issues that we faced – a high correlation between what students do in the exam and what they do in the school assessment. So, then we specified four tasks that forced the schools to do something that did not replicate the exam. We told them to try and do things that they would not normally do. All the different subjects will work it out themselves, but they have to try and show that they have got more coverage of the curriculum by putting the school-based assessment and the exam component together. We give them broad structural parameters but we let them do it. They can develop their own tasks. We do not judge them. The way we moderate the comparability between the different schools basically says, "Well, we trust you that you can assess your students in your school in the subject in these broad areas. We will take your advice, we won't change that advice, but we have to moderate it because we don't know if you're going to mark as hard or as easy as other schools." So, we have to have a statistical moderation procedure which sits on top of most of the systems which adjusts that school-based assessment in some way.

The HSC mark consists of the exam mark and the moderated assessment mark. Yet, you only use the ATAR for tertiary entrance. Why is that?

You get an HSC mark for every subject. But each student takes different subjects, some take 5, some take 7. But then, we do have a problem. It might be easier to get higher marks on something like Sport than on Applied Physics. So, we have to make

sure no student is disadvantaged or advantaged on the set of subjects they take. We take that HSC mark and then we scale it, so it takes account of the different subjects. This way, we end up with a scaled HSC mark. This moderation takes account of the easy and hard markers within a subject across schools. That makes them comparable. Once that scaling exercise is carried out, we add up their scores and end up with a single number. When we rank these numbers, here is the top student and here is the worst. The ATAR is the rank order of them. The top 0.1% of students have an ATAR of 99.9 and so on. And if I beat you by .01 of an ATAR rank, the community has to have great faith and trust that we know what we are doing. And we are very open because we can be challenged anytime. But, there is another layer of complexity, because each state can develop its own ATAR. The weakness in the whole thing is that we presume that the distribution of abilities across the states is the same, so that we can compare ATARs directly. So, we treat them as equal. But unlike Germany, most of our students go to university within their home state. We do not get many people from other states coming into NSW.

Can the ATAR be compared across time?

So, the ATAR is only used for tertiary entrance. But in NSW we also have a standard reference system. So, we actually have an end goal moderation system which allocates and describes performance levels from the basis of the skills that students have demonstrated across their exams and their school-based assessments. In our state, we would say that a performance band 6 student has demonstrated these skills. So, we can actually look at the performance of our students across time. And we can tell you whether our students are getting better or worse. We put a lot of money, time and effort into getting that. Using the professional judgement of examiners and teachers in the end goal method is our secondary exit. Because we use standards referencing, we have to align our marking rubrics to the standards and therefore pick out scripts from last year which are on the borderline between what we say a performance band 5 and a band 6 student is. That is the level everyone wants to get at in their exams. We have what we call standards packages that schools and students can look at. We say, "Students, if you want to be a band 6, this is the level you are going to need to get at. And this is where you are working at and how we are going to help you get there." It's a pretty sophisticated system.

How do you control for inter-rater reliability?

Inter-rater reliability is very strong. We go to great lengths to make sure it is in each state. Before a person is allowed to start marking for the day, they have to do what we call a calibration script, which we have already marked, to make sure they are marking consistently. Also, we have standardized scripts that go through to the markers constantly. Everybody marks the same script at the same time – they do not know they are doing it – but we have already marked it. If somebody is off key, we train them again, put them back on and check their past papers. And if we cannot train them, we cut them out of the system.

How do you ensure the validity of the exams?

We actually start any new curriculum by writing down what the standards are. What do we expect to see at the end of year 12 in a top Physics student or next level down? We always start with the standards. Then we build the curriculum to make sure that it will teach the standards. And then what we have to do is write our exams so that they are testing the curriculum which then gives information against the standards. So, when we design our exams, we have to make sure that there are questions on those exams which will enable the top students to be able to demonstrate they have those skills. And this is very hard, because people teach to our exams and students get tutored on them. So how do you test the higher order thinking that is required in that band 6? We are challenged all the time to make sure that our exams are really going to enable the top students to demonstrate that. So, in NSW we do a lot of work because otherwise we have a real validity problem. Nobody else, I think in Australia, does what we do. There are lots of people who believe that we should be having alternate learning journeys, that people can take different paths and not be as structured as we are. So, we have that sort of pressure. But if I walked out here and talked to the taxi driver, they will know about the HSC. They know it is the best system in the world. So, if you want to change that, if any government wants to change something like that, it is taking on a big risk because people just have faith in the system. They think they have the gold standard. We can have all sorts of curriculum reform and national curricula that go from K to 10. But you cannot touch year 11 and 12.

So, if people consider it the gold standard, is comparability in the exams an issue at all?

No, because people can see what we are doing and it makes sense to them. If students pick an easy subject, people can see they picked an easy subject. They get higher marks but then, when we produce the HSC score, that has been scaled back for that purpose. Not to say that they did not do well. They have done well there, but for tertiary entrance it does not work. Within a subject we are very rigorous on our procedures. People understand the moderation process and they understand the scaling process. So, we just have those processes in place. We get challenged on the standard setting and I think we should be doing more work to validate that the standards are right. If the minister asked me to put my hand on my heart and say that 16.3 % of students in English demonstrated these skills I think I might fail a lie detector. We have got no validation but we are starting that now. We will get alternate ways of seeing whether that 16 % is right or not.

Are there other existing or upcoming challenges?

We do not feel challenged at all by our system because we are on top of it. But there are challenges. The ATAR is losing its value because a lot of our universities just want people to come to them, so they do all sorts of things which devalues the ATAR. So that is a major problem for us. That affects our system. One of the challenges I have got at the moment is that some universities are offering students what we call early entry based on their year 11 result: "We will let you into our university now to stop the Uni-

versity of Sydney etc. from getting you." The problem with that is that those students then do not even try in year 12. And if they do not try, that affects the statistical moderation process and becomes quite a problem, which we have to try and solve. The other thing is that a lot of people believe that we should not be using this sort of entry at all and that we should be using more global sorts of measures. And that is attractive to some of our lower-level universities, of course. If people came to our government and said, "We think universities should decide which students they want", universities would automatically come to my center and say, "Write some exams for us. We're going to use some exams to get these people in because we don't believe it can be done fairly." Some rich people can go and put lots of money into portfolios and all sorts of fancy things, poor people cannot. This is just too big of an issue for us. So, we are grappling with problems all the time. But outside the system works well.

Is there a development towards online exams?

We promised ourselves we would be there about 15 years ago, we would have it all online. Not many systems have moved online in the high-stakes area. We have one online now – Science Extension – where we wanted the students to do a project and then the project is loaded into the system and the students do an exam with reference to their project. We have trialed that. It went okay but we still have big problems there. So, no, we have not gone online. One of the problems is that if we say we go online, the students across the state will do it in their exam rooms. We have got issues about internet access, which will blow up on us, but we are prepared to give that a go. But then we also have our schools who say, "We don't have the capacity to do it, we don't have enough computers." So, then you say, "Well, as an interim stage, we'll have some schools do it online and then some do paper and pencil." But then we have a comparability issue. You say, "Well, then let's have them do two separate subjects." But how do we make that transmission? I am a strong believer. We have our National Literacy and Numeracy Test for 3, 5, 7 and 9 online now. We are probably going to use automated essay scoring and get the results back within an hour of doing it in the next year or so. But we are not at the year 12 – not yet.

If I asked you again in 10 years, what do you think will have changed in NSW?

In 10 years, I do not think we will have changed much. I am not even sure if we will have gone online in 10 years. We have done a lot of exploration trying to get our exams more valid. The challenge I face is that because our exams are so high-stakes and released every year – because if not, people share them anyway – the teachers teach to those exams. And if you are an examiner, you cannot change it that much because people will argue that you are not matching the curriculum. It is that high-stakes out in the community. So, there is a lot of predictability about our exams. It is very hard to write these questions to assess higher order thinking, critical thinking, solving problems in different contexts etc., because ultimately, the teachers just teach the students the solution. You see that question, do this, this and this. And then we are not assessing the higher order skill. So, we are grappling with that and we're being challenged.

Empfehlungen zur vertiefenden Lektüre

Einen allgemeinen Überblick über die Sekundarstufe II in Australien gibt es unter https://www.australiancurriculum.edu.au/senior-secondary-curriculum/.

Alles Wissenswerte rund um das HSC findet sich auf den Internetseiten der NESA unter https://www.educationstandards.nsw.edu.au/wps/portal/nesa/11-12/hsc/about-HSC/. Anhand von Beispielen werden hier auch die Prozesse *Moderation* und *Standard Setting* im Detail dargestellt.

Unter https://www.uac.edu.au/future-applicants/atar wird die Definition und Funktionsweise des ATAR genauer beschrieben.

Eine wissenschaftliche Auseinandersetzung mit unterschiedlichen Aspekten zur Sekundarstufe II in Australien und dem HSC findet sich u. a. in folgenden Publikationen:

Bennett, J., Tognolini, J. & Pickering, S. (2012). Establishing and applying performance standards for curriculum-based examinations. *Assessment in Education: Principles, Policy & Practice, 19*(3), 321–339.

Hughes, J. (2019). The antecedents of the New South Wales Curriculum Review: an introduction to the New South Wales curriculum style. *Curriculum Perspectives, 39*(2), 147–157.

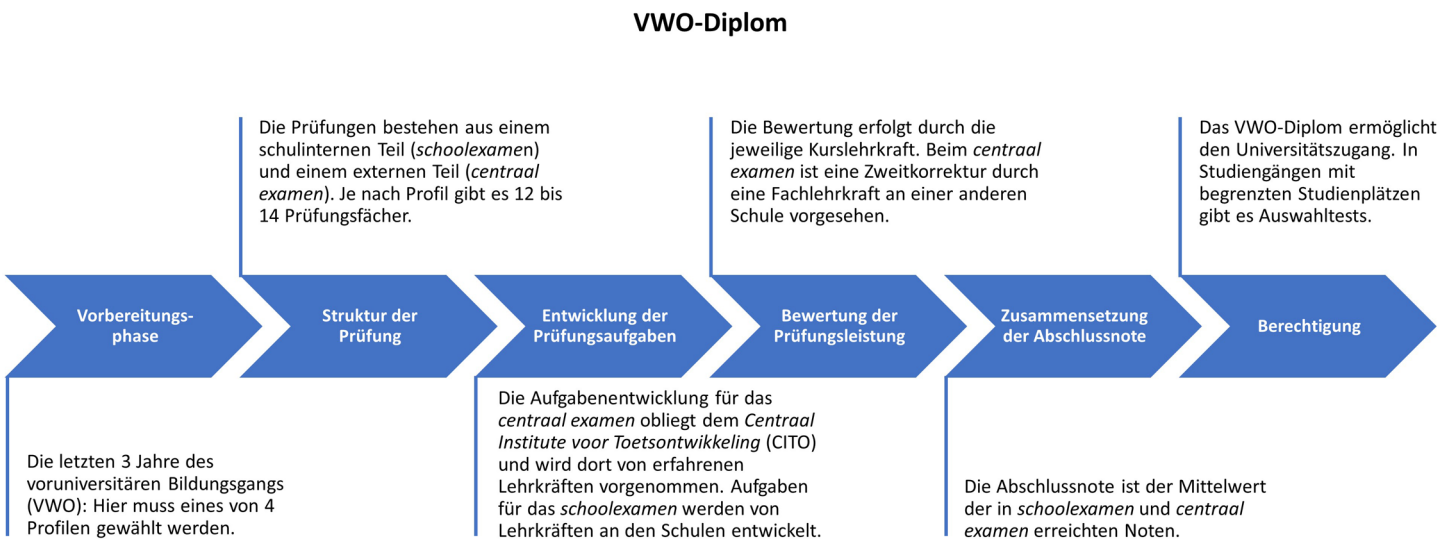
Savage, G. C. & O'Connor, K. (2019). What's the problem with 'policy alignment'? The complexities of national reform in Australia's federal system. *Journal of Education Policy, 34*(6), 812–835.

3 Das VWO-Diplom in den Niederlanden

"They have liberty to do whatever they like, but the central part is the same for all students."

Marieke van Onna

Das niederländische Prüfungssystem zeichnet sich dadurch aus, dass die Prüfungen am Ende der Sekundarstufe II sowohl einen zentralen, landesweit einheitlichen Teil als auch einen dezentralen Teil haben, der von den Schulen selbst gestaltet werden kann. Die Entwicklung und Evaluation der zentralen Prüfungsanteile wird im Auftrag des Colleges für Tests und Prüfungen (*College voor Toetsen en Examens*, CvTE) vom Zentralinstitut für Testentwicklung (*Centraal Institute voor Toetsontwikkeling*, CITO) geleitet, das sowohl nationale als auch internationale Dienstleistungen im Bereich *Educational Assessment* anbietet. Als Expertin für das Prüfungssystem wurde Marieke van Onna befragt, promovierte Psychometrikerin und wissenschaftliche Mitarbeiterin am CITO in Arnheim, wo sie für die Standardisierung und Evaluation der zentralen Abschlussprüfungen zuständig ist.


VWO-Diplom
Vorbereitungs-phase
Die letzten 3 Jahre des voruniversitären Bildungsgangs (VWO): Hier muss eines von 4 Profilen gewählt werden.
Struktur der Prüfung
Die Prüfungen bestehen aus einem schulinternen Teil (*schoolexamen*) und einem externen Teil (*centraal examen*). Je nach Profil gibt es 12 bis 14 Prüfungsfächer.
Entwicklung der Prüfungsaufgaben
Die Aufgabenentwicklung für das *centraal examen* obliegt dem *Centraal Institute voor Toetsontwikkeling* (CITO) und wird dort von erfahrenen Lehrkräften vorgenommen. Aufgaben für das *schoolexamen* werden von Lehrkräften an den Schulen entwickelt.
Bewertung der Prüfungsleistung
Die Bewertung erfolgt durch die jeweilige Kurslehrkraft. Beim *centraal examen* ist eine Zweitkorrektur durch eine Fachlehrkraft an einer anderen Schule vorgesehen.
Zusammensetzung der Abschlussnote
Die Abschlussnote ist der Mittelwert der in *schoolexamen* und *centraal examen* erreichten Noten.
Berechtigung
Das VWO-Diplom ermöglicht den Universitätszugang. In Studiengängen mit begrenzten Studienplätzen gibt es Auswahltests.

3.1 Überblick über die Organisation und Durchführung

Rahmenbedingungen	
Organisation	Die Gesamtverantwortung für das Bildungssystem liegt in der Hand des niederländischen Ministeriums für Bildung, Kultur und Wissenschaft. Das Ministerium legt den gesetzlichen Rahmen fest, an denen Schulen gebunden sind, und formuliert Bildungsstandards.
Qualifikationsphase	Die weiterführende Bildung beginnt im Alter von 12 Jahren. Etwa 20 % der Schüler:innen nehmen an den Prüfungen der voruniversitären Bildung (VWO) teil.[4] Das 6-jährige Programm ist das höchste Level der Sekundarstufe II und das abschließende Diplom der Zugang zur Universität. Die eigentliche Qualifikationsphase umfasst in der VWO die Schuljahre 4 bis 6 und ist durch Spezialisierung gekennzeichnet: Schüler:innen wählen eines von vier Profilen (*Nature and Science, Nature and Health, Economy and Society, Economy and Culture*). Einige Fächer (z. B. Niederländische Sprache und Literatur, Mathematik) sind Pflichtfächer; sie gibt es in allen Profilen. Andere Fächer sind Profilfächer; sie werden spezifisch nur im jeweiligen Profil unterrichtet.
Vorbereitung	
Prüfungsfächer	Die Kombination und Anzahl an Fächern, in denen die Schüler:innen Prüfungen ablegen, variiert zwischen den Profilen (zwischen 12 und 14 Fächer). In den meisten Fächern müssen sowohl die schulinternen als auch die schulexternen Prüfungen abgelegt werden. Für jedes Fach veröffentlicht das CvTE einen sogenannten Syllabus. Dieser enthält allgemeine Beschreibungen der einzelnen Domänen und spezifiziert außerdem, welche Domänen jeweils Bestandteil der schulinternen Prüfungen sind, und welche im Rahmen der schulexternen Examen geprüft werden.
Aufgabenentwicklung	Die schulinternen Prüfungen werden von den Schulen selbst oder von beauftragten Testinstituten erstellt. Die schulexternen Prüfungen werden nicht durch das CvTE erstellt, sondern durch das auf die Entwicklung von Testverfahren spezialisierte CITO. Das CITO beauftragt dabei erfahrene Lehrkräfte, die in Gruppen zusammenarbeiten und von einem/einer CITO-Prüfungsentwickler:in beaufsichtigt werden, mit der Entwicklung von Prüfungsaufgaben. Die endgültige Genehmigung der Prüfungsaufgaben erfolgt durch das CvTE, das dafür ebenfalls Lehrkräfte sowie Expert:innen aus dem Bildungsbereich (z. B. aus den Fachdidaktiken) einsetzt.
Durchführung der Prüfung	
Art, Anzahl und Format der Prüfungen	Die schulinternen Prüfungen bestehen aus zwei oder mehr Teilen pro Fach (mündlich, praktisch oder schriftlich). Die schulinternen Prüfungen müssen abgeschlossen und die Ergebnisse an die Schulaufsichtsbehörde übermittelt worden sein, bevor die schriftlichen schulexternen Prüfungen beginnen. Die Fragen sind entweder offen gestellt oder als Multiple-Choice-Fragen konzipiert. Die Schüler:innen müssen alle Aufgaben bearbeiten. Es gibt also keine optionalen Teile in den schriftlichen Prüfungen.

4 Alle weiteren Ausführungen in dieser Tabelle beziehen sich auf das VWO.

Zeitlicher Umgang und Hilfsmittel	Alle schriftlichen Prüfungen haben eine feste Bearbeitungszeit von drei Stunden. Ein niederländisches Wörterbuch ist für alle Prüfungen zugelassen. Schüler:innen mit einer anderen Herkunftssprache dürfen auch ein zweisprachiges Wörterbuch (Niederländisch/Muttersprache) nutzen. Im Fach Englisch ist ein einsprachiges Englischwörterbuch zugelassen, in den anderen modernen Fremdsprachen und in Friesisch sind zweisprachige Wörterbücher erlaubt. In den klassischen Sprachen (Latein, Altgriechisch) darf neben einem zweisprachigen Wörterbuch auch eine Grammatik genutzt werden. In Mathematikprüfungen dürfen der wissenschaftliche Taschenrechner, ein Geodreieck und Zirkel verwendet werden.
Korrektur und Bewertung der Prüfungsleistung	
Erwartungshorizonte und Bewertungsmuster	In den Syllabi für die Prüfungen ist festgelegt, wie die Noten berechnet werden. Erwartungshorizonte und Bewertungshinweise existieren für die schulexternen Prüfungen. Die Erwartungshorizonte werden von denselben Personen entwickelt, die auch die Prüfungen erstellt haben. Die endgültigen Versionen werden vom CvTE genehmigt, bevor sie an die Lehrkräfte ausgehändigt werden. Während der Bewertungsphase können, falls nötig, Änderungen an den Erwartungshorizonten durch das CvTE vorgenommen werden.
Korrektur und Bewertungsverfahren	Bewertet und benotet werden die schulinternen Prüfungen durch die jeweiligen Fachlehrer:innen. Die schulexternen Prüfungen werden ebenfalls von den entsprechenden Fachlehrer:innen bewertet und von einer weiteren Lehrkraft einer anderen Schule kontrolliert. Die Bewertungen werden mithilfe einer von CITO zur Verfügung gestellten Transformationsformel in Noten umgerechnet, wobei die Notenskala die Werte von 1 (sehr schlecht) bis 10 (exzellent) umfasst. Durch die Anwendung der Transformationsformel erfolgt ein *test equating*, welches sicherstellen soll, dass die Noten über die Jahre hinweg vergleichbar sind. Falls also eine Prüfung im Vergleich zu den Prüfungen der vorherigen Jahre etwas einfacher oder schwieriger ausfällt, so wird dies durch die Umrechnung ausgeglichen. Vor diesem Hintergrund erhebt CITO in jedem Prüfungsjahr flankierend zu den Prüfungen zusätzliche Daten, um mit deren Hilfe die Transformationsvorschrift für das *test equating* neu zu bestimmen. Die Schulen sind verpflichtet, die von CITO bereitgestellte Formel zu benutzen, um die bei den schulexternen Prüfungen erzielten Punkte in Noten umzurechnen.
Prüfungsergebnisse	
Zusammensetzung der Abschlussnote	Die Endnote für jedes Fach ist der Mittelwert der in der schulinternen und schulexternen Prüfung erzielten Noten. Generell ist eine 5,5 die minimal ausreichende Note zum Bestehen. Wenn ein Prüfling in allen Fächern Noten von 6 und höher erzielt, erhält sie oder er das VWO-Diplom. Ein Prüfling besteht jedoch auch dann, wenn in einem Fach nur die Note 4 oder 5 erreicht wird und in allen anderen Fächern Noten von 6 und höher erzielt werden. Ein Bestehen ist auch möglich, wenn in einem Fach die Note 4 und in einem zweiten Fach die Note 5 erreicht wird. Bestehensvoraussetzung ist hier, dass in allen anderen Fächern Noten von 6 und höher erreicht werden und nur eines der zwei Fächer, in dem die Minderleistungen erzielt wurden, ein Profilfach ist. Im Abschlusszeugnis der Schule werden die Ergebnisse in den schulinternen Prüfungen, die Ergebnisse in den schulexternen Prüfungen, die Endnoten für jedes Fach und die Gesamtnote ausgewiesen.
Berechtigung	Das VWO-Zertifikat erlaubt den Zugang zu universitärer und höherer Berufsausbildung.[5] Seit einigen Jahren können Universitäten neben dem VWO-Diplom eigene Zulassungsvoraussetzungen für Studiengänge mit begrenzten Plätzen wie z. B. Medizin festlegen. Für Studiengänge ohne Studienplatzbegrenzung müssen sie Auswahltests anbieten, die allerdings nicht sehr selektiv sind.

5 Der Zugang zu universitäter und höherer Berufsausbildung ist auch durch ein HAVO-Zertifikat möglich, das jährlich rund 28 % der niederländischen Schüler:innen erwerben.

3.2 Interview mit Marieke van Onna

What are the core aspects of your examination system?

In the Netherlands the core idea is that the school is the one who awards a diploma. But to ensure that each diploma is comparable across all the schools, the exam consists of a central part which is used for anchoring to make sure all schools use the same standards. So, there is a program, and all the schools should do what is in the program. It says which parts will be tested centrally and which parts they have to test themselves. For example, in the languages usually only the reading is tested centrally and the schools have to test the writing, listening and the speaking aspects. They have liberty to do whatever they like, but the central part is the same for all students. The program per subject is changed usually every three to five years. The SLO – the Netherlands' institute for curriculum development – is advising the government on which parts to test centrally and which parts to leave up to the schools. The schools already know two or three years in advance what the program is, so they can adapt to that. When there are major changes in the program, we also do a pilot. At a few schools, which would like to join the pilot, we start educating the children already according to the new program. And we do some pilot exams for one or two years to see whether the new program actually works. This way, we can still adjust it a bit before it is introduced to all schools.

What do you do to ensure the comparability of exam requirements and results?

Our primary concern is to ensure the comparability of exam grades between different years. To achieve that, we do test equating. So, in order to know the difference in difficulty between an exam in one year and the next year, we do some pilot testing. Some of our exams are tested two years before. We give them to schools so they can put them in some kind of school exam. They report the marks and scores on all items, which are related to some anchor items that are kept secret. We get teachers to sign that they will keep them secret. With this pre-test data we can estimate the difficulty of the exam for each time. So, the pre-testing is one thing we do. We also do post-tests, but that is only possible for subjects in which you can do very quick scoring, usually with multiple choice items like in the foreign languages. What we do is the following: We have the exam on Wednesday, for example, and then on Thursday we will select another population which is more or less equally able but not in an exam year. And we ask the school to administer parts of this exam and anchor items in a school exam or end of year test. We do that repeatedly and compare the difficulty of the exams by means of the anchor items which are in these post-tests. So that is the second thing we do. The third thing is that we have digital exams at the lowest levels, so not in VWO. Yet, pretests are not that reliable after all, because exam writers may alter the items between the pre-test and the end of school exam. Also, the number of students taking the pretests is usually low and we see school effects. Sometimes they have discussed a topic beforehand and do quite okay. Sometimes they have not discussed a topic yet and perform poorly. So, you get some differential item functioning in your pre-testing in comparison to your exam at the end of school. So, if the results of a pre- or posttest differ

too much from an assumed stable population ability, we try to find something in between and that would be checked and revised. Unless there is an indication that the pre-test was completely invalid, we usually stick to that.

Is there a debate on fairness in the exams?

The problem is getting into VWO. There is a lot of debate on that. Children from migrant backgrounds have lower probability of entering VWO. In addition, there is some degree of underperformance at primary education level in non-urbanized parts of The Netherlands. So, there is some debate on whether students have equal opportunities to enter the VWO. Already at the age of 12 they are sorted out more or less in different secondary educational tracks. We used to have this system in which you could always jump. If you had done an examination on a lower level, you could jump to the next level. But the jumping has been made increasingly difficult. Schools tend to set all kinds of restrictions now. So, there is a debate about the difficulty of entering the VWO program, but once students are in, there is not that much debate anymore.

Is digitalization an issue for the educational system?

We only do digital exams at the lowest levels because the secrecy of all the exam questions is much harder to ensure at higher levels. At higher exam levels, we do have Music, for example, which now is a hybrid system. Students have a laptop for listening to the music or watching a dance, but they still have to fill out a paper form, which is a bit odd because they could answer the questions directly on the computer. What we do see, however, is that the capacity of students to read long texts on paper is decreasing. For History or Geography, for example, they cannot use long texts anymore because then students get time problems, which is something that is worrying. We can only ask basic questions. Printing a long text for a higher level is getting more and more difficult. That is something we need to do more research on.

What are other existing challenges?

A lot of the developments that we see concern the universities, who have started to set up their own selection tests. They were asked to do so by the government, because the drop-out rate for first-year university students was too high. The universities needed some sort of selection to see whether students are really dedicated and capable of succeeding in a degree course. However, if you have 300 students coming in, you cannot have interviews with all of them. So, the universities give them some sort of selection test. That is far more selective than the exams and it is starting to get a much bigger role over time. Until now, the universities stick with this practice. But what if they start asking different things? Then the students have to choose whether they will study for the selection test or for the central examination. The highest-scoring students will come in, but if you have a very high scoring nerd in a selection test for Medicine, it does not mean that she/he will become a good doctor. You need all the other abilities as well. So, there is a lot of debate about these selection tests. We used to have a weighted random selection to enter Medicine. So, if you had a high score on your central exam, you would have a high weight and thus a higher chance of getting in. But de-

spite the weighting mechanism, people still had the impression that the selection process was like a lottery and considered it as unfair. Now they have this selection test and they say that is not fair either, because the tests are not measuring the right skills or motivation. So, these kinds of discussions keep coming and going.

Are there any future developments in sight?

It is not very concrete yet, but there is a program we have started already several years ago. It is called "Education 2032" – so what should education look like for students who turn 18 in 2032. It was supposed to be a cooperation between the government and all the teachers. But of course, the government came up with some plans on what to teach the children and the teachers could only react to that. And well, they did not agree. That is the problem. One aspect of the plan is that we should narrow down the central examination part and keep it to only a core of five subjects instead of this whole range of subjects, but then all these teachers of the non-core subjects start protesting. So, for example, the teachers of the relatively smaller languages – like French and German – are afraid that the central examination of these subjects will stop. They say: "Then we won't be important anymore, so students won't pick our subjects." So, it is a political debate in that sense. In the end, I think, there will be some upheaval in secondary education and some new design of the whole examination system.

Empfehlungen zur vertiefenden Lektüre

Einen allgemeinen Überblick über die Sekundarstufe II in den Niederlanden gibt es unter https://www.government.nl/topics/secondary-education/.

Detaillierte Informationen zum VWO-Diplom finden sich unter https://eacea.ec.europa.eu/national-policies/eurydice/netherlands/secondary-and-post-secondary-non-tertiary-education_en.

Unter https://www.examenblad.nl/item/vwo/2022/vwo können die zentralen Prüfungsaufgaben aus den vergangenen Prüfungsjahren für alle Fächer heruntergeladen werden.

Eine wissenschaftliche Auseinandersetzung mit unterschiedlichen Aspekten zur Sekundarstufe II in den Niederlanden und dem VWO-Diplom findet sich u. a. in folgenden Publikationen:

Béguin, A. & Ehren, M. (2011). Aspects of accountability and assessment in the Netherlands. *Zeitschrift für Erziehungswissenschaft, 14*(1), 25–36.

Klein, E. D. (2013). *Statewide Exit Exams, Governance, and School Development. An International Comparison*. Münster: Waxmann.

Rijn, P. W. van, Beguin, A. A. & Verstralen, H. H. F. M. (2012). Educational measurement issues and implications of high stakes decision making in final examinations in secondary education in the Netherlands. *Assessment in Education: Principles, Policy & Practice, 19*(1), 117–136.

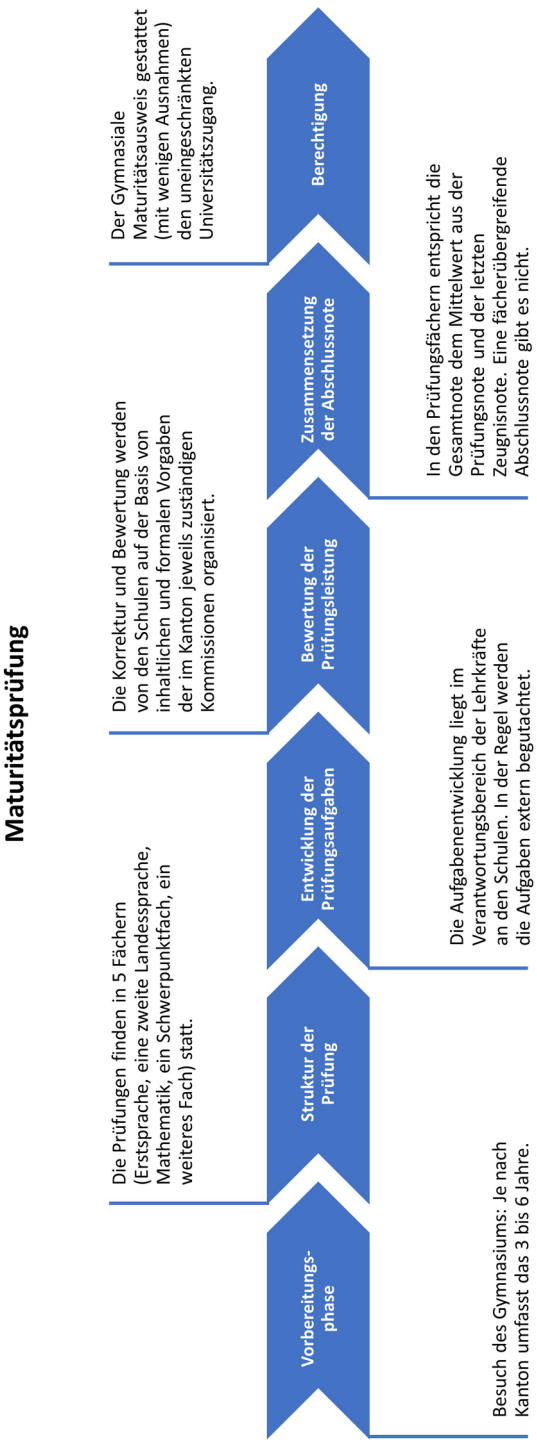
4 Die Maturitätsprüfung in der Schweiz

„Wer lehrt, der prüft."

Franz Eberle

Die Schulen in der Schweiz verfügen über eine hohe Autonomie. Dies betrifft nicht nur die Lehrfreiheit, sondern auch das Prüfungswesen und die Organisation der Maturitätsprüfungen[6]. Entsprechende Vorgaben für die Schulen sind kantonspezifisch geregelt, sodass sich die Ausgestaltung der Abschlussprüfungen sowohl zwischen den Kantonen als auch zwischen den Schulen innerhalb eines Kantons unterscheidet. Als Experte auf gesamtschweizerischer Ebene wurde Franz Eberle befragt, emeritierter Professor für Gymnasial- und Wirtschaftspädagogik an der Universität Zürich und Mitglied der Schweizerischen Maturitätskommission (SMK). Unter anderem war er für die Durchführung der Studie EVAMAR II (2005 bis 2008) verantwortlich, welche die zweite Phase der Evaluation der im Jahr 1995 eingeleiteten Maturitätsreform darstellt. Zudem war er Leiter eines Projekts zur Festlegung basaler fachlicher Kompetenzen für die allgemeine Studierfähigkeit im Fach Mathematik und in der Erstsprache.

6 In Bezug auf die Maturitätsprüfung existieren in der Schweiz von Kanton zu Kanton zum Teil recht unterschiedliche Bestimmungen und Verfahrensweisen. So ist etwa kantonspezifisch geregelt, ob und von wem die von den Schulen entwickelten Prüfungsaufgaben begutachtet werden. Je nach Kanton variiert auch die Art und Anzahl der für die Maturitätsprüfungen zuständigen Gremien und Kommissionen (z. B. der Bildungsrat, die kantonale Maturitätskommisson, die Schulkommissionen). Vor dem Hintergrund dieser Heterogenität werden in der Abbildung und der Tabelle zum Teil eher allgemeine Formulierungen verwendet (z. B. „in der Regel",„die im jew. Kanton zuständigen Kommissionen", „die [für die Begutachtung der Aufgaben zuständigen] Expertengruppen").


Maturitätsprüfung
Vorbereitungs-phase
Besuch des Gymnasiums: Je nach Kanton umfasst das 3 bis 6 Jahre.
Struktur der Prüfung
Die Prüfungen finden in 5 Fächern (Erstsprache, eine zweite Landessprache, Mathematik, ein Schwerpunktfach, ein weiteres Fach) statt.
Entwicklung der Prüfungsaufgaben
Die Aufgabenentwicklung liegt im Verantwortungsbereich der Lehrkräfte an den Schulen. In der Regel werden die Aufgaben extern begutachtet.
Bewertung der Prüfungsleistung
Die Korrektur und Bewertung werden von den Schulen auf der Basis von inhaltlichen und formalen Vorgaben der im Kanton jeweils zuständigen Kommissionen organisiert.
Zusammensetzung der Abschlussnote
In den Prüfungsfächern entspricht die Gesamtnote dem Mittelwert aus der Prüfungsnote und der letzten Zeugnisnote. Eine fächerübergreifende Abschlussnote gibt es nicht.
Berechtigung
Der Gymnasiale Maturitätsausweis gestattet (mit wenigen Ausnahmen) den uneingeschränkten Universitätszugang.

4.1 Überblick über die Organisation und Durchführung

Rahmenbedingungen	
Organisation	Die Steuerung des schulischen Bildungssystems liegt vorwiegend in der Hand der 26 Kantone. Die Organisation der schweizerischen Maturitätsprüfungen und die Begutachtung von Gymnasial- und Maturitätsfragen übernimmt die Schweizerische Maturitätskommission (SMK) – eine rund 25 Expert:innen umfassende, gemeinsame Kommission von Bund (Eidgenössisches Departement für Wirtschaft, Bildung und Forschung, WBF) und Kantonen (Schweizerische Konferenz der Kantonalen Erziehungsdirektoren, EDK). Auf gesamtschweizerischer Ebene geben das MAR (ein Reglement der EDK über die Anerkennung von gymnasialen Maturitätsausweisen) und die MAV (eine vom Bundesrat beschlossene Verordnung über die Anerkennung von gymnasialen Maturitätsausweisen) den rechtlichen Rahmen für die Erlangung von Maturitätsausweisen im Allgemeinen und für Maturitätsprüfungen im Speziellen vor.
Qualifikationsphase	Das Gymnasium ist die Schulart, an der die gymnasiale Maturität erworben wird. In der Regel umfasst das Gymnasium 4 bis 6 Jahre, in einigen Westschweizer Kantonen 3 Jahre. Als curriculare Grundlage dienen die Lehrpläne der Kantone, die sich in vielen Fällen an einem Rahmenlehrplan der EDK orientieren und einen inhaltlichen Rahmen setzen, der den einzelnen Gymnasien große Spielräume für die Schulentwicklung, die Profilbildung und die Unterrichtsgestaltung lässt (vgl. Brüggenbrock et al., 2016).
Vorbereitung	
Prüfungsfächer	Insgesamt gibt es je nach Kanton 13 bis 15 Maturafächer, inkl. der Maturaarbeit, d. h. eine von den Schüler:innen allein oder in einer Gruppe erstellte eigenständige schriftliche Arbeit (vgl. MAR, Art. 10), die mündlich präsentiert werden muss. Die 5 Prüfungsfächer sind die Erstsprache, eine zweite Landessprache oder eine zweite Kantonssprache, Mathematik, ein vom Prüfling gewähltes Schwerpunktfach, und ein weiteres Fach, für dessen Wahl kantonspezifische Regelungen gelten. Inhaltliche Vorgaben für die gymnasialen Maturitätsprüfungen erfolgen durch die im jeweiligen Kanton zuständige(n) Kommission(en). Vorgaben (z. B. im Sinne von Literaturlisten) auf gesamtschweizerischer Ebene gibt es nur für die Schweizerische Maturitätsprüfung, die zentral gestellt wird und zu der sich jeder/jede Bürger:in anmelden kann.
Aufgabenentwicklung	Die Aufgabenentwicklung erfolgt durch die einzelnen Lehrpersonen an den Schulen. In der Regel werden die entwickelten Aufgaben zur Begutachtung an Expertengruppen übermittelt. Durch „Gemeinsames Prüfen“, das auf eine Harmonisierung von Prozessen innerhalb einer Schule und ggf. auch zwischen Schulen zielt, soll die Vergleichbarkeit erhöht werden (vgl. Holmeier et al., 2017). Die Harmonisierung bezieht sich etwa auf die Festlegung von Inhalten, Kompetenzen und Anforderungen, auf die Entwicklung von Prüfungsaufgaben und auf die Festlegung von Korrekturrichtlinien. In vielen Kantonen inkludiert das „Gemeinsame Prüfen“ auch eine abschließende Qualitätskontrolle der entwickelten Aufgaben durch Lehrpersonen aus der Fachschaft, die nicht an der Aufgabenentwicklung beteiligt waren.
Durchführung der Prüfung	
Art, Anzahl und Format der Prüfungen	Die Prüfungen finden in schriftlicher Form statt. Mündliche Prüfungen können, abhängig vom jeweiligen Kanton, zusätzlich durchgeführt werden. Es werden überwiegend offene Aufgabenformate genutzt, einige Lehrpersonen erstellen aber auch Aufgabensets, die geschlossene Aufgaben (d. h. Multiple-Choice-Aufgaben) beinhalten.
Zeitlicher Umgang und Hilfsmittel	Die Vorgaben zu Dauer und Hilfsmittel erfolgen durch die im jeweiligen Kanton zuständige(n) Kommission(en).

Korrektur und Bewertung der Prüfungsleistung	
Erwartungshorizonte und Bewertungsmuster	Die „Bewertungsschlüssel" werden von Lehrpersonen entwickelt und zumeist extern begutachtet. Dabei müssen die inhaltlichen und formalen Vorgaben durch die im jeweiligen Kanton zuständige(n) Kommission(en) berücksichtigt werden.
Korrektur und Bewertungsverfahren	Die Korrektur und Bewertung der Prüfungsleistungen werden von den Schulen vor dem Hintergrund von inhaltlichen und formalen Vorgaben der im jeweiligen Kanton zuständigen Kommission(en) organisiert.
Prüfungsergebnisse	
Zusammensetzung der Abschlussnote	In den 5 Prüfungsfächern wird die Maturitätsnote zur Hälfte aus den Leistungen im letzten Ausbildungsjahr (= letzte Zeugnisnote) und den Leistungen in den Prüfungen gebildet. In den Nichtprüfungsfächern stellt die letzte Zeugnisnote im betreffenden Fach die Maturitätsnote dar. Eine fächerübergreifende Abschlussnote gibt es nicht.
Berechtigung	Die Leistungen in den Prüfungsfächern werden mit Noten von 1 (schlechteste Note) bis 6 (beste Note) bewertet, wobei es ganze und halbe Noten gibt. Alle Noten unter 4 stehen für ungenügende Leistungen und es existieren differenzierte Regelungen zum Bestehen. Ein Maturitätsausweis kann auch dann erlangt werden, wenn in einigen wenigen Fächern ungenügende Leistungen erzielt wurden. Der gymnasiale Maturitätsausweis gestattet den uneingeschränkten Zugang zu allen universitären Studiengängen. Für einzelne Fächer, wie z. B. Medizin, gibt es an manchen Universitäten zusätzliche Eignungstests.

4.2 Interview mit Franz Eberle

Wie würden Sie die Philosophie Ihres Prüfungssystems beschreiben?

Neben dem Ziel der allgemeinen Studierfähigkeit ist die Vorbereitung auf anspruchsvolle Aufgaben in der Gesellschaft – die „vertiefte Gesellschaftsreife" – das zweite Hauptziel der Maturität. Bei der grundsätzlichen Ausrichtung des Gymnasiums geht es eben nicht nur um die Voraussetzungen dafür, dass man in der Lage ist, irgendein Studium erfolgreich aufnehmen zu können, sondern eben dezidiert auch darum, in der Lage zu sein, sich später in verantwortungsvollen Positionen mit anspruchsvollen Aufgaben in der Gesellschaft sachkompetent befassen und maßgeblich zu deren Lösung beitragen zu können. Und das bedeutet eben, dass das Gymnasium die Legitimation hat, Inhalte zu unterrichten, die nicht unabdingbare Voraussetzung nur für die allgemeine Studierfähigkeit sind, sondern auch darüber hinausgehen. Der gymnasiale Bildungsauftrag geht also im Sinne der Förderung einer vertieften Gesellschaftsreife noch viel weiter als einfach die allgemeine Hochschulreife zu erwerben.

Was sind die zentralen Aspekte?

Das Prüfungssystem in der Schweiz ist sehr stark vom Bestreben nach Autonomie der einzelnen Schulen und der einzelnen Lehrpersonen geprägt. Man hört häufig in den Diskussionen, dass das Prinzip „Wer lehrt, der prüft" gelte, und da ist es einfach so, dass sehr viel der Professionalität der einzelnen Lehrpersonen überlassen ist und man auf deren Einschätzungskompetenzen vertraut. Das ist gymnasiale Kultur. Das heißt auch, dass es Unterschiede in der Ausgestaltung der Prüfungen gibt. Und das bedeutet natürlich, dass die Prüfungen – die Prüfungsanforderungen und Prüfungsergeb-

nisse – nicht immer vergleichbar sind. Wir haben eine ganze Reihe von Rahmenbedingungen, die die Vergleichbarkeit beeinträchtigen. Das beginnt bereits bei der Aufnahme in das Gymnasium. Jeder Kanton hat sein eigenes Aufnahmeverfahren für das Gymnasium. Die Anforderungen sind bereits zwischen den Kantonen nicht vergleichbar. Das spiegelt sich in relativ großen Unterschieden in den unterschiedlichen Aufnahme- und Maturitätsquoten wider, die nicht einfach darauf zurückgeführt werden können, dass die Schülerinnen und Schüler unterschiedlich leistungsfähig sind. Es gibt auch unterschiedliche Curricula in den Kantonen. Der jetzige Rahmenlehrplan wird eigentlich nicht ernst genommen, weil er viel zu unverbindlich ist. Manchmal haben wir sogar innerhalb der Kantone unterschiedliche Stundendotationen an den verschiedenen Schulen. Das ist zum Beispiel im Kanton Zürich der Fall. Dort ist die Schulautonomie im schweizerischen Vergleich am höchsten. Da hat jede Schule ihren eigenen Schullehrplan und ihre eigene Stundenstruktur. Und je nach Schule gibt es eben auch unterschiedliche Überzeugungen von Ansprüchen, Aufgabenschwierigkeiten und Bewertungen von Arbeiten.

Vergleichbar ist also nichts?

Es gibt natürlich immer andere Elemente, die trotzdem zu einer Mindestvergleichbarkeit führen. Da würde ich jetzt nennen, dass es überhaupt Zulassungsquoten zum Gymnasium gibt. Ich denke, das ist dann bereits wieder standardisierend. Wir haben ja quasi die besten Schülerinnen und Schüler im Gymnasium, in unterschiedlichem Ausmaß, und das bedeutet, dass der Unterricht und die Prüfungen in den Anforderungen und Schwierigkeiten nicht beliebig voneinander abweichen. Und dann gibt es ja noch die Lehrmittel, die je nach Fach auch ziemlich standardisierend sind. Zumindest in den Naturwissenschaften gibt es die Standardinstrumente und -werke, die dann weit verbreitet sind. Aber ja, insgesamt ist die Vergleichbarkeit erschwert. Es gibt zwar Zusammenarbeit innerhalb der Fachgemeinschaften und dem Verein Schweizerischer Gymnasiallehrerinnen und -lehrer, aber eine kantonsübergreifende, institutionalisierte Zusammenarbeit gibt es nicht.

Wird dieser Mangel an Vergleichbarkeit diskutiert?

Ja, das wird schon diskutiert. Der Wirtschaftsverband „Economy Suisse“ bringt die Notwendigkeit einer besseren Standardisierung immer wieder aufs Tablett. Es hat mal eine große und breite Diskussion gegeben, ob man am Gymnasium nicht auch Bildungsstandards einführen soll, ob man die Maturitätsprüfungen zentralisieren soll. Das Ergebnis dieser Diskussion war aber, dass sich die Kultur der Autonomie durchgesetzt hat. Dann haben die Ergebnisse unserer Evaluation der Maturitätsreform „EVAMAR II“ auch eine entsprechende Diskussion entfacht, auch medial. Die Feststellung, dass die Kantone mit höheren Maturitätsquoten schlechter abgeschnitten haben in unseren Tests, hat zu Diskussionen geführt. Ebenso, dass es relativ viele erfolgreiche Maturantinnen und Maturanten gibt, die in Mathematik und Erstsprache – d. h. in Bereichen von sogenannten basalen Kompetenzen für allgemeine Studierfähigkeit – relativ schlecht, also eigentlich ungenügend, abgeschnitten haben. Das ist immer noch eine Minderheit, muss man sagen, aber eigentlich eine zu große Minder-

heit, die in wichtigen fachlichen Kompetenzen ungenügende Voraussetzungen mitbringen für ein Studium. Das hat dann zur Diskussion geführt, dass wenigstens in den Bereichen von basalen fachlichen Kompetenzen für allgemeine Studierfähigkeit alle Maturantinnen und Maturanten mindestens genügend abschneiden, also dort keine Kompensationsmöglichkeiten mehr haben sollten.

Haben sich Konsequenzen aus dieser Diskussion ergeben?

Ja, es hat zu einem Folgeprojekt geführt, das auch an meinem Lehrstuhl geleitet wurde: „Ermittlung der basalen fachlichen Kompetenzen für allgemeine Studierfähigkeit in Erstsprache und Mathematik". Dort haben wir untersucht, in welchen Studienfächern man welches Wissen und Können in Mathematik und Erstsprache mitbringen muss. Dass es nur Mathematik und Erstsprache waren, war ein politischer Entscheid. Es war eine Diskussion, ob man in diesen Bereichen zentrale Tests machen soll. Aber das wurde am Schluss verworfen. Die Widerstände und Ängste gegen all die Nachteile vom Zentralisieren hatten die Oberhand hier. Aus dem Projekt resultierte ein Anhang zum Rahmenlehrplan, wo diese basalen Kompetenzen in Erstsprache und Mathematik für allgemeine Studierfähigkeit festgehalten sind. Und die Kantone sind jetzt gehalten, sicherzustellen, dass die Schulen bei diesem Wissen und Können ein mindestens genügendes Erreichen sicherstellen. Aber das ist eine relativ softe Vorgabe. Es laufen jetzt viele Projekte, die auch föderal integriert sind in den verschiedenen Kantonen. Und das ergibt natürlich dann bei diesen Kompetenzen eine bessere Vergleichbarkeit.

Gibt es weitere Herausforderungen?

Die Ergebnisse von „EVAMAR II", die zeigen, dass die Maturitätsprüfungen zum Teil unterschiedlich schwierig sind, je nach Kanton und je nach Schule, haben dazu geführt, dass diese Empfehlungen für Gemeinsames Prüfen dann daraus resultiert sind. Da hat man auch alles diskutiert: Soll man jetzt doch standardisieren, um bessere Vergleichbarkeit zu haben? Und wenn ja, auf welcher Ebene soll man ansetzen? Und das Resultat waren die Empfehlungen zum Gemeinsamen Prüfen, die eigentlich schulentwicklungsorientiert zu verstehen sind. Die Idee ist, dass man sich innerhalb von Fachschaften, in der Schule, zwischen verschiedenen Schulen, im Idealfall vielleicht sogar zwischen verschiedenen Kantonen auf gemeinsame qualitative Standards einigt und sich in der Diskussion eine gemeinsame Kultur des Verständnisses von gleichen Anforderungen und Bewertungen etabliert. Das wurde 2016 beschlossen und die Kantone mussten das entsprechend den Vorgaben umsetzen. Jeder Kanton hat das ein bisschen anders gelöst. Das ist ein Entwicklungsprojekt, das überhaupt noch nicht abgeschlossen ist. Aber es ist die Alternative zu zentralen Maturitätsprüfungen. Die Erkenntnis, dass man den Mangel an Vergleichbarkeit und die entsprechenden Ungerechtigkeiten, die dadurch verursacht werden, beheben muss, ist bei den Gymnasiallehrkräften schon überwiegend gewachsen. Und die Einsicht, dass es besser ist, von unten her aktiv zu werden, wenn man die Keule mit den standardisierten Prüfungen von oben vermeiden möchte, hat sich jetzt auch verbreitet. Also in meiner Wahrnehmung gibt es viele gute Entwicklungen in diese Richtung.

Welche zukünftigen Entwicklungen sehen Sie?

Im Hintergrund steht immer, dass man bald einmal ein sogenanntes „EVAMAR III" auf die Schiene bringen möchte, wo wieder gesamtschweizerisch die Kompetenzen erhoben und miteinander verglichen werden sollen. Aber das wird immer ein bisschen weiter hinausgeschoben. Was man immer wieder diskutiert, sind die Bestehensbedingungen. Bei uns ist die schlechteste Note 1 und die beste Note 6. Note 4 ist gerade noch genügend. Laut Kompensationsmodell darf man vier ungenügende Noten [d. h. Noten zwischen 1 und 3,5] haben. Jede Abweichung nach unten muss durch eine andere Fachnote nach oben doppelt kompensiert werden. Also mit einer 3 muss ich zum Beispiel in einem anderen Fach eine 6 haben oder in zwei anderen Fächern eine 5. Es gab immer wieder Diskussionen darüber, ob man diese Bestehensnormen ändern soll. Im Ergebnis von „EVAMAR II" hat man ja gesehen, dass es bei diesen basalen Kompetenzen eigentlich dringenden Handlungsbedarf gibt. Und aufgrund dieser Diskussionen, die sich jetzt über viele Jahre hinweg gezogen haben, hat man vor zwei Jahren beschlossen, mal eine Auslegeordnung[7] über die gymnasiale Maturität zu machen und zu schauen, was man alles besser machen könnte. Es war auch wirklich notwendig, dass auch die Westschweizer Kantone, also die Kantone in der Romandie, die Minimaldauer für das Gymnasium auf vier Jahre setzen. Und es war klar, dass es dafür endlich einen neuen Rahmenlehrplan braucht. Daraus sind dann vier Projekte entstanden, die jetzt unter dem Gesamttitel „Weiterentwicklung gymnasialer Maturität" laufen. Das erste Projekt ist ein neuer Rahmenlehrplan. Der soll am 1. August 2024 in Kraft gesetzt werden. Das zweite Projekt heißt „MAR/MAV", also die Überprüfung einzelner Artikel des Maturitätsanerkennungsreglements (MAR) und der -verordnung (MAV). Die Hauptaussage aus der Auslegeordnung war die, dass es eine Weiterentwicklung des Gymnasiums braucht, aber keine Revolution. Hier sind allerdings umfassende Vorschläge entstanden. Zum Beispiel hat man die curriculare Fächerstruktur unter die Lupe genommen und wie man die ändern könnte. Welche Fächer sollen obligatorisch bleiben aus dem Grundlagenbereich? Wie will man den Wahlpflichtbereich neu gestalten? Wäre es möglich, dass man eine Stufenordnung macht? Also die ersten zwei Jahre Grundlagen und dann die Jahre 3 und 4 des Gymnasiums als Vertiefungsstufe? Und da sind wir jetzt mittendrin. Da hat es eine erste Konsultation gegeben und die Ergebnisse werden jetzt ausgewertet. Das dritte Projekt ist „Governance". Aufgrund der Tatsache, dass die SMK keine periodische Überprüfung macht, müsste man in der überkantonalen Governance was ändern. Ein deklariertes Ziel dieses ganzen Projekts ist bessere Vergleichbarkeit. Und dann gibt es noch ein ganz kleines Projekt zur Erhöhung der Mindestdauer des Gymnasiums auf vier Jahre überall. Aber wichtig und spannend sind vor allem die Projekte „Rahmenlehrplan" und „MAR/MAV".

7 Schweizerisch, „dem (ersten, grundlegenden) Überblick dienende Zusammenstellung, zusammenfassende Darstellung aller relevanten Themenbereiche, Arbeitsergebnisse o. Ä." (https://www.duden.de/rechtschreibung/Auslegeordnung [24.02.2022]).

Was, glauben Sie, wird sich in 10 Jahren geändert haben?

Spätestens in 10 Jahren wird man eine erneute empirische Evaluation durchführen. Und dann wird man prüfen, ob diese basalen fachlichen Kompetenzen für allgemeine Studierfähigkeit tatsächlich in mindestens genügendem Ausmaß von allen Maturantinnen und Maturanten erreicht werden. Das ist das Ziel. Und es gibt viele Konzepte, die diese basalen Kompetenzen gezielter und bewusster auch in den anderen Fächern fördern möchten. Zweitens wird man prüfen, ob durch dieses Gemeinsame Prüfen die Anforderungen und Bewertungen auf hohem Niveau vergleichbarer geworden sind. Ich selbst gebe der Kraft der Weiterentwicklung von unten – also bottom-up – eine große Chance, aber ich bin noch nicht ganz sicher, ob es tatsächlich funktioniert. Im neuen Rahmenlehrplan gibt es auch ein extra Kapitel für transversale Bereiche. Das umfasst die Förderung von überfachlichen Kompetenzen, Interdisziplinarität, überfachlicher Wissenschaftspropädeutik, Digitalität und der Bildung für eine nachhaltige Entwicklung. Ich würde jetzt prognostizieren, wir haben im Jahr 2030 immer noch das Fächerprinzip, aber mit einer besseren Integration dieser Querbereiche. Was gleich bleiben wird – und das wird hochgehalten in der Schweiz – ist das Ziel der allgemeinen Studierfähigkeit, verknüpft mit dem grundsätzlich prüfungs- und numerus-clausus-freien Zugang zu allen universitären Studienfächern und zu den pädagogischen Hochschulen. Das ist mehrfach als politisches Ziel in den letzten Jahren bestätigt worden.

Empfehlungen zur vertiefenden Lektüre

Auf den Internetseiten der SMK finden sich unter https://www.sbfi.admin.ch/sbfi/de/home/bildung/maturitaet.html weiterführende Informationen zur Maturitätsprüfung; u. a. können dort auch die Maturitätsanerkennungsreglements (MAR) sowie die -verordnung (MAV) eingesehen werden.

Eine wissenschaftliche Auseinandersetzung mit unterschiedlichen Aspekten zur Sekundarstufe II in der Schweiz und der Maturitätsprüfung findet sich u. a. in folgenden Publikationen:

Brüggenbrock, C., Eberle, F. & Oelkers, J. (2016). Die jüngeren Entwicklungen des Gymnasiums und der Matura in der Schweiz. In J. Kramer, M. Neumann & U. Trautwein (Hrsg.), *Abitur und Matura im Wandel. Historische Entwicklungslinien, aktuelle Reformen und ihre Effekte* (S. 59–80). Wiesbaden: Springer.

Eberle, F. (2018). Die Maturitätsreform 1995. Intention, Evaluation der Wirkung und Anpassungsmassnahmen. In F. Imlig, L. Lehmann & K. Manz (Hrsg.), *Schule und Reform. Veränderungsabsichten, Wandel und Folgeprobleme*. Reihe Educational Governance (S. 213–227). Wiesbaden: Springer.

Eberle, F., Brüggenbrock, Ch., Rüede, Ch., Weber, Ch. & Albrecht, U. (2015). *Basale fachliche Kompetenzen für allgemeine Studierfähigkeit in Mathematik und Erstsprache*. Kurzbericht zuhanden der EDK. Zürich: Eigenverlag.

Literatur

Ackeren, I. van, Block, R., Klein, E. D. & Kühn, S. M. (2012). The impact of statewide exit exams: A descriptive case study of three German states with differing low stakes exam regimes. *Education Policy Analysis Archives, 20/8*, 1–28.

Aktionsrat Bildung (2011). *Gemeinsames Kernabitur. Zur Sicherung von nationalen Bildungsstandards und fairem Hochschulzugang*. Münster: Waxmann.

Holmeier, M., Maag Merki, K. & Hirt, C. (2017). *Gemeinsames Prüfen*. Wiesbaden: Springer.

Hoymann, T. (2005). *Umdenken nach dem PISA-Schock. Das gesamtdeutsche Zentralabitur als Motor für den Wettbewerb im Bildungsföderalismus*. Marburg: Tectum.

Klein, E. D. & Ackeren, I. van (2012). Challenges and Problems for Research in the Field of Statewide Exams. A Stock Taking of Differing Procedures and Standardization Levels. *Studies in Educational Evaluation, 37*(4), 180–188.

Klein, E. D., Kühn, S. M., Ackeren, I. V. & 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.

UNESCO – United Nations Educational, Scientific and Cultural Organization. (2006). International Standard Classification of Education, ISCED 1997 (S. 195–220). http://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-1997-en_0.pdf

Teil II:
Empirische Analysen

5 Evaluation der „Gemeinsamen Abituraufgabenpools der Länder“

Lars Hoffmann, Pauline Schröter & Petra Stanat

Zusammenfassung

Seit dem Prüfungsjahr 2017 stehen den Ländern in den Fächern Deutsch, Englisch, Französisch und Mathematik gemeinsame Abituraufgabenpools zur Verfügung, aus denen sie Prüfungsaufgaben entnehmen und in ihren schriftlichen Abiturprüfungen einsetzen können. Das Institut zur Qualitätsentwicklung im Bildungswesen hat den erstmaligen Einsatz von Aufgaben der Pools in den Abiturprüfungen der Länder empirisch evaluiert und diese Evaluation seither kontinuierlich fortgeführt. Der Schwerpunkt dieser Evaluation, deren Anlage, Methodik und Ergebnisse im vorliegenden Kapitel ausführlich erläutert werden, liegt auf der Frage, wie sich die aus den Pools entnommenen Aufgaben („Poolaufgaben“) im Vergleich zu den übrigen Abituraufgaben der Länder („landeseigene Aufgaben“) in Bezug auf verschiedene Evaluationskriterien bewähren. Die im Rahmen der Evaluation für die Prüfungsjahre 2017 und 2019 festgestellten Ergebnisse verdeutlichen, dass in den meisten Fällen nur geringe Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben bestehen. Eine Gegenüberstellung der Ergebnisse für beide Prüfungsjahre lässt zudem auf eine positive Entwicklung schließen, da sich die für das Jahr 2017 ohnehin schon geringe Anzahl an festgestellten Auffälligkeiten im Jahr 2019 nochmals deutlich reduzierte. Der Beitrag schließt mit einem Ausblick auf zukünftige Herausforderungen der Evaluation.

1 Einleitung

Im Oktober des Jahres 2012 hat die Kultusministerkonferenz (KMK) in ihrer 339. Sitzung die Bildungsstandards für die Allgemeine Hochschulreife in den Fächern Deutsch, Mathematik und fortgeführte Fremdsprache (Englisch/Französisch) verabschiedet (KMK, 2015; Stanat et al., 2016). Für die genannten Fächer war bereits im März desselben Jahres im Rahmen der 337. Sitzung der KMK der Aufbau eines gemeinsamen Abituraufgabenpools der Länder beschlossen worden, mit dessen Koordination das Institut zur Qualitätsentwicklung im Bildungswesen (IQB) betraut wurde.[1] In der Folge entwickelte das IQB in Abstimmung mit den Ländern zunächst eine „Konzeption für die Entwicklung und Nutzung von Abituraufgabenpools“, die in der

1 Auf die Entwicklung der Bildungsstandards für die Allgemeine Hochschulreife und den Aufbau des Gemeinsamen Abituraufgabenpools der Länder wird von Hoffmann, Schröter und Stanat in Beitrag 2 des vorliegenden Bandes genauer eingegangen.

342. Sitzung der KMK im Juni des Jahres 2013 als grundsätzlicher Verfahrensvorschlag verabschiedet wurde (Stanat & Pant, 2013). Diese Konzeption sah unter anderem eine empirische Fundierung der Entwicklung und Nutzung der Pools durch das IQB vor. Entsprechend hat das IQB den erstmaligen Einsatz von Aufgaben der gemeinsamen Abituraufgabenpools in den Abiturprüfungen der Länder im Prüfungsjahr 2017 empirisch evaluiert und diese Evaluation seither kontinuierlich fortgeführt (Hoffmann et al., 2018, 2020). Das vorliegende Kapitel gibt einen Überblick über die Anlage und Methodik dieser Evaluation und informiert über die zentralen Ergebnisse. Es schließt mit einem Ausblick auf zukünftige Herausforderungen der Evaluation.

2 Anlage und Methodik der Evaluation

2.1 Funktionen und Ziele der Evaluation

Die vom IQB durchgeführte Evaluation des Einsatzes von Aufgaben der Gemeinsamen Abituraufgabenpools in den Abiturprüfungen der Länder basiert auf einem Evaluationskonzept, das im Jahr 2017 vom IQB in Abstimmung mit den Ländern erarbeitet wurde. Dieses Konzept wurde seither stetig weiterentwickelt. Die Ausführungen in diesem Abschnitt beziehen sich auf die aktuelle Fassung des Konzepts, das aus dem Jahr 2021 stammt. Diese Fassung ist zwar für den Zeitraum von 2022 bis 2024 konzipiert, führt jedoch die wesentlichen Elemente der vorherigen Fassungen des Evaluationskonzepts fort.

In der Fachliteratur wird der Begriff der wissenschaftlichen Evaluation bzw. der Evaluationsforschung unterschiedlich definiert und verwendet (Wanzer, 2020). Das Konzept zur Evaluation des Einsatzes von Poolaufgaben in den Abiturprüfungen der Länder orientiert sich an dem von Döring und Bortz (2016b) vorgeschlagenen Begriffsverständnis, nach dem Evaluationsforschung „sozialwissenschaftliche Methoden [nutzt], um einen Evaluationsgegenstand (z. B. […] eine Maßnahme) unter Berücksichtigung der relevanten Anspruchsgruppen […] anhand bestimmter Evaluationskriterien […] und Maßgaben zu ihren Ausprägungen zu bewerten“ (ebd., S. 979). Evaluationsstudien lassen sich dabei u. a. hinsichtlich ihrer jeweiligen Funktionen differenzieren, wobei insbesondere die Unterscheidung von formativen und summativen Evaluationen (Scriven, 1972) bzw. von Prozess- und Ergebnisevaluationen gebräuchlich ist (DeGEval, 2017; Speck, 2016).

Formative Evaluationen werden in aller Regel nicht erst nach der finalen Umsetzung einer Maßnahme, sondern bereits parallel zum Prozess ihrer Einführung und Implementierung durchgeführt, weshalb sie manchmal auch als Prozessevaluationen bezeichnet werden. Formative Evaluationen zielen vor allem auf Optimierung ab. Sie richten sich in aller Regel primär an diejenigen Personen, die direkt an der Entwicklung einer Maßnahme beteiligt sind bzw. deren Durchführung verantworten. Diesen Personen sollen sie zum einen Rückmeldungen dazu geben, wie gut die Umsetzung der betreffenden Maßnahme gelingt. Zum anderen sollen sie Informationen bereitstellen, auf deren Grundlage die evaluierte Maßnahme weiterentwickelt und verbes-

sert werden kann. Demgegenüber haben summative Evaluationen den Charakter von Ergebnisevaluationen. Sie werden erst nach der finalen Umsetzung einer Maßnahme oder nach dem Abschluss ihrer Implementierung durchgeführt, sollen deren Resultate bilanzieren und dabei helfen, grundlegende Entscheidungen, wie etwa zur Weiterführung oder Ausweitung der evaluierten Maßnahme, zu treffen (Döring & Bortz, 2016b; DeGEval, 2017; Speck, 2016).

Konzeptuell ist die vom IQB durchgeführte Evaluation des Einsatzes der Poolaufgaben in den Abiturprüfungen der Länder derzeit formativ angelegt, d. h. die in ihrem Rahmen vorgesehenen Forschungstätigkeiten dienen dazu, den im Jahr 2017 begonnenen Implementationsprozess der Pools wissenschaftlich zu begleiten und empirisch fundierte Informationen dazu zu gewinnen, wie gut der Umsetzungsprozess gelingt. Im Mittelpunkt stehen dabei die in den Abituraufgabenpools bereitgestellten Aufgaben. So wird etwa untersucht, inwieweit und mit welchen Modifikationen diese Aufgaben zum Einsatz kommen und inwiefern sie sich in den Abiturprüfungen der Länder bewähren. Darüber hinaus soll die Evaluation die Qualitätssicherung und -entwicklung der Poolaufgaben unterstützen.

Das Evaluationskonzept hat eine Grundstruktur mit zwei Säulen. Die erste Säule bezieht sich auf die Evaluation der Qualität und Eignung der Poolaufgaben im Rahmen eines kontinuierlichen Monitorings. Dieser Teil der Evaluation wird im vorliegenden Kapitel näher beschrieben. Die zweite Säule des Konzepts bezieht sich auf die Durchführung von Evaluationsstudien zur Unterstützung der Qualitätssicherung und -entwicklung der Poolaufgaben. Dieser Teil der Evaluation ist nicht Gegenstand dieses Kapitels, jedoch werden im Beitrag 8 des vorliegenden Bandes zwei empirische Untersuchungen zur Objektivität bei der Bewertung von Abiturarbeiten im Fach Deutsch vorgestellt, die im Rahmen der zweiten Säule des Evaluationskonzeptes durchgeführt wurden.

2.2 Evaluationsbereiche und -kriterien des jährlich durchgeführten Monitorings

2.2.1 Evaluation der Nutzung der Aufgaben aus den Pools

Das kontinuierliche Monitoring begleitet den Einsatz der Poolaufgaben in den Abiturprüfungen der Länder bereits seit dem Prüfungsjahr 2017 und wird seither jährlich durchgeführt. Das Monitoring umfasst dabei die beiden Evaluationsbereiche *Nutzung der Aufgaben aus den Pools* sowie *Bewährung der Aufgaben aus den Pools*. Der erstgenannte Evaluationsbereich untergliedert sich nochmals in die *Evaluation der Entnahme der Aufgaben aus den Pools* und die *Evaluation der Modifikation der entnommenen Aufgaben aus den Pools*.

Evaluation der Entnahme der Aufgaben aus den Pools

Im Rahmen der Evaluation der Entnahme der Aufgaben aus den Pools wird für die Fächer Deutsch, Englisch, Französisch und Mathematik jeweils ermittelt, wie viele Aufgaben mit den Pools des jeweiligen Prüfungsjahres bereitgestellt wurden, wie

viele dieser Aufgaben entnommen wurden und wie viele Länder in ihren Abiturprüfungen Aufgaben aus den Pools eingesetzt haben. Als zentrales Evaluationskriterium für die Entnahme der Aufgaben aus den Pools wird für jedes der vier Fächer berechnet, von wie vielen Ländern die Poolaufgaben im Durchschnitt jeweils gemeinsam genutzt werden. Um die angestrebte Erhöhung der Vergleichbarkeit von Abiturprüfungsanforderungen zu erreichen, sollte sich dieser Quotient sukzessive erhöhen, sodass also letztlich möglichst viele Länder jeweils dieselben Aufgaben aus den Pools entnehmen und in ihren Abiturprüfungen einsetzen.

Evaluation der Modifikation der entnommenen Aufgaben aus den Pools

Wie in Beitrag 2 des vorliegenden Bandes erläutert, haben die Länder bis zu einer im Jahr 2025 endenden Übergangsfrist die Möglichkeit, die von ihnen aus den Pools entnommenen Aufgaben zu modifizieren bzw. an die jeweils in ihrem Land geltenden Regelungen zur Gestaltung von Abiturprüfungsaufgaben anzupassen, wobei die Maßgabe gilt, dass nur unbedingt erforderliche Modifikationen vorgenommen werden sollen. Diese Anpassungen werden im jährlich durchgeführten Monitoring ebenfalls untersucht. Als Evaluationskriterium kommt hierbei eine Klassifikation zum Einsatz, die zwischen den drei Modifikationsgraden „kaum relevante Veränderungen“, „relevante Veränderungen“ und „gravierende Veränderungen“ unterscheidet. Den Ausgangspunkt dieser Einstufung bilden Kategoriensysteme, die in Zusammenarbeit mit den Mitgliedern der für die Entwicklung der Poolaufgaben zuständigen AGs Aufgaben erarbeitet worden sind. In diesen fächerspezifischen AGs ist pro Land jeweils eine Expertin bzw. ein Experte aus der Schulpraxis mit umfangreichen Erfahrungen in der Erstellung von Abiturprüfungsaufgaben vertreten.[2]

Die Kategoriensysteme umfassen für jedes Fach eine überschaubare Anzahl allgemein formulierter Modifikationskategorien, denen die Veränderungen, welche die Länder an den von ihnen aus den Pools entnommenen Aufgaben vornehmen, zugeordnet werden. Auf der Grundlage von Beratungen und Abstimmungen in den AGs Aufgaben wurde festgelegt, welchem der drei Modifikationsgrade die Kategorien des Kategoriensystems jeweils entsprechen. Maßgeblich ist dabei, inwieweit zu erwarten ist, dass die jeweilige Modifikation die Anforderungen bzw. den Schwierigkeitsgrad der Aufgabe verändert. So wurde zum Beispiel im Fach Deutsch vereinbart, dass Modifikationen, die der Kategorie „Kürzung oder Verlängerung des von den Prüflingen zu bearbeitenden Textes“ zuzuordnen sind, jeweils als gravierende Veränderung klassifiziert werden. Demgegenüber gelten etwa Änderungen an der Struktur der den Aufgaben beigefügten Bewertungshinweise, die zum Beispiel in einigen Ländern traditionell (und abweichend von den im Pool zur Verfügung gestellten Aufgaben) als Kriterienraster formuliert sind, nur als relevante Veränderung. Als kaum relevant werden zum Beispiel Modifikationen eingestuft, die nur die Art der Darstellung bzw. das Layout der Erwartungshorizonte und Bewertungshinweise, nicht jedoch deren Inhalte betreffen.

2 Weitere Informationen zur Zusammensetzung der AGs Aufgaben finden sich in Beitrag 2 (Hoffmann, Schröter & Stanat) des vorliegenden Bandes.

In jedem Prüfungsjahr werden die Ländervertreterinnen und -vertreter in den AGs Aufgaben der vier Fächer gebeten, Angaben zu den Modifikationen zu machen, die in ihrem jeweiligen Land an den eingesetzten Poolaufgaben vorgenommen wurden. Dies geschieht auf der Grundlage der skizzierten Kategoriensysteme, d. h. die Ländervertreterinnen und -vertreter informieren nicht nur über die von ihrem Land durchgeführten Änderungen, sondern ordnen diesen auch den betreffenden Modifikationsgrad zu. Darüber hinaus sind sie aufgefordert, jeweils in knapper Form die Gründe der in ihrem Land vorgenommenen Modifikationen darzulegen.

Die von Ländervertreterinnen und -vertretern in den AGs Aufgaben zur Verfügung gestellten Informationen werden vom IQB in einem mehrschrittigen Prozess ausgewertet. Zunächst werden die Angaben zu den vorgenommenen Modifikationen nachvollzogen und dann jeweils einer der drei übergeordneten Kategorien „Veränderungen an den Materialien", „Veränderungen an den Aufgabenstellungen" und „Veränderungen an den Vorgaben zur Bewertung" zugeordnet. Für jede aus den Pools eingesetzte Aufgabe wird also verzeichnet, inwiefern sie im Hinblick auf die genannten drei übergeordneten Kategorien kaum relevant, relevant oder gravierend modifiziert wurde. Hierbei wird auch berücksichtigt, dass sich manche Modifikationen gegenseitig bedingen können (z. B. kann eine Veränderung der Aufgabenstellung eine Veränderung des Erwartungshorizonts erforderlich machen). Es wird angestrebt, insbesondere den Anteil von als gravierend eingestuften Modifikationen sukzessive zu reduzieren und nach dem Ende der genannten Übergangsfrist im Prüfungsjahr 2025 keine Veränderungen mehr vorzunehmen.

Die von den Ländervertreterinnen und -vertretern in den AGs Aufgaben mitgeteilten Gründe für die durchgeführten Modifikationen werden ebenfalls zunächst nachvollzogen und anschließend mittels quantitativer Inhaltsanalysen ausgewertet (Döring & Bortz, 2016a). Dabei wird aus den angegebenen Gründen zunächst ein Kategoriensystem abgeleitet und anschließend ermittelt, wie viele der genannten Gründe den einzelnen Kategorien zugeordnet werden können. Hierbei steht insbesondere im Fokus, wie häufig die Modifikationen mit einer Anpassung der Aufgabe an landesspezifische Regelungen begründet werden. Ferner ist von Interesse, ob und zu welchem Anteil Modifikationen durch fachliche Erwägungen (im Sinne von fachlichen Zweifeln an der Qualität einer Aufgabe oder einzelner Aufgabenaspekte) motiviert sind.

2.2.2 Evaluation der Bewährung der Aufgaben aus den Pools

Datengrundlage

Ebenfalls seit dem Prüfungsjahr 2017 evaluiert das IQB, wie sich die aus den Pools entnommenen Aufgaben in den Abiturprüfungen der Länder bewähren. Dieser Teil der Evaluation basiert maßgeblich auf Daten, die in bundesweiten Schulstichproben erhoben werden. Hierzu wählt das IQB in Zusammenarbeit mit den Ländern in jedem Prüfungsjahr pro Land 20 Schulen aus und erhebt dort pro Fach und Anforderungsniveau, für das vom jeweiligen Land Poolaufgaben entnommen und in der Abi-

turprüfung eingesetzt werden, jeweils Daten zu einem Kurs.[3] In den meisten Ländern kommt hierbei ein vom IQB entwickeltes onlinebasiertes Eingabeinstrument zum Einsatz, das aus zwei Teilen besteht.[4] Im ersten Teil der Eingabemaske werden die Lehrkräfte der für die Evaluation ausgewählten Kurse um anonymisierte Angaben zu ihren Schülerinnen und Schülern gebeten. Insbesondere sollen sie eintragen, welche Ergebnisse die Schülerinnen und Schüler bei einzelnen Prüfungsaufgaben und in der Prüfung insgesamt erzielt haben. Des Weiteren werden zur Erfassung von Vorleistungen Halbjahres- und Klausurnoten erfasst, die die einzelnen Schülerinnen und Schüler des Kurses im Rahmen der Qualifikationsphase, also in den beiden Jahrgangsstufen vor den Abiturprüfungen, erhalten haben. Ferner werden einige wenige Hintergrunddaten zu den Schülerinnen und Schülern (z. B. Geschlecht und Sprachhintergrund) erhoben.

Im zweiten Teil der Eingabemaske werden die Lehrkräfte um Einschätzungen zu den Aufgaben der schriftlichen Abiturprüfungen gebeten, wobei ihnen in der Regel[5] nicht bekannt ist, welche Aufgaben aus den Pools stammen und welche Aufgaben von ihrem Land selbst entwickelt und nur dort eingesetzt wurden (sogenannte „landeseigene Aufgaben“). Der Fokus dieser Lehrkräftebefragung lag bis zum Prüfungsjahr 2021 auf Einschätzungen zum Anspruch der Aufgaben und zum wahrgenommenen Nutzen der zu den Aufgaben bereitgestellten Erwartungshorizonte. Diese Einschätzungen werden anhand von Items auf einer fünfstufigen Likert-Skala vorgenommen. Zudem haben die Lehrkräfte die Möglichkeit, in einem Freitextfeld Kommentare und Hinweise zu den Abiturprüfungsaufgaben zu äußern. Die weiterentwickelte Fassung des Evaluationskonzeptes sieht einen Ausbau der Lehrkräftebefragung ab dem Prüfungsjahr 2022 vor. Es ist geplant, eine größere Anzahl von Lehrkräften zu befragen und neben dem Anspruch der Aufgaben und der Nützlichkeit der Erwartungshorizonte weitere Aspekte der Aufgabenqualität einschätzen zu lassen. Diese Aspekte betreffen etwa Einschätzungen zur Eignung der Situierung von Aufgabenstellungen[6] (insb. bei Aufgaben im Fach Deutsch), zur inhaltlichen Differenzierung von Teilaufgaben (insb. bei Aufgaben des Kompetenzbereiches Schreiben in den Fächern Englisch und Französisch) oder zur Angemessenheit der zur Verfügung stehenden Bearbeitungszeit (insb. bei Aufgaben im Fach Mathematik).

Ebenfalls ab dem Prüfungsjahr 2022 sollen im Rahmen des kontinuierlichen Monitorings auch die in den Ländern für die Abiturprüfungsaufgaben zuständigen Abiturkommissionen um eine Beurteilung der Qualität der jeweils von ihrem Land eingesetzten Prüfungsaufgaben gebeten werden.

3 Die Auswahl der Schulen erfolgt dabei mit einem anonymisierten Verfahren, d. h., das IQB hat keine Kenntnis davon, welche konkreten Schulen an der Evaluation teilnehmen.

4 Einige wenige Länder verzichten ganz oder in Teilen auf die Verwendung des vom IQB entwickelten Eingabeinstruments. Stattdessen führen sie eigene Erhebungen durch, deren Daten sie dem IQB übermitteln.

5 Eine Ausnahme bildet das Land Rheinland-Pfalz, in dem nur die aus den Pools eingesetzten Prüfungsaufgaben zentral gestellt und die übrigen Prüfungsaufgaben von den Fachlehrkräften der jeweiligen Schulen entwickelt werden. Den Lehrkräften ist mithin bekannt, welche Prüfungsaufgaben aus den Pools stammen und welche nicht.

6 Bei einer Situierung handelt es sich, vereinfacht ausgedrückt, um die Einbettung einer Aufgabenstellung in einen authentischen Kontext. Im Fach Deutsch wird beispielsweise in manchen Aufgaben angegeben, an wen der im Rahmen der Aufgabenstellung zu verfassende Text zu adressieren ist (z. B: Leserinnen und Leser eine Schülerzeitung) und welche Funktionen dieser Text erfüllen sollte (z. B. Informieren).

Evaluationskriterien

Wie in Beitrag 2 des vorliegenden Bandes beschrieben, bestehen die schriftlichen Abiturprüfungen der Länder in den Fächern Deutsch, Mathematik, Englisch und Französisch seit dem Prüfungsjahr 2017 in den meisten Fällen sowohl aus landeseigenen Aufgaben als auch aus Aufgaben, die aus den Pools entnommen wurden. Vor diesem Hintergrund wird im Rahmen des kontinuierlichen Monitorings untersucht, wie sich die Poolaufgaben im Vergleich zu den landeseigenen Aufgaben bewähren. Dieser Vergleich erfolgt bezogen auf sechs Kriterien:

(1) *Auswahl der Aufgaben*: In vielen Ländern haben die Prüflinge die Möglichkeit, in den schriftlichen Abiturprüfungen zwischen verschiedenen Aufgaben zu wählen. Die schriftlichen Abiturprüfungen im Fach Deutsch sind dabei so strukturiert, dass die Prüflinge aus einem je nach Land und Anforderungsniveau zwei bis fünf Prüfungsaufgaben umfassenden Aufgabenset eine Aufgabe zur Bearbeitung auswählen. Mithin entscheiden sie sich also implizit, d. h. ohne Kenntnis der Herkunft der Aufgaben, für oder gegen die Bearbeitung einer aus dem Pool stammenden Aufgabe. Vor diesem Hintergrund wird für das Fach Deutsch ermittelt, zu welchem Anteil die Prüflinge Aufgaben aus den Pools auswählen. Zudem wird untersucht, ob die Poolaufgaben vorrangig von bestimmten Gruppen von Schülerinnen und Schülern gewählt werden oder ob die Auswahl weitgehend unabhängig von Hintergrundmerkmalen der Prüflinge (z. B. Vorleistungen in der Qualifikationsphase, Geschlecht) erfolgt. In vielen Ländern haben die Prüflinge auch in den Abiturprüfungen der Fächer Englisch, Französisch und Mathematik die Möglichkeit, zwischen verschiedenen Aufgaben zu wählen. Letztendlich bearbeiten sie zumeist ein Aufgabenset, das sowohl aus dem Pool stammende Aufgaben als auch landeseigene Aufgaben umfasst. Die Analysen erfolgen für die einzelnen Länder getrennt, da sich die Art der Wahlmöglichkeiten zwischen den Ländern unterscheidet und länderübergreifende Analysen daher nicht möglich sind.

(2) *Empirische Schwierigkeit der Aufgaben:* Weiteren Aufschluss darüber, ob sich die Aufgaben der Pools bewährt haben, gibt die empirische Schwierigkeit der Aufgaben, die anhand der bei den Aufgaben erzielten Lösungsquoten bestimmt wird.[7] In Bezug auf dieses Kriterium wird für alle vier Fächer untersucht, wie erfolgreich die Poolaufgaben bearbeitet werden und inwiefern sich Unterschiede in den für die Poolaufgaben und die landeseigenen Aufgaben ermittelten Lösungsquoten feststellen lassen.

(3) *Trennschärfe der Aufgaben:* Die Trennschärfe einer Aufgabe wird über die Korrelation der jeweils bei dieser Aufgabe erzielten Ergebnisse mit den Gesamtergebnissen der Abiturprüfung berechnet. Das Kriterium der Aufgabentrennschärfe bil-

7 Je nach Land und Fach werden die bei einer Aufgabe erzielten Ergebnisse unterschiedlich erfasst, und zwar entweder in Form von Notenpunkten (NP) oder in Form von Bewertungseinheiten (BE), wobei hierfür in einigen Ländern auch andere Termini gebräuchlich sind. Ein wesentlicher Unterschied ist hierbei, dass NP auf einer Skala mit einem fixen Maximum erfasst werden (15 NP), während BE anhand einer Skala vergeben werden, deren Maximum (d. h. die Anzahl der maximal erreichbaren BE) je nach Aufgabe variiert. Um die in BE erfassten Ergebnisse über Aufgaben hinweg miteinander vergleichen zu können, erfolgt im Rahmen der Evaluation eine Umrechnung der jeweils erzielten BE in Lösungsquoten, die als Quotient aus der bei einer Aufgabe erreichten und den maximal bei dieser Aufgabe erreichbaren BE ermittelt werden.

det also ab, wie gut die bei einer bestimmten Aufgabe erzielten Ergebnisse mit dem Gesamtergebnis der Prüfung korrespondieren. Das Kriterium der Trennschärfe ist nur für die schriftlichen Abiturprüfungen in den Fächern Mathematik, Englisch und Französisch sinnvoll bestimmbar, da die Prüflinge hier, wie bereits erläutert, nicht nur – wie im Fach Deutsch – eine einzelne Aufgabe, sondern ein ganzes Set von Aufgaben bearbeiten müssen. Im Zuge der Evaluation wird für die drei genannten Fächer untersucht, inwiefern Poolaufgaben und landeseigene Aufgaben hinsichtlich ihrer Trennschärfe differieren. Substanzielle Unterschiede würden dabei darauf hinweisen, dass Poolaufgaben und landeseigene Aufgaben jeweils unterschiedliche Facetten erfassen. Eine solche Diskrepanz kann entweder prüfungsdidaktisch intendiert und mithin explizit beabsichtigt sein, oder auf nichtintendierte Faktoren (z. B. missverständlich formulierte Arbeitsaufträge, widersprüchliche Angaben im Erwartungshorizont, extrem leichte oder extrem schwierige Aufgaben) zurückgehen. Auffälligkeiten, die in Bezug auf das Evaluationskriterium der Trennschärfe identifiziert werden, weisen also darauf hin, dass eine vertiefende inhaltliche Analyse der Prüfungsaufgaben und ihrer Zusammenstellung vorgenommen werden sollte.

(4) *Kriteriale Validität der Aufgaben in Bezug auf Vorleistungen:* Als Indikator für die kriteriale Validität der Aufgaben aus den Pools wird untersucht, wie hoch die bei den Prüfungsaufgaben erreichten Ergebnisse mit den zuvor in der Qualifikationsphase erzielten Leistungen korrelieren. Der Indikator bildet also ab, inwieweit die bei einer bestimmten Aufgabe erzielten Ergebnisse mit den in den Halbjahren der gymnasialen Oberstufe erreichten Vorleistungen korrespondieren. Zur Berechnung dieses Kriteriums werden die Ergebnisse der Abiturprüfung sowohl mit den Klausurnoten als auch mit den Halbjahresnoten aus der Qualifikationsphase in Bezug gesetzt. Somit wird sowohl der Zusammenhang mit punktuellen, ausschließlich schriftlich erzielten Vorleistungen (Klausuren) untersucht als auch der Zusammenhang mit Vorleistungen, die über einen längeren Zeitraum mündlich und schriftlich erbracht wurden (Halbjahresnoten). Für alle vier Fächer wird im Zuge der Evaluation geprüft, inwiefern sich Poolaufgaben und landeseigene Aufgaben hinsichtlich der Höhe der für sie berechneten Validitätskoeffizienten unterscheiden. Geringere Werte lassen sich hierbei als ein Indiz dafür interpretieren, dass die für die Lösung der betreffenden Aufgabe erforderlichen Kompetenzen im Unterricht der Qualifikationsphase offenbar weniger umfassend bzw. vertiefend entwickelt wurden.

(5) *Lehrkräfteeinschätzungen zu den Aufgaben:* Zusätzlich zu den Informationen, die sich auf Schülerinnen und Schüler beziehen, wird für alle vier Fächer ermittelt, wie die Lehrkräfte den Anspruch der Prüfungsaufgaben beurteilen. Im Fokus steht dabei, wie die Einschätzungen zum Anspruch der Aufgaben aus den Pools im Vergleich zu den landeseigenen Aufgaben ausfallen. Wie bereits oben erwähnt, wird die im Rahmen der Evaluation durchgeführte Lehrkräftebefragung ab dem Prüfungsjahr 2022 Einschätzungen zu weiteren Aspekten der Aufgabenqualität umfassen.

(6) *Lehrkräfteeinschätzungen zu den Erwartungshorizonten:* Im Rahmen der Bewertung von Prüfungsarbeiten wird für alle vier Fächer ermittelt, wie die Lehrkräfte die Nützlichkeit der zu den Aufgaben bereitgestellten Erwartungshorizonte beurteilen. Im Fokus steht dabei auch hier, wie die Einschätzungen für die Aufgaben aus den Pools im Vergleich zu den Einschätzungen für die landeseigenen Aufgaben ausfallen.

2.3 Auswertung des jährlich durchgeführten Monitorings

Wie eingangs erläutert, verfolgt die vom IQB durchgeführte Evaluation der Gemeinsamen Abiturprüfungspools der Länder zum gegenwärtigen Zeitpunkt primär formative Ziele. Dies gilt in besonderem Maße für das jährlich durchgeführte Monitoring, das sich an einen begrenzten Adressatenkreis wendet. Er umfasst erstens die direkt am Projekt Beteiligten und dabei insbesondere die für die Entwicklung der Poolaufgaben zuständigen Mitglieder der AGs Aufgaben, zweitens die in den Ländern für die gymnasiale Oberstufe und die Abiturprüfungen zuständigen Personen sowie drittens die verantwortlichen politischen Entscheidungsträger der Länder. Von Bedeutung ist dabei, dass für die genannten drei Adressatengruppen unterschiedliche Evaluationsergebnisse relevant sind.

Die Mitglieder der AGs Aufgaben benötigen möglichst spezifische und detaillierte Informationen dazu, welche Ergebnisse für jede einzelne von ihnen entwickelte Aufgabe im Hinblick auf die oben skizzierten Evaluationskriterien erzielt wurden. Nur aus solchen spezifischen Rückmeldungen ist es sinnvoll möglich, Schlussfolgerungen für die zukünftige Aufgabenentwicklung abzuleiten. Im Unterschied dazu ist für die in den einzelnen Ländern für die Abiturprüfungen zuständigen Personen vor allem relevant, wie sich die eingesetzten Poolaufgaben in den Prüfungen ihres jeweiligen Landes bewährt haben. Der Fokus liegt hierbei insbesondere auf der Frage, inwiefern in Bezug auf die Ergebnisse der Prüflinge und die Einschätzungen der Lehrkräfte bedeutsame Unterschiede zwischen den Poolaufgaben und den landeseigenen Aufgaben festzustellen sind. Diese Informationen sind auch für die Bildungspolitik relevant. Für die politischen Entscheidungsträger auf Ebene der KMK ist zudem von Interesse, wie gut der Implementationsprozess der Pools insgesamt voranschreitet. Dementsprechend benötigen sie auch aggregierte, d. h. über die einzelnen Länder und ggf. über mehrere Prüfungsjahre zusammengefasste Rückmeldungen zur Nutzung, Modifikation und Bewährung der Poolaufgaben.

Im Rahmen der Auswertung der jährlichen Monitorings wird den jeweiligen Interessen der drei genannten Adressatengruppen Rechnung getragen, indem die erhobenen Evaluationsdaten in unterschiedlicher Weise und jeweils adressatengerecht aufbereitet werden. So wird den Mitgliedern der AGs Aufgaben auf Anfrage zu allen im betreffenden Prüfungsjahr eingesetzten Poolaufgaben jeweils ein Dossier zur Verfügung gestellt, aus dem hervorgeht, ob die betreffende Aufgabe unverändert oder modifiziert eingesetzt wurde und inwiefern in Bezug auf die Evaluationskriterien zur Bewährung Auffälligkeiten festgestellt wurden. Sollte die betreffende Aufgabe zudem Gegenstand von Freitextäußerungen der befragten Lehrkräfte gewesen sein, werden

auch diese Rückmeldungen dem Dossier beigefügt. Zur Unterstützung der Qualitätssicherung und -entwicklung sollen die zur Verfügung gestellten Informationen in den AGs Aufgaben einen Austausch dazu anstoßen, wie sich die entwickelten Aufgaben bewährt haben und hinsichtlich welcher Aspekte es gegebenenfalls noch Gestaltungsbedarfe gibt.

Zudem verfasst das IQB für die einzelnen Länder Berichte mit den jeweiligen landesspezifischen Evaluationsergebnissen. Diese Berichte sind insbesondere an die in den Ländern für die Abiturprüfungen zuständigen Personen adressiert. Sie enthalten Informationen dazu, wie die vom betreffenden Land an den Poolaufgaben durchgeführten Modifikationen einzustufen sind. Auch geben sie einen Überblick, wie sich die Poolaufgaben in den Abiturprüfungen des betreffenden Landes bewährt haben.

Darüber hinaus erstellt das IQB in jedem Jahr einen primär an die politischen Entscheidungsträger in der KMK adressierten Bericht zur Nutzung, Modifikation und Bewährung der Poolaufgaben in allen Ländern. Dieser Bericht fasst die Evaluationsergebnisse der Länder zusammen und stellt diese in anonymisierter Form dar. So ist aus dem Bericht etwa ablesbar, in welchem Umfang die Poolaufgaben länderübergreifend genutzt und modifiziert werden. Die Zusammenfassung der Evaluationsergebnisse zur Bewährung der Poolaufgaben erfolgt unter Verwendung von metaanalytischen Methoden. Hierbei wird jedes Einzelergebnis (d. h. jeder statistische Kennwert, der für eine Aufgabe aus dem Pool in einem einzelnen Land berechnet wurde) als eine „Studie“ in die Auswertung einbezogen. Zudem werden verschiedene Differenzierungen vorgenommen (z. B. nach Aufgabenarten oder Kompetenzbereichen). Als Ergebnis der Metaanalysen wird jeweils ein über alle Länder, Anforderungsniveaus und Poolaufgaben zusammengefasster Effekt berechnet, der zum Beispiel angibt, ob die Aufgaben aus den Pools (nach Einschätzung der Lehrkräfte) in einem bestimmten Fach insgesamt weniger anspruchsvoll oder anspruchsvoller als die landeseigenen Aufgaben waren. Um dem öffentlichen Interesse an den Gemeinsamen Abituraufgabenpools der Länder Rechnung zu tragen, werden die primär an die politischen Entscheidungsträger adressierten Teile des Berichts, die sich auf die Bewährung der Poolaufgaben beziehen, trotz der gegenwärtig formativen Ausrichtung der Evaluation auf den Internetseiten des IQB veröffentlicht.

3 Überblick über zentrale Ergebnisse der Evaluation der Bewährung der Poolaufgaben

Die Evaluation der Bewährung der Poolaufgaben erfolgte erstmals im Prüfungsjahr 2017 und dabei zunächst für alle vier Fächer. Nach einer umfassenden Auswertung dieser Daten, die im Jahr 2018 erfolgte, wurde für die folgenden Prüfungsjahre eine alternierende Untersuchung für die sprachlichen Fächer und das Fach Mathematik beschlossen. Entsprechend wurde im Prüfungsjahr 2019 zunächst die Bewährung der Poolaufgaben in den Fächern Deutsch, Englisch und Französisch evaluiert. Im Prüfungsjahr 2020 sollte dann eine Untersuchung der Bewährung der Poolaufgaben im

Fach Mathematik erfolgen, die aufgrund der Corona-Pandemie jedoch abgebrochen werden musste und auf das Prüfungsjahr 2021 verschoben wurde. Da die Analysen dieser Daten und die darauf bezogene Berichtlegung noch nicht abgeschlossen ist, basieren die in diesem Kapitel berichteten Evaluationsergebnisse zur Bewährung der Aufgaben aus den Pools auf den Ergebnissen für die Prüfungsjahre 2017 und 2019. Tabelle 1 gibt einen Überblick über die Anzahl der in den Evaluationsstudien für die Prüfungsjahre 2017 und 2019 einbezogenen Prüfungsarbeiten.

Tabelle 1: Anzahl der in den Evaluationsstudien für die Prüfungsjahre 2017 und 2019 in den Fächern Deutsch, Englisch, Französisch und Mathematik einbezogenen Prüfungsarbeiten

	Deutsch	Englisch	Französisch	Mathematik
Prüfungsjahr 2017	5167	6602	1852	8094
Prüfungsjahr 2019	6903	8589	1346	–

Nachfolgend werden die wichtigsten länder- und aufgabenübergreifenden Ergebnisse der Evaluation zur Bewährung der Poolaufgaben zusammengefasst. Der Fokus der Darstellung liegt dabei auf den festgestellten fachspezifischen Auffälligkeiten, aus denen wichtige Impulse für die weitere Aufgabenentwicklung abgeleitet werden konnten.

3.1 Ergebnisse der Evaluation zur Bewährung der Poolaufgaben im Fach Deutsch

Im Rahmen der Evaluationsstudie zur Bewährung der Poolaufgaben im Prüfungsjahr 2017 wurde für das Fach Deutsch festgestellt, dass sich die von den Ländern eingesetzten Poolaufgaben in Bezug auf die empirische Schwierigkeit insgesamt betrachtet statistisch signifikant von den landeseigenen Aufgaben unterschieden. So erzielten Prüflinge, die eine Poolaufgabe zur Bearbeitung ausgewählt hatten, ein im Mittel um rund 0.4 Notenpunkte schwächeres Prüfungsergebnis als Schülerinnen und Schüler, deren Wahl auf eine landeseigene Aufgabe fiel. Dieser Unterschied verschwand allerdings, wenn im Rahmen der Analysen zusätzlich für die von den Prüflingen in der Qualifikationsphase erzielten Vorleistungen kontrolliert wurde: Prüflinge mit eher schwachen Vorleistungen hatten in der Prüfung verstärkt Aufgaben zu pragmatischen Texten gewählt, während Schülerinnen und Schüler, die in der Qualifikationsphase eher gute Leistungen erzielt hatten, mehrheitlich Aufgaben der Art Interpretation literarischer Texte bevorzugten. Diese Aufgabenart war allerdings bei den von den Ländern eingesetzten Poolaufgaben deutlich unterrepräsentiert, da die Länder vornehmlich Aufgaben der Arten Analyse pragmatischer Texte, Erörterung pragmatischer Texte und materialgestütztes Verfassen argumentierender/informierender Texte aus dem Pool entnommen hatten. Die festgestellte Diskrepanz zwischen den bei den Poolaufgaben und bei den landeseigenen Aufgaben erzielten Prüfungsergebnissen war also kein Hinweis darauf, dass die Poolaufgaben besonders schwierig gewesen waren. Der Unterschied konnte vielmehr darauf zurückgeführt werden, dass die Poolaufga-

ben eher von Prüflingen gewählt wurden, die weniger leistungsstark waren und dementsprechend auch weniger gute Prüfungsergebnisse erzielten.

Ein anderes Befundmuster als im Prüfungsjahr 2017 wurde in Bezug auf die empirische Aufgabenschwierigkeit im Rahmen der Evaluationsstudie zum Prüfungsjahr 2019 ermittelt. Insgesamt betrachtet unterschieden sich hier die Poolaufgaben und die landeseigenen Aufgaben nicht statistisch signifikant in den im Mittel erreichten Ergebnissen. Tatsächlich waren die von den Ländern im Prüfungsjahr 2019 entnommenen Poolaufgaben anders zusammengesetzt als noch im Prüfungsjahr 2017. Sie umfassten nun einen höheren Anteil an Aufgaben zu literarischen Texten und wurden mithin nicht verstärkt von eher leistungsschwachen Prüflingen bevorzugt.

In Bezug auf das Evaluationskriterium der kriterialen Validität wurden in den beiden Prüfungsjahren 2017 und 2019 kaum Auffälligkeiten festgestellt. Die für die Poolaufgaben und landeseigenen Aufgaben ermittelten Validitätskoeffizienten lagen jeweils im Intervall $r \in [.65; .85]$. Die Koeffizienten sind als hoch einzustufen und weisen insgesamt darauf hin, dass die im Fach Deutsch aus dem Pool entnommenen Aufgaben gut mit den in der Qualifikationsphase erzielten Leistungen korrespondieren. In beiden Prüfungsjahren fielen die in Bezug auf die Klausurnoten ermittelten Validitätskoeffizienten tendenziell etwas höher aus als die in Bezug auf die Halbjahresnoten berechneten Werte. Dieses Befundmuster zeigte sich in ähnlicher Form auch in den Fächern Englisch, Französisch und Mathematik und dürfte darauf zurückzuführen sein, dass es sich sowohl bei den Klausuren als auch bei den Abiturprüfungen um schriftlich zu erfüllende Anforderungen handelt und die angelegten Bewertungsmaßstäbe recht ähnlich sein sollten. Demgegenüber umfassen Halbjahresnoten auch Einzelnoten, die für die Bewältigung von Anforderungen anderer Art (insbesondere bei mündlichen Leistungskontrollen oder Referaten) erteilt werden. In Bezug auf die Halbjahresnoten wurde im Prüfungsjahr 2017 ein zwar statistisch signifikanter, aber nicht sehr großer Unterschied zwischen den für die Poolaufgaben ($r = .69$) und die landeseigenen Aufgaben berechneten Validitätskoeffizienten ermittelt ($r = .74$). In Bezug auf die Klausurnoten wurde hingegen keine statistisch signifikante Differenz festgestellt. Im Prüfungsjahr 2019 fanden sich für keines der beiden Kriterien statistisch signifikante Unterschiede in der Ausprägung der für die Poolaufgaben und die landeseigenen Aufgaben gefundenen Validitätskoeffizienten.

In den beiden Prüfungsjahren 2017 und 2019 schätzten die jeweils befragten Lehrkräfte den Anspruch von Poolaufgaben und landeseigenen Aufgaben etwa gleich hoch ein. Statistisch signifikante Unterschiede fanden sich hingegen im Hinblick auf die Lehrkräfteeinschätzungen zu den Erwartungshorizonten: Bei den Poolaufgaben wurde deren Nützlichkeit im Prüfungsjahr 2017 signifikant niedriger eingeschätzt als bei den landeseigenen Aufgaben, wobei die zu diesem Unterschied berechnete Effektstärke ($g = 0.28$) gemäß den in der Fachliteratur gängigen Konventionen als klein einzustufen ist (Borenstein et al., 2021). Eine nach Aufgabenarten differenzierte Betrachtung der Ergebnisse zeigt, dass die festgestellte Diskrepanz vor allem auf die weniger positiven Einschätzungen der Lehrkräfte zur Nützlichkeit der Erwartungshorizonte bei den Poolaufgaben zum materialgestützten Verfassen argumentierender Texte und

bei den Aufgaben zur Analyse pragmatischer Texte (jeweils $g = .50$, mittlerer Effekt) zurückzuführen sind. Ein anderes Befundmuster wurde wiederum im Rahmen der Evaluationsstudie zum Prüfungsjahr 2019 ermittelt. Hier beurteilten die Lehrkräfte die Nützlichkeit der Erwartungshorizonte der Aufgaben aus dem Pool als ähnlich hoch wie die der Erwartungshorizonte der landeseigenen Aufgaben.

Tabelle 2 fasst die in den Evaluationsstudien zu den Prüfungsjahren 2017 und 2019 für das Fach Deutsch festgestellten Auffälligkeiten nochmals in knapper Form zusammen.

Tabelle 2: Überblick über die in den Evaluationsstudien zu den Prüfungsjahren 2017 und 2019 für das Fach Deutsch festgestellten Auffälligkeiten

Evaluationskriterium	Auffälligkeiten
Auswahl der Aufgaben	Die Poolaufgaben wurden häufiger von leistungsschwächeren Prüflingen gewählt (2017). Dieses Ergebnis geht darauf zurück, dass unter den eingesetzten Poolaufgaben solche Aufgabenarten überrepräsentiert waren, die bevorzugt von leistungsschwächeren Prüflingen gewählt werden.
Empirische Schwierigkeit der Aufgaben	Keine Auffälligkeiten (sofern für das Prüfungsjahr 2017 die Überrepräsentation bestimmter Aufgabenarten berücksichtigt wird).
Kriteriale Validität der Aufgaben	In Bezug auf die Halbjahresnoten fiel der für die Poolaufgaben ermittelte Validitätskoeffizient signifikant geringer aus als der für die landeseigenen Aufgaben bestimmte Kennwert, wobei der Unterschied gering war (2017).
Lehrkräfteeinschätzungen zu den Aufgaben	Keine Auffälligkeiten.
Lehrkräfteeinschätzungen zu den Erwartungshorizonten	Die Erwartungshorizonte der Poolaufgaben wurden als weniger nützlich beurteilt als die Erwartungshorizonte der landeseigenen Aufgaben (2017).

Anmerkung: Jeweils in Klammern ist die Jahreszahl der Evaluationsstudie vermerkt, in der die betreffende Auffälligkeit festgestellt wurde.

3.2 Ergebnisse der Evaluation der Bewährung der Poolaufgaben im Fach Englisch

Analog zum Vorgehen im Fach Deutsch wurde im Rahmen der Evaluationsstudien zur Bewährung der Poolaufgaben in den Prüfungsjahren 2017 und 2019 auch für das Fach Englisch ermittelt, ob sich Poolaufgaben und landeseigene Aufgaben in Bezug auf die empirische Aufgabenschwierigkeit unterschieden. In beiden Prüfungsjahren wurden dabei keine statistisch signifikanten Unterschiede festgestellt. Bei einer nach Aufgaben der drei Kompetenzbereiche Hörverstehen, Sprachmittlung und Schreiben differenzierten Analyse fiel jedoch im Prüfungsjahr 2017 auf, dass die Prüflinge bei den Poolaufgaben zum Hörverstehen im Mittel signifikant besser abgeschnitten hatten als bei den übrigen Prüfungsaufgaben. Die hierzu ermittelte Effektstärke ist als klein einzustufen ($g = .27$), der Unterschied fällt also kaum ins Gewicht. Im Prüfungsjahr 2019 konnte ein solcher Effekt nicht festgellt werden.

Die für die Poolaufgaben einerseits und die landeseigenen Aufgaben andererseits berechneten Trennschärfekoeffizienten lagen in den Prüfungsjahren 2017 und

2019 jeweils im Intervall von $r \in [.84; .90]$ und waren somit sehr hoch. In beiden Prüfungsjahren unterschied sich die Höhe dieser Koeffizienten nicht signifikant zwischen Poolaufgaben und landeseigenen Aufgaben. Eine nach Kompetenzbereichen differenzierte Analyse zeigte allerdings in beiden Prüfungsjahren, dass die ermittelten Koeffizienten für die Poolaufgaben zum Hörverstehen signifikant geringer ausgeprägt sind als für die Poolaufgaben zu den Kompetenzbereichen Sprachmittlung und Schreiben. Ein sehr ähnliches Bild fand sich in den Prüfungsjahren 2017 und 2019 in Bezug auf das Evaluationskriterium der kriterialen Validität. Auch hier lagen die für die Poolaufgaben und landeseigenen Aufgaben ermittelten Koeffizienten in einem als hoch einzustufenden Bereich ($r \in [.70; .80]$) und auch hier wurden für die Poolaufgaben zum Hörverstehen signifikant geringere Kennwerte gefunden als für die Poolaufgaben zur Sprachmittlung und zum Schreiben. Insgesamt scheinen also die Poolaufgaben im Fach Englisch sowohl das Gesamtergebnis der Prüfung als auch die in der Qualifikationsphase erzielten Leistungen gut abzubilden. Die bei der nach Kompetenzbereichen differenzierten Betrachtung festgestellten Besonderheiten im Befundmuster dürften darauf zurückzuführen sein, dass die Prüfungsaufgaben der drei Kompetenzbereiche in den Ländern eine unterschiedlich lange Tradition haben und Aufgaben zum Hörverstehen von einigen Ländern erst seit dem Prüfungsjahr 2017 in der schriftlichen Abiturprüfung eingesetzt werden.

Die Auswertung der Lehrkräfteeinschätzungen zum Anspruch der Aufgaben ergab insgesamt sowohl im Prüfungsjahr 2017 als auch im Prüfungsjahr 2019 keine statistisch signifikanten Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben. Im Prüfungsjahr 2017 zeigte allerdings eine nach Kompetenzbereichen differenzierte Betrachtung, dass die Lehrkräfte die Poolaufgaben zum Hörverstehen signifikant anspruchsvoller einschätzten als die übrigen Prüfungsaufgaben. Die ermittelte Effektstärke ist als mittel einzustufen ($g = .56$). Bemerkenswert ist dieser Befund vor allem vor dem Hintergrund der Ergebnisse zur empirischen Aufgabenschwierigkeit, die zeigen, dass die Schülerinnen und Schüler die aus dem Pool eingesetzten Aufgaben zum Hörverstehen insgesamt gut bewältigt haben. Im Prüfungsjahr 2019 wurde eine derartige Überschätzung des Anspruchs der Aufgaben zum Hörverstehen nicht mehr festgestellt.

Im Hinblick auf die von den Lehrkräften beurteilten Nützlichkeit der zu den Aufgaben bereitgestellten Erwartungshorizonte wurden in den beiden betrachteten Prüfungsjahren sowohl in der Gesamtschau als auch bei einer nach Kompetenzbereichen getrennten Betrachtung keine statistisch signifikanten Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben festgestellt.

Die in den Evaluationsstudien zu den Prüfungsjahren 2017 und 2019 für das Fach Englisch festgestellten Auffälligkeiten sind in Tabelle 3 zusammengefasst.

Tabelle 3: Überblick über die in den Evaluationsstudien zu den Prüfungsjahren 2017 und 2019 für das Fach Englisch festgestellten Auffälligkeiten

Evaluationskriterium	Auffälligkeiten
Empirische Schwierigkeit der Aufgaben	Bei den Poolaufgaben zum Hörverstehen erzielten die Prüflinge im Mittel signifikant bessere Ergebnisse als bei den übrigen Prüfungsaufgaben (2017).
Trennschärfe der Aufgaben	Der für die Poolaufgaben zum Hörverstehen ermittelte Trennschärfekoeffizient fiel signifikant geringer aus als die für die Poolaufgaben zur Sprachmittlung und zum Schreiben berechneten Kennwerte (2017 und 2019).
Kriteriale Validität der Aufgaben	Der für die Poolaufgaben zum Hörverstehen ermittelte Validitätskoeffizient fiel signifikant geringer aus als die für die Poolaufgaben zur Sprachmittlung und zum Schreiben berechneten Kennwerte (2017 und 2019).
Lehrkräfteeinschätzungen zu den Aufgaben	Die Poolaufgaben zum Hörverstehen wurden im Vergleich zu den übrigen Prüfungsaufgaben als deutlich anspruchsvoller eingeschätzt (2017).
Lehrkräfteeinschätzungen zu den Erwartungshorizonten	Keine Auffälligkeiten.

Anmerkung: Jeweils in Klammern ist die Jahreszahl der Evaluationsstudie vermerkt, in der die betreffende Auffälligkeit festgestellt wurde.

3.3 Ergebnisse der Evaluation der Bewährung der Poolaufgaben im Fach Französisch

Im Fach Französisch legen bundesweit betrachtet deutlich weniger Schülerinnen und Schüler eine schriftliche Abiturprüfung ab als in den Fächern Deutsch, Englisch und Mathematik. Die für das Fach Französisch ermittelten Evaluationsergebnisse basieren dementsprechend auf erheblich kleineren Stichproben als die Befunde zu den anderen Fächern und sind daher weniger belastbar.

In Bezug auf das Evaluationskriterium der empirischen Aufgabenschwierigkeit wurden für das Fach Französisch in den Evaluationsstudien zu den Prüfungsjahren 2017 und 2019 insgesamt keine signifikanten Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben festgestellt. Bei einer nach Kompetenzbereichen differenzierten Betrachtung wurde für das Prüfungsjahr 2017 festgestellt, dass die Prüflinge bei den Poolaufgaben zum Hörverstehen signifikant schwächer abgeschnitten hatten als bei den übrigen Prüfungsaufgaben, wobei die ermittelte Effektstärke als klein bis mittel einzustufen ist ($g = .44$). Für das Prüfungsjahr 2019 fanden sich hingegen keine Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben.

In Bezug auf die beiden Kriterien der Trennschärfe und der kriterialen Validität wurden in den Prüfungsjahren 2017 und 2019 insgesamt sowohl für die Poolaufgaben als auch für die landeseigenen Aufgaben jeweils als hoch einzustufende Koeffizienten ermittelt (Trennschärfe: $r \in [.84; .91]$; kriteriale Validität: $r \in [.68; .77]$). Die Unterschiede in der Ausprägung der für die Poolaufgaben und landeseigenen Aufgaben bestimmten Kennwerte fielen dabei jeweils gering aus und waren statistisch nicht signifikant. Bei einer nach Kompetenzbereichen differenzierten Analyse wurden allerdings für die Poolaufgaben zum Hörverstehen jeweils signifikant geringere Trenn-

schärfe- und Validitätskoeffizienten festgestellt als für die Poolaufgaben der Kompetenzbereiche Sprachmittlung und Schreiben. Wie im Fach Englisch dürften auch die Auffälligkeiten im Befundmuster für die Poolaufgaben zum Hörverstehen im Fach Französisch darauf zurückzuführen sein, dass einige Länder erst seit dem Prüfungsjahr 2017 Aufgaben dieser Art in der schriftlichen Abiturprüfung einsetzen.

Die Auswertung der Lehrkräfteeinschätzungen zum Anspruch der Aufgaben im Fach Französisch und zur Nützlichkeit der bereitgestellten Erwartungshorizonte ergab in den beiden Prüfungsjahren 2017 und 2019 insgesamt keine signifikanten Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben. Bei einer nach Kompetenzbereichen differenzierten Analyse wurde jedoch für das Prüfungsjahr 2017 festgestellt, dass die Lehrkräfte die Poolaufgaben zum Hörverstehen im Vergleich zu den anderen Prüfungsaufgaben als signifikant anspruchsvoller beurteilten ($g = 1.01$, starker Effekt). Somit waren die Lehrkräfteeinschätzungen zum Anspruch der Poolaufgaben des Kompetenzbereiches Hörverstehen im Fach Französisch, anders als im Fach Englisch, konsistent mit den Ergebnissen zur empirischen Schwierigkeit dieser Aufgaben. Im Prüfungsjahr 2019 wurde der Anspruch der Poolaufgaben zum Hörverstehen im Fach Französisch hingegen nicht mehr signifikant höher beurteilt, was den Ergebnissen zur empirischen Schwierigkeit dieser Aufgaben entspricht.

Der Überblick der in den Evaluationsstudien zu den Prüfungsjahren 2017 und 2019 für das Fach Französisch festgestellten Auffälligkeiten findet sich in Tabelle 4.

Tabelle 4: Überblick über die in den Evaluationsstudien zu den Prüfungsjahren 2017 und 2019 für das Fach Französisch festgestellten Auffälligkeiten

Evaluationskriterium	Auffälligkeiten
Empirische Schwierigkeit der Aufgaben	Bei den Poolaufgaben zum Hörverstehen erzielten die Prüflinge im Mittel signifikant schwächere Ergebnisse als bei den übrigen Prüfungsaufgaben (2017).
Trennschärfe der Aufgaben	Der für die Poolaufgaben zum Hörverstehen ermittelte Trennschärfekoeffizient fiel signifikant geringer aus als die für die Poolaufgaben zur Sprachmittlung und zum Scheiben berechneten Kennwerte (2017 und 2019).
Kriteriale Validität der Aufgaben	Der für die Poolaufgaben zum Hörverstehen ermittelte Validitätskoeffizient fiel signifikant geringer aus als die für die Poolaufgaben zur Sprachmittlung und zum Schreiben berechneten Kennwerte (2017 und 2019).
Lehrkräfteeinschätzungen zu den Aufgaben	Die Poolaufgaben zum Hörverstehen wurden im Vergleich zu den übrigen Prüfungsaufgaben als deutlich anspruchsvoller eingeschätzt (2017).
Lehrkräfteeinschätzungen zu den Erwartungshorizonten	Keine Auffälligkeiten.

Anmerkung: Jeweils in Klammern ist die Jahreszahl der Evaluationsstudie vermerkt, in der die betreffende Auffälligkeit festgestellt wurde.

3.4 Ergebnisse der Evaluation der Bewährung der Poolaufgaben im Fach Mathematik

Im Rahmen der Evaluationsstudie zur Bewährung der Poolaufgaben im Prüfungsjahr 2017 wurde für das Fach Mathematik ein signifikanter Unterschied zwischen der für die Poolaufgaben und die landeseigenen Aufgaben ermittelten empirischen Aufgabenschwierigkeit gefunden. So erzielten die Prüflinge bei den Poolaufgaben im Mittel weniger gute Leistungen als bei den landeseigenen Aufgaben ($g = .27$, kleiner Effekt). Dabei zeigte sich in einer nach Sachgebieten differenzierten Analyse, dass der festgestellte Unterschied insbesondere auf die Poolaufgaben im Sachgebiet Stochastik zurückging. Während hier für die Aufgaben aus dem Pool geringere Lösungsquoten als für die landeseigenen Aufgaben festgestellt wurden ($g = .41$, kleiner bis mittlerer Effekt), waren in den übrigen Sachgebieten (d. h. für Analysis und Analytische Geometrie/Lineare Algebra) keine statistisch signifikanten Unterschiede zu verzeichnen.

Für die beiden Kriterien der Trennschärfe und der kriterialen Validität wurden insgesamt sowohl für die Poolaufgaben als auch für die landeseigenen Aufgaben jeweils als hoch einzustufende Koeffizienten ermittelt (Trennschärfe: $r \in [.78; .79]$; kriteriale Validität: $r \in [.59; .65]$). Statistisch signifikante Unterschiede fanden sich dabei nur bei einer nach Sachgebieten differenzierten Analyse. So wurden für die Poolaufgaben zum Sachgebiet Stochastik jeweils deutlich geringere Koeffizienten ermittelt als für die Poolaufgaben zu den übrigen Sachgebieten. Die Poolaufgaben zum Sachgebiet Stochastik hingen also mit dem Gesamtergebnis der Prüfung und mit den in der Qualifikationsphase erzielten Leistungen weniger eng zusammen als die Poolaufgaben zu den übrigen Sachgebieten.

Für die Lehrkräfteeinschätzungen zum Anspruch der Aufgaben und zu den bereitgestellten Erwartungshorizonten wurden keine signifikanten Unterschiede zwischen Poolaufgaben und landeseigenen Aufgaben festgestellt.

Tabelle 5 fasst die wenigen in der Evaluationsstudie zum Prüfungsjahr 2017 für das Fach Mathematik festgestellten Auffälligkeiten nochmals in knapper Form zusammen.

Tabelle 5: Überblick über die in der Evaluationsstudie zum Prüfungsjahr 2017 für das Fach Mathematik festgestellten Auffälligkeiten

Evaluationskriterium	Auffälligkeiten
Empirische Schwierigkeit der Aufgaben	Bei den Poolaufgaben erzielten die Prüflinge im Mittel signifikant schwächere Ergebnisse als bei den landeseigenen Aufgaben. Statistisch lässt sich dieser Unterschied vor allem auf die geringen Lösungsquoten bei den Poolaufgaben zum Sachgebiet Stochastik zurückführen.
Trennschärfe der Aufgaben	Der für die Poolaufgaben zum Sachgebiet Stochastik ermittelte Trennschärfekoeffizient fiel signifikant geringer aus als die Kennwerte, die für die Poolaufgaben der anderen Sachgebiete bestimmt wurden.
Kriteriale Validität der Aufgaben	Der für die Poolaufgaben zum Sachgebiet Stochastik ermittelte Validitätskoeffizient fiel signifikant geringer aus als die Kennwerte, die für die Poolaufgaben der anderen Sachgebiete bestimmt wurden.

(Fortsetzung Tabelle 5)

Evaluationskriterium	Auffälligkeiten
Lehrkräfteeinschätzungen zu den Aufgaben	Keine Auffälligkeiten.
Lehrkräfteeinschätzungen zu den Erwartungshorizonten	Keine Auffälligkeiten.

4 Fazit und Ausblick

Für die bis dato durchgeführten Evaluationsstudien zur Bewährung der Aufgaben aus den Gemeinsamen Abituraufgabenpools der Länder lässt sich insgesamt festhalten, dass in Bezug auf die sechs Evaluationskriterien in den meisten Fällen nur geringe Unterschiede zwischen Abituraufgaben aus den Pools und landeseigenen Abituraufgaben gefunden wurden. In den Fächern Deutsch, Englisch und Französisch, für die Daten zur Bewährung der aus den Pools entnommenen und in den Abiturprüfungen der Länder eingesetzten Aufgaben sowohl zum Prüfungsjahr 2017 als auch zum Prüfungsjahr 2019 erhoben wurden, zeigte sich zudem ein positiver Trend. So wurden in diesen Fächern im Prüfungsjahr 2019 weniger Auffälligkeiten festgestellt als noch im Prüfungsjahr 2017, wobei auch hier die Unterschiede zwischen den Poolaufgaben und den landeseigenen Aufgaben überwiegend gering waren. Eine ähnliche Entwicklung zeigt sich bei einer längsschnittlichen Betrachtung der Evaluationsergebnisse zur Nutzung der Aufgaben der Pools. So stieg die Anzahl der Aufgaben aus den Pools, die in mehreren Ländern eingesetzt wurden, im Laufe der Jahre an. Ein kleiner Einschnitt war in den Prüfungsjahren 2020 und 2021 zu verzeichnen, in denen einige Länder die zuvor vereinbarten gemeinsamen Prüfungstermine als Reaktion auf die Auswirkungen der Corona-Pandemie verschoben hatten und daher keine Poolaufgaben einsetzen konnten. Ein positiver Trend zeigt sich ebenfalls in Bezug auf die an den Poolaufgaben vorgenommenen Modifikationen. Hier ist im Laufe der Jahre insbesondere der Anteil an als gravierend eingestuften Modifikationen zurückgegangen. Die Grundlagen hierfür wurden im Rahmen des in Beitrag 2 skizzierten Annäherungsprozesses der Länder gelegt, im Zuge dessen eine zunehmende Angleichung landesspezifischer Prüfungsvorgaben erfolgt.

Die beschriebenen Trends verdeutlichen, dass der Implementationsprozess der Gemeinsamen Abituraufgabenpools der Länder in den letzten Jahren weiter vorangeschritten ist. Wie bereits weiter oben erwähnt, wurde die Evaluation konzeptionell bereits an diesen Entwicklungsverlauf angepasst. Zukünftig dürften weitere Anpassungen am Evaluationskonzept vorzunehmen sein. Folgerichtig erscheint vor allem, dass die Evaluation parallel zum weiteren Voranschreiten des Implementationsprozesses sukzessive die Frage in den Blick nehmen sollte, inwiefern die von der KMK für die Gemeinsamen Abituraufgabenpools formulierten Ziele erreicht werden. Neben der Sicherung der Qualität von Abiturprüfungsaufgaben und einer länderübergreifenden

Ausrichtung der Aufgabenstellungen an den Bildungsstandards für die Allgemeine Hochschulreife umfassen diese Zielvorgaben vor allem auch, dass die Vergleichbarkeit des Anforderungsniveaus der Abituraufgaben gewährleistet werden soll (KMK, 2017; Stanat et al., 2016; Stanat & Pant, 2013). Insbesondere mit der Fokussierung der letztgenannten Zielsetzung sollte in den nächsten Jahren eine Akzentverschiebung der bislang formativ ausgerichteten Evaluation hin zu einer summativ-bilanzierenden Evaluation stattfinden.

Literatur

Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. (2021). *Introduction to Meta-Analysis*. John Wiley & Sons.

DeGEval – Gesellschaft für Evaluation (2017). *Standards für Evaluation. Erste Revision 2016*. https://www.degeval.org/fileadmin/Publikationen/DeGEval-Standards_fuer_Evaluation.pdf

Döring, N. & Bortz, J. (2016a). Datenerhebung: Dokumentenanalyse. In N. Döring & J. Bortz (Hrsg.), *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl., S. 533–577). Springer.

Döring, N. & Bortz, J. (2016b). Evaluationsforschung. In N. Döring & J. Bortz (Hrsg.), *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl., S. 975–1036). Springer.

Hoffmann, L., Schröter, P. & Stanat, P. (2018). *Evaluation von Aufgaben der Pools für das Prüfungsjahr 2017. Ergebnisse zur Bewährung der Aufgaben*. https://www.iqb.hu-berlin.de/abitur/evaluation/PoolsfrdasPrfung.pdf [22.06.2021]

Hoffmann, L., Schröter, P. & Stanat, P. (2020). *Evaluation von Aufgaben der Pools für das Prüfungsjahr 2019. Ergebnisse zur Bewährung der Aufgaben*. https://www.iqb.hu-berlin.de/abitur/evaluation/PoolsfrdasPrfung_1.pdf [22.06.2021]

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik. (2015). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2015/2015_06_11-Gesamtstrategie-Bildungsmonitoring.pdf

KMK – Ständige Konferenz der Kultusminister der Länder der Bundesrepublik. (2017). *FAQs – Gemeinsamer Abituraufgabenpool der Länder*. https://www.kmk.org/fileadmin/Dateien/pdf/Bildung/AllgBildung/FAQs-Abiturpool.pdf

Scriven, M. (1972). Die Methodologie der Evahation. In C. Wulf (Hrsg.), *Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen* (S. 60–91). Piper & Co.

Speck, K. (2016). Programm-, Prozess- und Produktevaluation. In H. Marburger, C. Griese & T. Müller (Hrsg.), *Bildungs-und Bildungsorganisationsevaluation: Ein Lehrbuch* (S. 83–104). De Gruyter.

Stanat, P., Becker-Mrotzek, M., Blum, W. & Tesch, B. (2016). Vergleichbarkeit in der Vielfalt. Bildungsstandards der Kultusministerkonferenz für die Allgemeine Hochschulreife. In J. Kramer, M. Neumann & U. Trautwein (Hrsg.), *Abitur und Matura im Wandel* (S. 29–58). Springer.

Stanat, P. & Pant, H. A. (2013). *Konzeption für die Entwicklung und Nutzung eines Pools von Abiturprüfungsaufgaben (Arbeitsfassung)*. http://docplayer.org/49069143-Konzeption-fuer-die-entwicklung-und-nutzung-eines-pools-von-abiturpruefungsaufgaben-arbeitsfassung.html

Wanzer, D. L. (2020). What Is Evaluation? Perspectives of How Evaluation Differs (or Not) From Research. *American Journal of Evaluation, 42*(1), 28–46. https://doi.org/10.1177/1098214020920710

Tabellenverzeichnis

6 Implementation der Gemeinsamen Abituraufgabenpools der Länder im schulischen Mehrebenensystem – Projektskizze, theoretisch-konzeptuelle Grundlagen und erste empirische Befunde

Alexander Groß & Svenja Mareike Schmid-Kühn

Zusammenfassung

Seit dem Prüfungsjahr 2017 sind ländergemeinsam entwickelte schriftliche Abiturprüfungsaufgaben für die Fächer Deutsch, Mathematik, Englisch und Französisch in allen Bundesländern fester Bestandteil des Abiturprüfungsverfahrens. Mit diesen ist das Ziel verbunden, bundesweit vergleichbare Prüfungsanforderungen zu gewährleisten. Aufgrund der Kulturhoheit obliegt es den Bundesländern gleichzeitig vollumfänglich selbst, auf welche Art und Weise sie diese sogenannten Poolaufgaben letztlich in ihrer Prüfung zum Einsatz bringen. Die sich hieraus ergebende Frage nach dem Reformerfolg dieser Standardisierungsmaßnahme bildet die Grundlage des hier vorgestellten, qualitativ-explorativen Forschungsprojekts. Konkret wird im Rahmen einer Interviewstudie der gesamte Implementationsprozess dieser *Gemeinsamen Abituraufgabenpools der Länder* untersucht, wobei alle von der Einführung betroffenen Akteure verschiedener Systemebenen in die Analyse miteinbezogen werden. Dieser Beitrag stellt die Grundkonzeption des Forschungsprojekts sowie erste Untersuchungsergebnisse auf KMK-Ebene vor, wo der Implementationsprozess über eine Ländervereinbarung zur Entwicklung gemeinsamer Prüfungsaufgaben seinen Anfang nahm. Hierbei sind sowohl der seitens der KMK angestrebte sowie der beobachtete Umgang mit diesem neuen Instrument der Poolaufgaben auf Länderebene und die in diesem Zusammenhang entsprechend nachweisbaren Handlungslogiken von besonderem Interesse. Als konzeptueller Rahmen der Analyse dient der Ansatz der Educational Governance, ergänzt um zentrale Prämissen der Organisationstheorie des soziologischen Neo-Institutionalismus.

1 Einleitung

1.1 Ausgangslage

Das Abitur bzw. die Allgemeine Hochschulreife (AHR) ist der höchste zu vergebende Schulabschluss in Deutschland und berechtigt formal als einziger zur Aufnahme eines jeden Studiums an einer deutschen Universität (vgl. zur Geschichte der AHR

den Beitrag 1 von Klemm in diesem Band). Dabei kann festgehalten werden, dass im Bundeslandvergleich nicht bzw. nur eingeschränkt von einer entsprechenden Gleichwertigkeit des von allen Ländern vergebenen Abschlusszertifikats gesprochen werden kann; das Bundesverfassungsgericht konstatiert beispielsweise, dass die inhaltlichen Anforderungen zwischen den Bundesländern im schriftlichen Abitur so heterogen sind, dass dies in der Konsequenz zu einer „defizitären, länderübergreifenden Vergleichbarkeit der Abiturnoten" (BVerfG, 1BvL 3/14, Absatz 248) führt. Eine wesentliche bildungspolitische Reaktion auf dieses vielfach konstatierte Defizit stellen die aus diesem Grund in die Gesamtstrategie der Kultusministerkonferenz (KMK) zum Bildungsmonitoring aufgenommenen *Gemeinsamen Abituraufgabenpools der Länder* (in der Folge auch: [Aufgaben-]Pools bzw. Poolaufgaben) als zusätzliches Instrument zur Herstellung vergleichbarer Anforderungen dar. Inhaltlich anschlussfähig an die Beiträge 2 und 5 von Hoffmann, Schröter und Stanat in diesem Band wird in diesem Beitrag ein empirisches Forschungsprojekt zu der Implementation dieser Pools in ausgewählten Bundesländern vorgestellt; hierbei sollen sowohl die konzeptuell-theoretische Fundierung des explorativen Vorhabens erläutert als auch erste Analyseergebnisse aus dem Gesamtprojekt vorgestellt werden.

1.2 Die Herstellung von Vergleichbarkeit durch die Implementation Gemeinsamer Abituraufgabenpools

Hinter der Einführung der gemeinsamen Abituraufgabenpools der Länder verbirgt sich das grundsätzliche Ziel aller Bundesländer, das angesprochene Vergleichbarkeitsdefizit durch die gemeinsame Entwicklung und den entsprechenden Einsatz standardgesicherter Abiturprüfungsaufgaben aufzulösen. Diese Aufgaben werden unter Zusammenarbeit aller 16 Bundesländer für die Fächer Deutsch, Mathematik, Englisch und Französisch entwickelt[1] und für den Einsatz in der schriftlichen Abiturprüfung der Bundesländer bereitgestellt. Zur Verfügung stehen diese Aufgaben seit dem Prüfungsjahr 2017, womit deren Einführung gleichzeitig mit der Implementation der Bildungsstandards für die AHR im Schuljahr 2016/17 vollzogen wurde. Konkret soll so deren erfolgreiche Implementation unterstützt werden, was in der Konsequenz zu vergleichbaren Anforderungen in den Abiturprüfungen der Länder führen soll (Stanat & Pant, 2013; KMK, 2017; vgl. vertiefend den Beitrag 2 von Hoffmann et al. in diesem Band). Seit 2017 setzen alle Bundesländer nun (auch) Poolaufgaben in ihrer schriftlichen Abiturprüfung ein.

Diese auf formaler Ebene klar nachweisbaren ländergemeinsamen Standardisierungsbemühungen treffen nunmehr auf jenen Standardisierungskontext, aufgrund dessen das hier aufzulösende funktionale Defizit in Bezug auf die Gleichwertigkeit von Abschlüssen überhaupt erst virulent werden konnte: Aufgrund der föderalen Struktur der Bundesrepublik Deutschland, welche als „Kulturhoheit der Länder" (Art. 30 GG; Art. 70 GG) besonders auch in der Bildungspolitik zum Tragen kommt (Hepp, 2011), obliegen sämtliche strukturellen und inhaltlichen Entscheidungen, das Abitur allgemein und konkret auch die Abiturprüfung betreffend, grundsätzlich den

1 Ab dem Prüfungsjahr 2025 erweitert um die Fächer Biologie, Chemie und Physik.

Bundesländern selbst.[2] Jene zugewiesenen regulativen Handlungskompetenzen ermöglichen also länderspezifische Akzent- und Schwerpunktsetzungen (Tarkian & Thiel, 2016, S. 3), wenn es um die tatsächliche Umsetzung gemeinsam getroffener Beschlüsse auf Ebene der KMK in konkrete Regelungen auf Länderebene geht. In Bezug auf den somit als rein normativ geprägten Reformimpetus zu charakterisierenden gemeinsamen KMK-Beschluss kann festgehalten werden, dass die Bundesländer vollständig selbst darüber entscheiden können, welche und wie viele der Aufgaben aus dem Pool in der Prüfung letztlich verwendet werden. Darüber hinaus können diese aktuell auch noch „aufgrund länderspezifischer Regelungen […] Anpassungen vornehmen" (KMK, 2017, Ziff. 2), bevor die Poolaufgaben in der Prüfung zum Einsatz kommen (vgl. vertiefend Beitrag 2 von Hoffmann et al. in diesem Band). Diese gemeinsame Ländervereinbarung wurde im Oktober 2020 dahin gehend ergänzt, dass ab dem Jahr 2025 solche Modifikationen von den Ländern nicht mehr vorgenommen werden (sollen) (KMK, 2020; vgl. Beitrag 2 von Hoffmann et al. in diesem Band).

Insgesamt kann aufgrund solch umfassender Autonomien auf Bundeslandebene gleichwohl nicht ex ante von der Herstellung von Vergleichbarkeit durch die Implementation der Pools ausgegangen werden; auf struktureller Ebene sind die Poolaufgaben zwar fester Bestandteil des Abiturprüfungsverfahrens aller Bundesländer, die Art und Weise des Einsatzes in der Abiturprüfung obliegt jedoch einzig den Ländern selbst.

An dieser Stelle der Unsicherheit über den bildungspolitischen Reformerfolg der Einführung der Abituraufgabenpools ist das nachfolgend skizzierte Forschungsprojekt *Implementation der Gemeinsamen Abituraufgabenpools der Länder* angesiedelt. In diesem Projekt wurden Interviews mit Expertinnen und Experten geführt, die auf verschiedenen Ebenen des schulischen Mehrebenensystems (vgl. Abschnitt 2.1) mit der Implementation der Abituraufgabenpools befasst sind (Vertreterinnen und Vertreter der KMK und ausgewählter Länderministerien sowie Schulleitungen, Fach(konferenz)leitungen und Lehrkräfte aus verschiedenen Bundesländern; zum genauen Forschungsdesign vgl. Abschnitt 3). Im weiteren Verlauf wurden diese noch zusätzlich ergänzt um Interviews mit den (externen) Vertreterinnen und Vertretern wesentlicher Interessengruppen (Direktor:innenkonferenz, Lehrkräfteverbände/Gewerkschaften, Elternvereinigung, Schüler:innenkonferenz), um so einen möglichst umfassenden und multiperspektivischen Blick auf die Implementation der Poolaufgaben zu erhalten. Hiermit greift das Projekt ein diesbezüglich bestehendes Forschungsdesiderat auf und ist gleichzeitig anschlussfähig an die kontinuierliche Evaluation der Pools durch das Institut zur Qualitätsentwicklung im Bildungswesen (IQB) (vgl. Beitrag 2 von Hoffmann et al. in diesem Band).

Das Feld der wesentlichen Stakeholder im Kontext der Einführung der Poolaufgaben stellt sich entsprechend wie folgt dar, wobei zunächst die internen und unmittelbar von der Reform tangierten Akteure im Fokus der Untersuchung stehen.

2 Eine beobachtbare Folge dessen ist, dass sich auch auf Ebene der strukturellen Ausgestaltung der gymnasialen Oberstufe und der geltenden formal-organisatorischen Rahmenbedingungen für die Abiturprüfung in den Ländern ebenso eine entsprechende Vergleichbarkeitsherausforderung manifestiert. Vgl. hierzu den Beitrag von Schmid-Kühn & Groß in diesem Band.

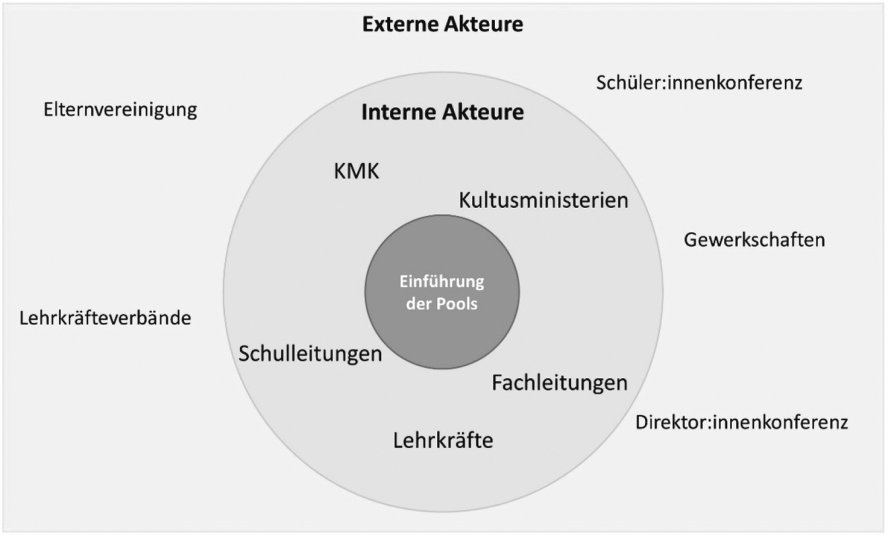


Abbildung 1: Wesentliche Stakeholder im Kontext der Einführung der Aufgabenpools[3] (eigene Darstellung)

In diesem Beitrag werden erste empirische Ergebnisse aus dem Projekt vorgestellt – fokussiert wird die Perspektive der KMK, wo die Einführung der Pools von allen Ländern einstimmig beschlossen und der Implementationsprozess entsprechend angestoßen wurde. Bevor die für die Analyse forschungsleitenden Fragestellungen dargelegt werden, ist es, aufgrund des hier genuin qualitativen Forschungszugangs und eines entsprechend angenommenen ko-konstituierenden Verhältnisses von Theorie und Empirie (z. B. Strübing, 2018; siehe auch Abschnitt 3), vorab geboten, die Analyse auch entsprechend konzeptuell-theoretisch zu fundieren.

2 Konzeptuelle Rahmung und theoretische Fundierung

2.1 Die Einführung der Aufgabenpools als bildungspolitisches Governance-Vorhaben

Die dargelegten Reformbemühungen auf Ebene der KMK stehen im Kontext der grundsätzlichen Frage nach der tatsächlichen Reformierbarkeit des Schulsystems (z. B. Dietrich, 2018). Allgemein gilt, dass die Einführung der Pools eine bildungspolitisch „verordnete“ Innovation darstellt, die als Steuerungsvorhaben fest in dessen Strukturlogik verankert ist. Der Begriff der Steuerung ist allerdings insofern irreführend, als diesem kein unilaterales Top-down-Verständnis im Sinne eines reibungsfreien Durchregierens (Bormann, 2014, S. 159) zugrunde liegt. Vielmehr muss analytisch der Komplexität des deutschen Bildungssystems Rechnung getragen werden, in

3 Darüber hinaus haben sich im Verlauf der Interviews die an dieser Stelle ausgeklammerten Wirtschaftsverbände noch als zusätzlich wesentliche Akteure herauskristallisiert.

welchem grundsätzlich nicht von solchen Prozessen plandeterminierter Beeinflussung ausgegangen werden kann, sondern den Möglichkeiten direkter Steuerbarkeit im Gegenteil klare Grenzen gesetzt sind. Dies liegt darin begründet, dass tatsächlich „soziale Ordnung und Leistung im Bildungswesen unter der Perspektive der Handlungskoordination zwischen verschiedenen Akteuren in komplexen Mehrebenensystemen" (Altrichter & Maag Merki, 2016, S. 8) zustande kommt. Diese Erweiterung des Steuerungsbegriffs stellt die inhaltlich-institutionelle Grundlage dar, auf welcher solche bildungspolitischen Reformvorhaben wie die Einführung der Aufgabenpools stattfinden und einzuordnen sind. Hiermit wird zunächst der Tatsache Rechnung getragen, dass an Veränderungsprozessen im Bildungswesen immer *verschiedene Akteure* beteiligt sind, die sich auf *unterschiedlichen Systemebenen* verorten; diese Akteure agieren jedoch nicht abgekoppelt voneinander, sozusagen als separate Entitäten, sondern beziehen sich handelnd, über diese Ebenen hinweg, reziprok *auf*einander (Handlungskoordination) und stehen darüber hinaus in unterschiedlichen Abhängigkeitsverhältnissen *zu*einander (Interdependenz). Jene Interdependenzen gründen sich auf die im schulischen Mehrebenensystem formalhierarchisch geprägten und gleichsam *asymmetrisch zugewiesenen Verfügungsrechte* (Altrichter, 2008), wie für die Handlungspotenziale der Bundesländer im Kontext der Implementation der Pools bereits angedeutet. Jene Steuerungsrealitäten gilt es bei der Analyse aktiv zu berücksichtigen, wobei auf dieser Grundlage zusammenfassend von einer Weiterentwicklung des klassischen Steuerungsansatzes in Richtung eines Multilevel-Governance-Verständnisses unterschiedlicher Interaktionsmuster und Koordinationsmechanismen (Benz, 2007, S. 297) gesprochen werden kann. Diese Grundidee, welche die konzeptuelle Basis des sogenannten Educational-Governance-Ansatzes (Altrichter & Heinrich, 2007) bildet, bringt die angesprochenen Unsicherheiten über den Reformerfolg der Einführung der Pools mit individuellen Handlungslogiken und sense-making-Prozessen (Weick, 1995) auf Akteursebene in Verbindung. Hieraus resultiert die wesentliche Erkenntnis, dass bei der Einführung einer solchen Neuerung analytisch unterschieden werden muss zwischen einem Gestaltungswillen durch Normsetzung, welche in diesem Fall auf Ebene der KMK stattfindet, und den sich daraus entwickelnden tatsächlichen Handlungsrealitäten auf den weiteren Systemebenen. Mögliche zutage tretende Implementationsbrüche in diesem Prozess der Institutionalisierung, welche die für die AHR angestrebte bundesweite Standardsicherung aktiv gefährden, sind dann das Resultat spezifischer (transintentionaler) Übersetzungs- (Sahlin & Wedlin, 2008) und Rekontextualisierungsprozesse (Fend, 2008) im Umgang mit dieser neuen Institution; sie entstehen aus einer Kombination aus handlungsleitenden Sinnlogiken und individuellen Einstellungen auf der einen und den den jeweiligen Akteuren tatsächlich zuerkannten Handlungspotenzialen auf der anderen Seite. Auf Basis der hieraus abzuleitenden Erkenntnis, dass Bildungsreformen institutionell bereits per se ein hohes Risiko des Scheiterns innewohnt (z. B. Maag Merki, 2018), interessiert in einem nächsten Schritt entsprechend, wie sich das schulische Mehrebenensystem konkret und mit Bezug zur Einführung der Pools konstituiert und welche Verfügungsrechte die entsprechenden Akteure der unterschiedlichen Ebenen

hier jeweils zugewiesen bekommen. Die Grundlage der in dieser Forschungsarbeit vorgenommenen Mehrebenenanalyse bildet die Aufgliederung des Schulsystems in eine „Makro-", „Intermediäre", „Meso-" und „Mikroebene" (Altrichter & Heinrich, 2007; Kühn, 2010).

Mit dieser Unterteilung wird der Notwendigkeit Rechnung getragen, dass KMK (Makro) und Kultusministerien (Intermediär) strukturell und auch analytisch distinkt voneinander unterscheidbar sein müssen. Das liegt darin begründet, dass hier auf den ersten Blick zwei Instanzen existieren, die das grundsätzliche Potenzial und die Legitimation besitzen, das Gesamtsystem steuernd zu beeinflussen, womit beide auf Makroebene zu verorten wären (Gasterstädt, 2019, S. 30). Der hier für die Analyse zentrale Unterschied zwischen diesen Akteuren besteht jedoch darin, dass auf Makroebene der KMK die Implementation der Aufgabenpools zwar gemeinsam zwischen den Ländern vereinbart wurde, ohne dass jedoch, wie bereits zu Beginn erläutert, diese Absichtserklärung für die Länder verbindliche Vorgaben zu Art und Umfang des Einsatzes der Aufgaben in den schriftlichen Abiturprüfungen beinhaltet hätte bzw. diese überhaupt aufgrund deren Kulturhoheit hätte beinhalten können; eine solche Entscheidung ist einzig durch politische Regelsetzung zu treffen und obliegt den jeweiligen Kultusministerien. Die Notwendigkeit einer trennscharfen Abgrenzung dieser beiden Ebenen voneinander basiert also auf unterschiedlich zugewiesenen Handlungspotenzialen im Umgang mit den Pools. Auf dieser Grundlage ist die auf KMK-Ebene aggregierte Norm sowie die auf Ebene der Kultusministerien und auf Basis individueller Sinn- und Handlungslogiken stattfindenden Übersetzungsprozesse jener Norm in (bildungspolitische) Handlung, von besonderem Interesse. Konkret finden an dieser Stelle im Institutionalisierungsprozess wesentliche Rekontextualisierungen statt, welche direkten Einfluss darauf haben, in welchem Maße die den Pools inhärenten Standardisierungspotenziale tatsächlich genutzt werden. Oder kurz: Auf Intermediärer Ebene entscheidet sich, wenn nicht ausschließlich, dann doch maßgeblich, ob durch gemeinsame Poolaufgaben bundesweit vergleichbare Abituranforderungen entstehen (können) oder nicht. Diese bereits angedeutete Steuerungsrealität findet sich dann auch entsprechend in der Unterteilung der Ebenen wieder, womit auch strukturell zum Ausdruck kommt, dass alle nachfolgenden Übersetzungsprozesse der Akteure weiterer Systemebenen sich immer auf die bereits rekontextualisierte Norm/Regel der Kultusministerien beziehen. Gerade für die Akteure auf Schulebene hat die Implementation der Poolaufgaben so überhaupt erst auf Länder- und nicht bereits auf Makroebene der KMK begonnen. Fend erklärt diesen Praxiszusammenhang theoretisch so, dass letztlich immer die „übergeordnete Ebene für die untergeordnete als Kontext präsent ist" (2008, S. 181) und sich Akteure jeweils auf den entsprechenden Input „von oben" aktiv handelnd beziehen. In diesem Fall sind hiermit die konkreten Regelsetzungen seitens der Kultusministerien gemeint, welche die institutionelle Handlungsgrundlage der Akteure auf Mesoebene (Schulleitungen, Fach[konferenz]leitungen) und Mikroebene (Lehrkräfte) darstellen. Eine entsprechende Visualisierung dieser komplexen Mehrebenenlogik mit konkretem Bezug zu den Aufgabenpools stellt sich auf dieser Grundlage nun wie folgt dar:

Das schulische Mehrebenensystem	Zentrale Akteure bei der Einführung der Aufgabenpools	Steuerungs-/Handlungspotenziale
Makroebene	Die Amtschefskommission *Qualitätssicherung in Schulen* auf Ebene der **KMK**	Artikulation/Bündelung (vermeintlich) geteilter Überzeugungen (normativ) als **nicht verbindliche Empfehlung**
Intermediäre Ebene	Für die Einführung und Umsetzung der Pools fachlich Verantwortlichen der **Kultusministerien**	umfassende Handlungsautonomie durch „Kulturhoheit der Länder" (**Regelsetzung**)
Mesoebene	(Erweiterte) **Schulleitungen** bzw. **Fach(konferenz)leitungen** der von der Einführung der Poolaufgaben betroffenen Fächer	**Adressaten der Regelsetzung**; schulinterne Qualitätssicherung bei professioneller Autonomie der Lehrkräfte
Mikroebene	Von der Einführung betroffene **Fachlehrkräfte**	graduelle Entscheidungsautonomie (Klatetzki, 2012)

Abbildung 2: Implementation der Abituraufgabenpools im schulischen Mehrebenensystem – Steuerungs-/Handlungspotenziale zentraler Akteure (eigene Darstellung)

Diese Darstellung fokussiert die entsprechenden grenzüberschreitenden und reziproken Prozesse der Handlungskoordinationen zwischen den Systemeben (z. B. Altrichter, Heinrich & Soukup-Altrichter, 2011), wobei hier das Augenmerk letztlich auf der Art und Weise des Managements vorhandener Interdependenzen liegt, welches im Kontext bildungspolitischer Reformen wesentlichen Einfluss auf deren Erfolg hat.

In Bezug auf das in diesem Beitrag vorgestellte Forschungsvorhaben wird auf Grundlage jenes Reformkontextes für die Einführung der Pools deutlich, dass deren erfolgreiche Implementation im Sinne der intendierten Sicherung allgemeingültiger Qualitätsstandards der Allgemeinen Hochschulreife möglich, aber nicht ex ante gesichert ist. Dieser Erkenntnis und den entsprechenden bis hierhin dargelegten theoretischen Überlegungen in diesem Zusammenhang folgend, ist das gesamte Forschungsprojekt zunächst als Governance-Studie angelegt mit dem konkreten Ziel, Handlungslogiken und Rezeptionsprozesse im Mehrebenensystem bezüglich des Umgangs mit den Pools als Veränderungsimpuls systematisch zu analysieren und dabei alle maßgeblichen Entscheidungsprozesse in ihrer Vielschichtigkeit und besonders auch im Kontext grenz- und ebenenüberschreitender Verflechtungen (Bosche & Lehmann, 2014, S. 241) zu berücksichtigen. Um jene Potenziale auch tatsächlich zu nutzen, wird das gesamte Mehrebenensystem, inklusive des bisher wissenschaftlich vernachlässigten, nichtsdestoweniger bildungspolitisch relevanten Akteurs der KMK in die Analyse einbezogen; ergänzt um die erwähnten externen Stakeholder wesentlicher Interessengruppen wird somit dem Implementationsprozess in seiner gesamten Komplexität analytisch Rechnung getragen.

Die in diesem Beitrag dargestellten ersten Ergebnisse fokussieren als Ausschnitt nun entsprechend jene Normsetzungen auf Makroebene und die von dort beobachteten Übersetzungsprozesse im Wege der Implementation, also die unmittelbare Über-

tragung der auf KMK-Ebene aggregierten Norm in konkrete Handlungen in Form von Regelsetzungen auf Länderebene. Hierbei liegt ein erster Forschungsfokus auf der Herausarbeitung bzw. dem Sichtbarmachen der Entscheidung selbst („Wie"), ergänzt um eine Perspektive, in welcher dann nicht die reine Beschreibung der beobachtbaren Handlung, sondern die diesen Handlungsprozessen und -koordinationen zugrundeliegenden Erklärungen („Warum") in den Analysefokus rücken.

Eine solche analytische Berücksichtigung nachweisbarer Begründungsmuster bedarf gleichzeitig notwendigerweise auch einer entsprechenden theoretischen Erweiterung, da mit den Prämissen der Educational Governance Zusammenhänge lediglich dargestellt und nicht in Bezug auf deren „ursächlichen Impulse und treibenden Faktoren der Erzeugung" (Langer, 2015, S. 47; Bosche, 2013; Graß, 2015) analysiert werden können.

2.2 Die Erklärung komplexer Handlungszusammenhänge auf Basis des soziologischen Neo-Institutionalismus

Ein solcher analytische Einbezug der Perspektive des „Warum" muss hier also nicht nur forschungspraktisch geleistet, sondern notwendigerweise ebenso theoretisch fundiert werden; hierfür werden zentrale Prämissen des soziologischen Neo-Institutionalismus (NI) (Meyer & Rowan, 1977; DiMaggio & Powell, 1983; Zucker, 1977) in die Mehrebenenanalyse miteinbezogen. Grundsätzlich besteht das analytische Potenzial dieser organisationstheoretischen Erweiterung darin, komplexe Zusammenhänge zwischen verschiedenen (organisationalen) Ebenen nun sowohl beschreiben *als auch* entsprechende Begründungsmuster herausarbeiten zu können (Muslic, 2017, S. 29; Senge, 2011). Die Vertreterinnen und Vertreter des NI propagieren in diesem Kontext zunächst die Abkehr von der Klassifikation einer Organisation (wie der Schule) als „technisch-rationalem Instrument" (Walgenbach, 2019, S. 301) und der hier zugrunde gelegten Vorstellung, dass organisationales Handeln, orientiert an Max Weber (1980), immer vernunftgeleitet und effizient im Sinne einer zu erfüllenden Aufgabe oder eines zu lösenden Problems, folglich reibungsfrei erfolgt.

Auf Basis dieser sich auch hier wieder manifestierenden Steuerungsskepsis fokussiert der NI analytisch konkret die „Entstehung und Verbreitung von Institutionen" (Süß, 2009, S. 190). Institutionen stellen hierbei das Bindeglied zwischen (Schul-)Organisation und Gesellschaft dar und bezeichnen beständig zur Anwendung gebrachte und gleichzeitig veränderbare soziale Regeln. Diese zentralen gesellschaftlichen Strukturausprägungen (Türk, 1997, S. 158) bilden den Kontext, in welchem die Einführung der Aufgabenpools einzuordnen ist. Dabei gilt allgemein, dass eine Organisation wie die Schule, mit ihren unterschiedlichen Systemebenen und dem Handeln der jeweils dort verorteten Akteure, grundsätzlich gesellschaftlich eingebettet ist und vorhandene Institutionen als Regelsysteme maßgeblich die hier stattfindenden Handlungskoordinationen beeinflussen. Die bereits angesprochenen Rekontextualisierungen von Akteursseite basieren folglich so auf der einen Seite auf entsprechend zuerkannten institutionellen Handlungsspielräumen wie dargestellt, gleichzeitig ist jenes Handeln im Umkehrschluss gleichzeitig auch immer mehr oder weniger insti-

tutionell beschränkt und zwar durch die bereits im Governance-Ansatz figurierten unterschiedlich zugewiesenen Verfügungsrechte. Jene grundsätzliche Orientierung an der strukturationstheoretischen Idee einer Dualität von Struktur und Handlung (Giddens, 1979) trägt auf der einen Seite zu einer Theoretisierung der angesprochenen Handlungskoordinationen im schulischen Mehrebenensystem bei (Langer, 2015) und greift gleichzeitig auch den wesentlichen Gedanken einer grundsätzlichen Reform- und Steuerungsskepsis top-down wieder auf. Hieraus resultiert, dass neben der herausgestellten Notwendigkeit, gleichzeitig nun auch die theoretisch fundierte Möglichkeit besteht, in Bezug auf die Einführung der Pools sowohl vorhandene Regelungsstrukturen als auch die Innenperspektive von Akteuren zu betrachten, wobei dieses Interdependenzverhältnis einer sogenannten „embedded agency" (Seo & Creed, 2002) sowohl als „Restriktion als auch als Handlungschance" (Niedlich, 2019, S. 355) für diese zu verstehen ist.[4] Mit einem solchen bewussten Einbezug der Akteursperspektive, welche in der Educational Governance analytisch eher vernachlässigt wird (Graß, 2015), und der gleichzeitigen Berücksichtigung entsprechender Strukturvorgaben des schulischen Mehrebenensystems (Muslic, 2017) kann letztlich die intendierte Erweiterung der Analyseperspektive vom „Wie" auf das „Warum" geleistet werden. Bezogen auf die Einführung der Pools bedeutet das konkret, dass die Handlungen der individuellen Akteure auf den unterschiedlichen Systemebenen und die nachweisbaren Handlungskoordinationen als individueller und gleichzeitig institutionell beeinflusster Prozess zu verstehen sind. Besonders der Blick auf die Art und Weise der Umsetzung von Vorgaben und die Existenz und das Ausnutzen institutioneller Freiräume in diesem Zusammenhang können mithilfe zentraler neo-institutionalistischer Analysekategorien multidimensional beschrieben und mit dem Fokus auf entsprechende Kausalzusammenhänge analysiert werden. Mithilfe jener konkreten, aus dem Theoriegerüst des NI emergierten Handlungsstrategien als Möglichkeiten der Interdependenzbewältigung (Schimank, 2007), können folglich der die spezifischen Handlungskoordinationen prägende institutionelle Erwartungszusammenhang (Herbrechter & Schemmann, 2019, S. 195) offengelegt sowie individuelle Rekontextualisierungen und Übersetzungen auf Basis subjektiver Handlungslogiken in die Analyse miteinbezogen werden.

Bevor konkret auf die verschiedenen Strategien des Umgangs mit den Pools eingegangen werden kann, ist es wesentlich darzustellen, dass Institutionalisierungsprozesse grundsätzlich auf drei Ebenen nachweisbar sind: der *regulativen*, der *normativen* und der *kulturell-kognitiven* (Scott, 2014). Institutionen vereinen in der Regel alle drei Elemente in sich und konstituieren sich letztlich auf Basis der einzelnen Merkmalsausprägungen; anhand jener spezifischen institutionellen Ausprägungen kann dann auch die zentrale Frage nach der jeweiligen legitimatorischen Basis, also der Art und Weise der Anerkennung einer Institution beantwortet werden (Suchman, 1995; Scott, 2014).

4 Diese, hier notwendigerweise stark verkürzte, Darstellung der theoretischen Grundlagen des NI basiert auf der grundsätzlichen Vorstellung eines seitens der Akteure interpretativen und auf spezifischen Handlungspotenzialen beruhenden Umgangs mit institutionellen Anforderungen. Für eine umfassende Darstellung der wissenschaftstheoretischen Basis dieser akteursorientierten Erweiterung grundlegender neo-institutionalistischer Konzepte vgl. Groß (i. V.).

Steht bei einer Institution vornehmlich das *regulative Element* im Vordergrund, bildet die Basis dieser Institution die explizit festgeschriebene Regel („what has to be done"). Diese Regel liegt häufig in Form von (staatlichen) Gesetzen vor und muss notwendigerweise befolgt werden; Nicht-Befolgung führt meist zu einer entsprechenden Sanktion. Im konkreten Fall verfügen, wie erläutert, ausschließlich die Kultusministerien über diese Möglichkeit der institutionellen Ausprägung der Pools. Handlungen werden in diesem Zusammenhang dann als legitim betrachtet, wenn sie den geltenden Regularien entsprechen, oder anders herum: Wenn der Umgang mit einer Institution auf einer regulativen Legitimitätszuschreibung beruht, hält sich der Akteur an geltende (gesetzliche) Vorgaben, die sich, inklusive möglicher Handlungsspielräume, aus der konkreten regulativen Ausprägung der Institution ergeben; eine gleichzeitige inhaltliche Legitimation ist an dieser Stelle ebenso möglich, aber nicht notwendigerweise vorhanden.

Ist eine Institution hingegen hauptsächlich *normativ* geprägt, stellt diese eine Erwartung dahin gehend dar, was auf Basis eigener (Wert-)Vorstellungen als (institutionell) angemessen („how things should be done") eingeschätzt wird. Wird dieser Vorstellung entsprochen, dann geschieht dies auf Basis geteilter Normen und Sinnlogiken in Bezug auf den vorliegenden Gegenstand und man weist diesem bzw. der entsprechenden Institution normative Legitimität zu. Diese Legitimation tangiert konkret die Inhaltsebene und weist damit über die reine Erfüllung regulativer Vorgaben hinaus. Im Fall der Pools wird deutlich, dass vor der eigentlichen regulativen Institutionalisierung auf Länderebene, die KMK diese normativ ausprägt, im Sinne der Zielerreichung legitimiert und daraus eine als angemessen bewertete Regelsetzung ableitet, ohne das regulative Element jedoch tatsächlich aktiv gestalten zu können.

Während diese innere Überzeugung als klar werteorientiert einzuordnen ist, können die erwähnten Sinnlogiken auch an einem spezifischen, ggf. auch kurzfristig realisierbaren Individualinteresse eines Akteurs ausgerichtet sein. Hierbei erhofft man sich, durch die neue Institution selbst bzw. den eigenen Umgang mit dieser individuelle Ziele zu erreichen, wie z. B. länderseitig eine externe Anerkennung eigener Standardisierungsbemühungen im Kontext des Abiturs. Ist eine solche Logik analytisch nachweisbar, weist man den Pools entsprechend *pragmatische* Legitimität (Suchman, 1995) zu.

Das dritte, *kulturell-kognitive* Element einer Institution verweist auf deren Stabilität und Persistenz. Konkret liegen bei spezifischen Akteuren potenziell als selbstverständlich erachtete Handlungsroutinen in Bezug auf einen entsprechenden Gegenstand vor („the way we do things"), die eine existierende Institution umfassend stabilisieren und deshalb nur schwer ablösbar machen, da ein alternatives Handeln schlicht nicht/kaum vorstellbar ist (Zucker, 1977). Solche taken-for-granted-Annahmen (Jepperson, 1991) darüber, wie qualitativ hochwertige Bildung zu erzeugen ist, können grundsätzlich bei allen Akteuren des schulischen Mehrebenensystems vorliegen, z. B. in Form spezifischer und tradierter Fächerlogiken (Beispiel: „Taschenrechnerfrage" in Mathematik; „(Literatur) Kanonfrage" in Deutsch). Je nach zugewiesenen Verfügungsrechten stellen derartig stark verwurzelte Vorstellungen ggf. ein Implementationshin-

dernis dar, da sie dann potenziell zu einem transintentional-rekontextualisierenden Umgang mit den Pools führen, wenn auf Basis jener Selbstverständlichkeiten gehandelt wird und diese einer Herstellung vergleichbarer Anforderungen im Weg stehen. Dies ist insbesondere dann wahrscheinlich, wenn entsprechende Legitimitätszuschreibungen für eine neue Institution ausbleiben, da dem vorherrschenden System selbstverständlicher („Das haben wir immer schon so gemacht.“) Deutungsmuster (Walgenbach & Meyer, 2008, S. 64) nicht entsprochen wird und gleichzeitig auf regulativer Ebene keine/kaum Sanktionen für den individuell-transintentionalen Umgang mit der institutionellen Neuerung zu befürchten sind. Das lose gekoppelte schulische Mehrebenensystem (Weick, 1976) mit z. T. umfassend institutionalisierten Freiräumen für die Akteure der unterschiedlichen Systemebenen (vgl. Abbildung 2), bietet hier einen entsprechenden Gestaltungsrahmen für solche Übersetzungsprozesse.

Insgesamt gehen mit jenen institutionellen Ausprägungen wie gesagt immer auch entsprechende Legitimitätszuschreibungen für die Aufgabenpools als Institution einher, auf deren Basis sich dann der konkrete Umgang mit dieser bildungspolitischen Innovation manifestiert. Dieser Zusammenhang lässt sich folgendermaßen visualisiert darstellen:

Institutionelle Elemente	Art und Weise der Wirkung	Form der Legitimität	Basis der Zuschreibung
Regeln, Gesetze	Regulativ → Zwang und Sanktion	Regulative Legitimität	Vorgaben in Form von Regeln/Gesetzen werden erfüllt
Normen	Normativ → Wertebasis des ‚Angemessenen‘	Normative Legitimität	Spezifischen Werten und Normen wird institutionell entsprochen
Geteilte kulturelle Deutungsmuster	Kulturell-kognitiv → unhinterfragte Selbstverständlichkeit	Kulturell-kognitive Legitimität	taken-for-granted Annahmen wird institutionell entsprochen
		Pragmatische Legitimität	Akteure erreichen durch Institution individuelle Ziele

Abbildung 3: Elemente von Institutionen und der Prozess der Legitimitätszuweisung (eigene Darstellung)

In Bezug auf die Einführung der Aufgabenpools hängt der Reformerfolg zunächst ganz konkret und maßgeblich davon ab, welche Normen und ggf. unhinterfragten Selbstverständlichkeiten der regulativen Institutionalisierung auf Intermediärer Ebene der Kultusministerien zugrunde liegen und ob diese sich von der auf KMK-Ebene gemeinsam aggregierten Norm unterscheiden. Darüber hinaus ist von analytischer Bedeutung, wie auf Schulebene mit der seitens der Kultusministerien regulativ

ausgeprägten Neuerung handelnd umgegangen wird; im Fokus steht hier also insgesamt die Frage, ob und wenn ja, inwiefern, im Sinne des angestrebten Ziels der bundesweiten Standardsicherung, transintentionale und kontraproduktive Rekontextualisierungen auf Basis divergierender Normen bzw. taken-for-granted-Annahmen nachweisbar sind.

Als Folge jener potenziellen (De-)Legitimationen der Pools lassen sich auf Handlungsebene und von Akteursseite konkrete Mechanismen der Rekontextualisierung nachweisen. Diese erwähnten, ebenso dem neo-institutionalistischen Theoriegebäude entnommenen Handlungsstrategien für den Umgang mit dieser Reformaufforderung manifestieren sich entweder als *Anpassung* an eine institutionell bindende Anforderung (regulativ) bzw. einen entsprechenden normativen Reformimpetus, mit dem ebenso entsprechende Erwartungen verbunden sind, oder im Sinne einer aktiven *Trennung* von institutioneller Formal- („policy") und tatsächlicher Handlungsebene („practice").

Im ersten Fall der handelnden Entsprechung kann von einer im Sinne der Zielerreichung intentionalen Rekontextualisierung gesprochen werden, die als sogenannter Isomorphieprozess (DiMaggio & Powell, 1983) Auswirkungen auf sowohl Struktur- als auch Handlungsebene zeitigt. Das bedeutet, dass Akteure mit ihrem Handeln versuchen, den institutionellen Vorgaben so handelnd zu entsprechen wie intendiert. Angelehnt an die verschiedenen Elemente der Institution und die Arten der Legitimation, kann diese Konformität entweder gesetzlich (regulativ) erzwungen werden bzw. auf Basis einer normativen/kulturell-kognitiven Legitimitätszuschreibung (man ist inhaltlich derselben Ansicht) entstehen. Wenn die Pools also beispielsweise so auf Länderebene eingeführt werden (Regel), wie auf KMK-Ebene gemeinsam und im Sinne der Herstellung von Vergleichbarkeit verhandelt (Norm), konstituiert und manifestiert sich diese (Isomorphie-)Entscheidung auf den dargestellten institutionellen Ebenen und im Kontext spezifischer Legitimitätszuweisungen. Derselbe Handlungszusammenhang gilt grundsätzlich ebenso für die Akteure auf Schulebene und deren Umgang mit den neuen zentralen Vorgaben, dort allerdings, aufgrund der geltenden Formalhierachie (z. B. Fuchs, 2004), mit stärker regulativ geprägten institutionellen Voraussetzungen.

Neben diesem Ansatz der angenommenen Institutionenkonformität auf Akteursseite, welcher für sich genommen eine deutliche „pro-Reform"-Tendenz aufweist, stellt die Idee der Entkopplung von Struktur und Handlung (Meyer & Rowan, 1977) den entsprechenden handlungspraktischen und theoretischen Gegenentwurf hierzu dar. Hier wird konkret der Tatsache Rechnung getragen, dass Reaktionen von Akteursseite auf geäußerte institutionelle Ansprüche wesentlicher Akteure in der Praxis nicht ausschließlich gleich, sondern regelmäßig auch unterschiedlich ausfallen; dies liegt an der bereits thematisierten grundsätzlichen Schwierigkeit, Steuerungsprozesse überhaupt und speziell auch in einem schulischen System loser Kopplungen zu technologisieren bzw. zu kontrollieren (Wacker, 2008; Terhart, 2000). Konkret greift das Theorem der Entkopplung die handlungspraktische Realität auf, dass auf struktureller Ebene eine Regelsetzung vorgenommen wird, welcher handelnd

hingegen ggf. gar nicht oder nicht wie intendiert entsprochen wird. Im konkreten Fall existieren möglicherweise länderinterne modi operandi in Form tradierter Handlungsroutinen und stabiler kultureller Muster (Langer & Brüsemeister, 2019, S. 772) zum schriftlichen Abiturprüfungsverfahren, welche umfassend legitimiert sind und auf inhaltlicher Ebene entsprechend nicht abgelöst werden sollen. In der Konsequenz homogenisiert man so, aus Gründen spezifischer (angestrebter) Legitimitätszuschreibungen, Handeln zwar auf regulativer Ebene, stabilisiert aber gleichzeitig weiterhin im Prozess des Handelns selbst individuelle Normen bzw. reproduziert unhinterfragte, institutionelle Selbstverständlichkeiten. Indem man eine solche Konformitätsfassade („ceremonial facade"; Meyer & Rowan, 1977) herstellt, entkoppelt man die gezeigte Handlung („talk"), i. d. R. in Form einer regulativ vorhandenen Institutionalisierung, von den tatsächlich realisierten Inhalten („action"). Die Möglichkeit, diesen Prozess des sogenannten „window dressings" mit dem theoretischen Begriffsapparat des NI beschreiben zu können, ist deshalb für die Analyse des Forschungsgegenstandes der Abituraufgabenpools in besonderem Maße fruchtbar, da auf Sichtebene der beobachtbaren Handlung ein isomorpher Prozess der Homogenisierung nachweisbar ist und zwar insofern als alle Bundesländer, obwohl keinerlei rechtliche Verpflichtung dazu besteht, die Pools trotzdem regulativ wie erläutert implementieren; gleichzeitig besteht jedoch auch die Möglichkeit, dass in den Bundesländern eigene, bis dato tradierte, also unhinterfragt selbstverständliche Handlungsroutinen in Bezug auf das Abitur aufgrund ihrer institutionellen Persistenz nicht einfach abgelöst wurden, sondern trotzdem weiter zur Anwendung kommen. Solche potenziell nachweisbaren Entkopplungshandlungen sind genauso wenig prädeterminiert wie isomorphe Prozesse, beide stellen jedoch eine jeweils *mögliche Reaktion* auf die entsprechende Reformaufforderung dar, welche nunmehr mithilfe des neo-institutionalistischen Theoriegerüsts auch beschrieben *und* erklärt werden kann.

Die dargestellten institutionellen (Governance-)Realitäten machen es an dieser Stelle also notwendig, neben den ggf. nachweisbaren Isomorphieprozessen, auch mögliches transintentionales Verhalten von Akteuren und die dergestalte Ausschöpfung zuerkannter Handlungspotenziale analytisch berücksichtigen zu können. Dabei schließen sich diese beiden unterschiedlichen neo-institutionalistischen Konzepte nicht aus, sondern ko-existieren unter der forschungsleitenden Prämisse, dass die nachweisbaren Übersetzungsprozesse und Rekontextualisierungen „may not only lead to homogenization but also to variation […]" (Sahlin & Wedlin, 2008, S. 219). Diese neo-institutionalistische Fundierung von Handlungs*alternativen* im Umgang mit den Pools als multidimensionale Institution lässt sich wie folgt visualisieren:

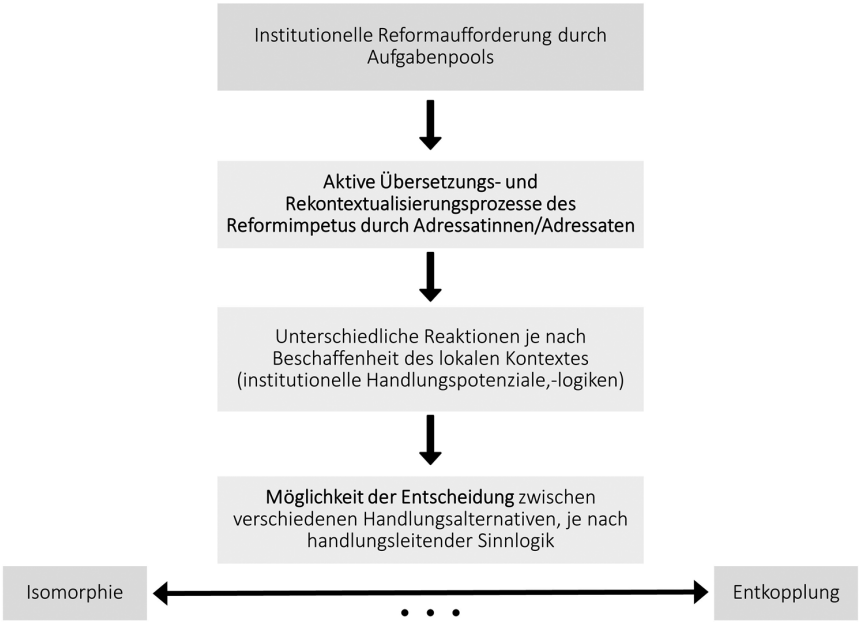


Abbildung 4: Prozess der Übersetzung einer Reformaufforderung in Handlung (eigene Darstellung)

Für die Einführung der Aufgabenpools stellt sich also letztlich die Frage nach der Nachweisbarkeit und der Genese solcher Isomorphie- bzw. Entkopplungsprozesse, wobei insbesondere die hier zugrunde gelegten Handlungslogiken der Akteure der verschiedenen Systemebenen von analytischem Interesse sind, da diese den tatsächlichen Reformerfolg maßgeblich (mit) beeinflussen.

3 Forschungsdesign und zentrale Fragestellungen

Das gesamte Forschungsprojekt folgt den Prinzipien des interpretativen Paradigmas (Wilson, 1980) und ist als explorativ-entdeckendes Vorhaben (Brüsemeister, 2008; Lamnek & Krell, 2016) angelegt. Ein solch genuin qualitatives Vorgehen wird vor allem dann zur Anwendung gebracht, wenn der Forschungsgegenstand noch weitgehend unerforscht und auch theoretisch noch wenig strukturiert ist und bis dato diesbezüglich noch keine Formulierung von Hypothesen hat stattfinden können (z. B. Döring & Bortz, 2016; Lamnek & Krell, 2016). Dies ist bei der Einführung der Pools und in Bezug auf die sich hier manifestierenden Implementations- und Handlungslogiken der Akteure auf den verschiedenen Systemebenen der Fall. An dieser Stelle soll explizit darauf hingewiesen werden, dass die analytischen Potenziale, welche sich aus dieser qualitativ-explorativen Ausrichtung des Forschungsprojekts ergeben, in der Herausarbeitung subjektiv gemeinten Sinns und des Nachvollzugs unterschiedlicher Akteursperspektiven auf den Forschungsgegenstand der Aufgabenpools bestehen. Hieraus ergibt sich, dass die so entstehenden Analyseergebnisse die Komplexität und

Mannigfaltigkeit gesellschaftlicher Wirklichkeit (Mannheim, 1980; Schütz, 1971) berücksichtigen, gleichzeitig aber nicht die Herstellung statistischer Repräsentativität und einer entsprechenden Verallgemeinerbarkeit der Ergebnisse zum Ziel haben (können); eine solche könnte auf Grundlage der in diesem Forschungsprojekt geleisteten Exploration und mithilfe eines dann genuin quantitativen Zugangs, z. B. in Form einer in allen Bundesländern durchgeführten (teil-)standardisierten Fragebogenstudie (z. B. Döring & Bortz, 2016), geleistet werden, wobei mit einer solch triangulativen Vorgehensweise ein klar komplementär-integratives Verhältnis qualitativer und quantitativer Forschungsparadigmen zum Tragen kommt (vgl. z. B. Kruse, 2015).

Empirische Grundlage dieses Beitrags sind Daten aus einer qualitativen Interviewstudie: Im Rahmen der Studie wurden, orientiert an der Struktur des Mehrebenensystems und mit konkretem Fokus auf den gesamten Prozess der Einführung der Pools, leitfadengestützte Einzelinterviews mit Expertinnen und Experten (Bogner, Littig & Menz, 2014) aller Ebenen geführt. Hierbei wurden, neben den Vorsitzenden der Amtschefskommission *Qualitätssicherung in Schulen* (AC1; AC2), in drei verschiedenen Bundesländern jeweils Akteure auf Intermediärer- und Meso-/Mikroebene interviewt, sodass insgesamt Datenmaterial aus 55, jeweils ca. einstündigen Expertinnen- und Experteninterviews, durchgeführt in den Jahren 2018 und 2019 zur Einführung der Pools vorliegt. Diese Fallauswahl stellt sich visualisiert folgendermaßen dar:

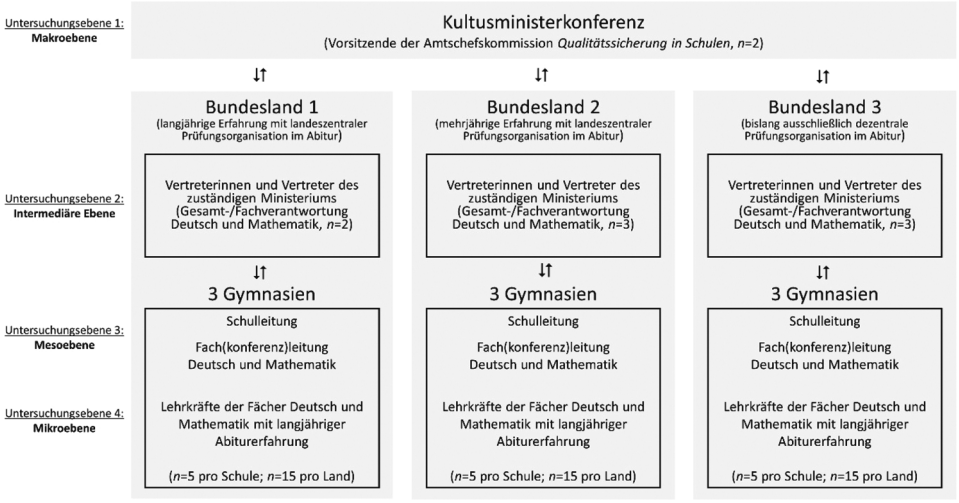


Abbildung 5: Fallauswahl (eigene Darstellung)

Die der Fallauswahl zugrunde liegende Samplingstrategie folgt allgemein dem Prinzip der qualitativen Repräsentation (Kruse, 2015), wobei durch das sogenannte „intensity sampling" (Patton, 1990) bewusst heterogene Fälle ausgewählt und so gleichwohl das vorhandene Varianzspektrum des zu untersuchenden Forschungsgegenstandes berücksichtigt werden kann. In dem hier vorgestellten Forschungsprojekt wurden entsprechend drei Bundesländer mit möglichst unterschiedlichen Traditionen zur Or-

ganisation und Durchführung der Abiturprüfung in die Analyse einbezogen. Durch die Auswahl und Analyse von Fällen mit möglichst unterschiedlichen (institutionellen) Rahmenbedingungen und der so antizipierten verschieden gelagerten Sinn- und Handlungslogiken auf Akteursseite in Bezug auf die Einführung der ländergemeinsamen Abituraufgabenpools soll erreicht werden, dass trotz geringer Fallzahl und des explorativen Vorgehens, eine hohe Aussagekraft der Daten in Bezug auf das Erkenntnisinteresse und die Forschungsfragen erzielt wird.

Die vorliegenden Verbaldaten wurden in vollem Umfang („full transcription") und nach dem semantisch-inhaltlichen Transkriptionssystem (Dresing & Pehl, 2018) verschriftlicht. Die Auswertung des Materials folgte dem Vorgehen der inhaltlich strukturierenden qualitativen Inhaltsanalyse (Kuckartz, 2016) und dem hiermit verbundenen Ziel der regelgeleiteten Systematisierung von Kommunikationsinhalten. Die hierzu notwendigen Kategorien zur Kodierung der Interviewdaten wurden in einem (theoriebasiert) deduktiv-(datengesteuert-)induktiven Verfahren entwickelt (Döring & Bortz, 2016, S. 542), das Kategoriensystem wurde durch das Autorenteam in einem mehrstufigen Prozess der Erprobung und Diskussion (konsensuelle Codierung z. B. Kuckartz, 2018, S. 206 ff.) mehrmals überarbeitet.

Die auf dieser Grundlage generierten bundeslandspezifischen Erkenntnisse („within case") werden des Weiteren noch fallübergreifend („cross case") und mit dem Ziel der Herstellung einer analytischen Gesamtperspektive auf die Pools und die entsprechenden komplexen Strukturzusammenhänge und Prozessverläufe (Pflüger, Pongratz & Trinczek, 2017, S. 390) aufeinander bezogen, sodass das hier gewählte Vorgehen insgesamt als „multiple-case"-Fallstudie (Yin, 2018) einzuordnen ist.

Auf Basis der bisherigen praxisbezogenen Erläuterungen und der theorieorientierten Konkretisierung des Forschungsgegenstandes ergeben sich für die in diesem Beitrag fokussierten Akteure der KMK nunmehr folgende forschungsleitenden Fragestellungen:

1. Welche konkreten Intentionen werden auf KMK-Ebene mit dem gemeinsamen Beschluss zur Einführung der Abituraufgabenpools verfolgt und in welchem Maße sollen die den Pools inhärenten umfassenden Standardisierungspotenziale regulativ ausgeschöpft werden?
2. Welchen rekontextualisierenden Umgang mit der gemeinsam aggregierten (Standardisierungs-)Norm nimmt man seitens der Amtschefs auf Länderebene wahr?
3. Welche Rolle spielt die Kulturhoheit der Länder als Strukturprinzip für eine erfolgreiche, bundesweite Standardisierung von Anforderungen über ländergemeinsame Aufgabenpools?

Die Unterscheidung spezifischer Forschungsfragen an dieser Stelle zeigt den inhaltlichen „roten Faden" der Analyse auf; gleichzeitig sind die hier angesprochenen Aspekte jedoch analytisch nicht distinkt voneinander abgrenzbar, sondern als komplexe Formen der Handlungskoordination nur im jeweiligen Bezug zueinander zu verstehen. Entsprechend steht auch der hier virulent werdende institutionelle Gesamtzu-

sammenhang im Fokus der Analyse, was zugleich bedeutet, dass die Beantwortung der Fragen im Ergebnisteil nicht kursorisch erfolgt. Die Darstellung der Ergebnisse berücksichtigt vielmehr jene Reziprozität und ist so insgesamt an einer Logik der Nachvollziehbarkeit der Forschung (z. B. Jansen, 2019) ausgerichtet. Zur Ergebnisdarstellung ist zu bemerken, dass die Analyse und Interpretation immer mithilfe von und im Rückgriff auf die dargelegten theoretischen Konzepte und Terminologien erfolgt, womit eine entsprechende „theoretische Generalisierbarkeit der Ergebnisse" (Aljets, 2015, S. 122) erreicht werden kann. Diese für den Ergebnisteil prägende Art der Darstellung geht gleichzeitig mit einer ständigen Rückbindung der Interpretation an die vorliegenden Daten, verstanden als Prozess des methodisch kontrollierten Fremdverstehens (Kelle, 2019, S. 167), einher. Die Forschungsfragen können so sowohl theoretisch fundiert als auch mit explizitem Bezug auf die getätigten Interviewaussagen beantwortet werden. Der Einbezug solcher Akteursaussagen wird praktisch realisiert durch die Darstellung entsprechend aussagekräftiger Zitate. Diese sollen als illustratives Moment (Kruse, 2015, S. 637) dazu beitragen, die theoretisch orientierten Befunde praktisch zu belegen und so zu verdeutlichen, „how the findings and interpretations have arisen from the data" (Eldh, Årestedt & Berterö, 2020, S. 3). Zitate tragen folglich zur intersubjektiven Überprüfbarkeit der bis dahin noch in erster Linie als „theoretische Abstraktionen" (Przyborski & Wohlrab-Sahr, 2014, S. 409) vorliegenden Ergebnisse bei, können aber gleichzeitig nicht für sich selbst sprechen und bedürfen einer konzeptuell-theoretischen Basis.

Insgesamt basiert die im nächsten Abschnitt vorgenommen aktive Deutung der Daten somit auf einem angenommenen reziproken Verhältnis von Theorie und Empirie (Kalthoff, 2019).

4 Erste empirische Ergebnisse zur Einführung der Gemeinsamen Abituraufgabenpools aus Sicht der KMK

4.1 Institutioneller Kontext der Standardisierungsbemühungen

Zunächst ist in Bezug auf die in diesem Beitrag fokussierten Akteure der KMK festzuhalten, dass die Amtschefskommission ein *„Bindeglied" [AC2: 12*[5]*]* zwischen der reinen Fach- und der politischen Entscheidungsebene innerhalb der KMK darstellt; hiermit haben die beiden Amtschefs den umfassendsten (Experten-)Blick auf die angesprochenen Standardisierungsbemühungen.

Auf Amtschefs-Ebene weist man zunächst auf die grundsätzliche Bedeutung des Forschungsgegenstandes hin und betont, dass das Abitur *„ein bildungspolitisch brisantes und immer sehr virulentes Thema" [AC1: 12]* ist; alle getroffenen Entscheidungen in diesem Kontext erzielen folglich eine unmittelbare Außenwirkung und ziehen entsprechende (De-)Legitimationen wesentlicher Akteure nach sich. Dies liegt nach eige-

5 An dieser Stelle wird immer auf den jeweiligen Absatz im entsprechenden Interviewtranskript Bezug genommen, womit der Notwendigkeit Rechnung getragen wird, dass Zitate immer klar im entsprechenden Datenmaterial verortet werden müssen (z. B. Kuckartz, 2016).

ner Einschätzung darin begründet, dass zunächst grundsätzlich *„alle gesellschaftlichen Gruppierungen" [AC2: 24]* an den Entwicklungen zum Abitur interessiert sind, da mit der AHR eben nicht nur ein Schulabschluss vergeben wird, sondern dass dieser *„Abschluss […] gleichzeitig eine Berechtigung" [AC1: 12]* in Form des Hochschulzugangs darstellt. In Bezug auf die hiermit in Verbindung gebrachte Selektionsentscheidung, die in der öffentlichen Wahrnehmung großen Einfluss auf den weiteren Lebensweg hat und in diesem Kontext sogar z. T., so die Wahrnehmung der Amtschefs, als notwendige Bedingung für *„Lebensglück und die Perspektive im Leben" [AC1: 12]* betrachtet wird, kommt ein entsprechender *„Gerechtigkeitsaspekt" [AC2: 24]* zum Tragen; konkret betonen aus Sicht der Amtschefs hier insbesondere Eltern, die das Abitur *„ganz massiv im Blick" [AC2: 24]* haben, auf der einen Seite die Wichtigkeit, *„dass jedes ihrer Kinder möglichst auch das Abitur macht" [AC1: 12]*, wobei auf der anderen Seite gleichzeitig das jeweils landesinterne Abitur gleichzeitig als qualitativ hochwertig gelten soll und konkret *„nicht weniger wert sein darf als in anderen Ländern" [AC1: 12]*; dies könnte sich wiederum auf die tatsächliche Studienmöglichkeit für die eigenen Kinder, gerade *„wenn es um NC-Fächer geht" [AC2: 24]*, entsprechend negativ auswirken. Hier wird deutlich, dass bei jeder Veränderung des regulativen Elements der Institution, unmittelbar entsprechende normative und/oder pragmatische Legitimitätszuschreibungen für die AHR zur Disposition stehen. Beide hier formulierten institutionellen Ansprüche und die sich manifestierenden Sinnlogiken tangieren letztlich dieselbe Frage, und zwar die nach den tatsächlich in den Bundesländern gestellten inhaltlichen Anforderungen und den damit assoziierten Kompetenzniveaus in der Abiturprüfung selbst und in der vorgeschalteten Qualifikationsphase der Oberstufe. Diesbezüglich stellt man auf Amtschefs-Ebene fest: *„Das Abitur leistet nicht das, was es soll" [AC2: 22]*. Diese normativ-institutionelle De-Legitimation bezieht sich darauf, dass, trotz der auf Ebene der Formalstruktur nachweisbaren regulativ-isomorphen Bemühungen aller Bundesländer, inhaltliche Anforderungen bundesweit über die Implementation entsprechender Bildungsstandards zu homogenisieren und so letztlich auch die AHR nicht nur formal, sondern auch inhaltlich vergleichbar zu machen *(„Die Kompetenzen, die jemand erworben hat, müssen absolut vergleichbar sein." [AC2: 28])*, man mit der tatsächlichen Standardisierungsleistung dieser Institution noch nicht zufrieden ist *(„Ich hatte mir auch ganz ehrlich mehr von denen [den Bildungsstandards] versprochen, dass mehr Vergleichbarkeit entsteht." [AC2: 28])*. Diese institutionelle Be- bzw. Entwertung bedeutet die Ablösung einer bis dato auf KMK-Ebene geltenden Sinnlogik, auf Basis welcher eine Vergleichbarkeit der Anforderungen gar nicht aktiv hergestellt werden muss, sondern sich zu großen Teilen quasi automatisch in allen Bundesländern ergibt *(„Wir haben sehr stark, sehr lange [darauf] gesetzt: Jedes Land macht es so, wie es das für richtig hält und am Ende kommt da schon was Vernünftiges bei raus." [AC1: 10])*. In der Praxis wurde vielmehr eine entsprechende Entkopplung von Formal- und Inhaltsstruktur beobachtet, und zwar insofern als gemeinsam vereinbarte Standards zwar regulativ in allen Ländern institutionalisiert waren, aber gleichzeitig nicht in dem Maße auch inhaltlich zur Anwendung kamen, wie das auf Grundlage dieses regulativen Status quo erwartbar und möglich gewesen wäre *(„Wir wussten ja gar nicht, wie*

unterschiedlich zwischen den Ländern auch Standards sind." [AC2: 14]). Um jene – gleichfalls mit einem klaren externen Legitimationsdefizit aus der gesellschaftlichen Öffentlichkeit assoziierte *(„Die Gesellschaft akzeptiert das nicht mehr, diese große Varianz." [AC2: 18])* – Entkopplung in Form einer policy-practice-gap (Bromley & Powell, 2012) zu überwinden, die Implementation der Bildungsstandards entsprechend erfolgreich zu gestalten und letztlich auch auf inhaltlich-qualitativer Ebene das Kompetenzniveau der Schülerinnen und Schüler so zu homogenisieren, dass bundesweit einheitlich eine entsprechende *„Studierfähigkeit" [AC1: 40]* nachgewiesen werden kann, findet auf KMK-Ebene ein neuerlicher, normativ geprägter sense-making-Prozess statt, welcher folgender Standardisierungslogik folgt: Jenem funktionalen Defizit soll über eine De-Institutionalisierung tradierter (Prüfungs-)Normen auf Länderebene und einer entsprechend deutlicher standardsichernd ausgerichteten gemeinsamen Re-Institutionalisierung des schriftlichen Abiturprüfungsverfahrens begegnet werden. Hierbei gelten die Bildungsstandards weiter als notwendige *(„Bildungsstandards: Grundvoraussetzung." [AC2: 30])*, gleichsam nicht hinreichende *(„Bildungsstandards alleine konnten es nicht sein." [AC2: 30])* Institution zur Herstellung von Vergleichbarkeit. Der *„nächste folgerichtige Schritt" [AC1: 70]* der Ergänzung in diesem Standardisierungskontext betrifft dann konkret die Entwicklung und den Einsatz gemeinsamer Abiturprüfungsaufgaben *„mit denen man diese Standards überprüfen kann" [AC1: 44]*.

4.2 Die Gemeinsamen Abituraufgabenpools als institutionelle Standardisierungsmaßnahme

Die Entwicklung und der Einsatz ländergemeinsamer Abituraufgaben wird in diesem institutionellen Kontext also auch aus praktischer Sicht als notwendige bildungspolitische Innovation und entsprechend nächster Schritt zur Herstellung bundesweit vergleichbarer Anforderungen eingeordnet. Dieser basiert nach Aussage der Amtschefs auf einem von allen Bundesländern gemeinsam gestalteten und verantworteten Prozess der *Über*arbeitung individueller, von den einzelnen Ländern zunächst intern entwickelten Abiturprüfungsaufgaben. Jene Aufgaben werden dann von allen Ländern gemeinsam, in einem *„mehrere Schleifen" [AC2: 99]* umfassenden Prozess der Qualitätssicherung dahin gehend geprüft, ob diese für den Prüfungseinsatz als geeignet eingeschätzt werden *(„Welche taugen?" [AC2: 99])*. Eine gemeinsame inhaltliche Modifikation ist an dieser Stelle nicht die Ausnahme, sondern vielmehr die Regel, denn die Aufgaben werden *„überarbeitet, ergänzt, herangezogen und andere verworfen" [AC2: 99]*. Am Ende dieses Qualitätssicherungsprozesses, welcher vom IQB koordiniert, allerdings nicht von diesem inhaltlich verantwortet wird *(„Die [Aufgaben] werden nicht vom IQB erstellt." [AC2: 32])*, werden die so entstandenen Aufgaben von Seiten der KMK umfassend inhaltlich-normativ legitimiert und im Sinne der beschriebenen Intention als zur Herstellung vergleichbarer Anforderungen geeignete Institution eingeschätzt *(„Am Ende sind es vergleichbare Aufgaben vom Schwierigkeitsgrad, von den Kompetenzen und so weiter." [AC2: 99])*. Insgesamt kann auf dieser Basis also von einer umfassenden inhaltlichen Beteiligung aller Bundesländer am Prozess der Aufgaben-

erstellung gesprochen werden. Der hier gelegte Fokus auf eine entsprechend *gemeinsame* und gleichzeitig qualitativ hochwertige Standardisierungsarbeit *(„Jedes Land schickt seine Besten." [AC2: 12])* im Sinne eines symbiotischen Prozesses der Zusammenarbeit (Gräsel & Parchmann, 2004) kommt strukturell klar zum Ausdruck. In diesem Kontext kann man also in Bezug auf die Art und Weise der Institutionalisierung des Prozesses der Länderzusammenarbeit, an dessen Ende die fertigen Prüfungsaufgaben stehen, von einem idealtypischen Vorgehen sprechen, wenn es um die gemeinsame Institutionalisierung einer standardsichernden Maßnahme wie den Abituraufgabenpools geht. Ein entsprechender isomorpher Übersetzungsprozess dieser gemeinsam aggregierten Norm in (regulative) Handlung, also der flächendeckende Einsatz der in die Pools eingegebenen Aufgaben in den Abiturprüfungen der Länder, womit gleichzeitig eine entsprechend hohe Standardisierungsleistung einherginge, wäre auf dieser Basis dann der nächste logische Schritt. In der praktischen Umsetzung auf Länderebene ist dann zunächst auch ein entsprechender institutioneller Isomorphieprozess nachweisbar: Die Pools bzw. die entsprechenden Aufgaben werden *von allen* Bundesländern als fester Bestandteil der eigenen schriftlichen Abiturprüfung zur Anwendung gebracht, man entspricht also auf Intermediärer Ebene regulativ dem normativ geprägten Re-Institutionalisierungsimpetus auf KMK-Ebene. Gleichwohl werden bereits an dieser Stelle und im Sinne der angestrebten Standardisierung transintentionale institutionelle Entkopplungsprozesse deutlich. Diese vorgenommenen Rekontextualisierungen durch die Kultusministerien tangieren dann nicht die Frage der Implementation selbst („ob"), sondern betreffen die *Art und Weise* des Einsatzes der Poolaufgaben in der eigenen Prüfung („wie"). Hier kommt zum Ausdruck, dass der durch die Amtschefs artikulierte institutionelle Gestaltungswille auf KMK-Ebene, welcher aufgrund entsprechend fehlender regulativer Verfügungsrechte immer genuin auf das normative Element einer Institution beschränkt sein muss, keine eigene Standardisierungsleistung zeitigt. Konkret wird vonseiten der Amtschefs dann auch darauf hingewiesen, dass regelmäßig von Länderseite *„die Aufgaben unterschiedlich stark gezogen und unterschiedlich stark verändert wurden" [AC1: 18]*, womit man gleichzeitig die in der entsprechenden, im Oktober 2020 wie erläutert diesbezüglich novellierten, KMK Vereinbarung (KMK, 2017) gemeinsam beschlossenen Freiräume in der regulativen Umsetzung entsprechend nutzt. Die beiden wesentlichen hier angesprochenen Aspekte, bei welchen die Länderakteure ihre zuerkannten regulativen Verfügungsrechte entsprechend ausnutzen und so die den Pools inhärenten Standardisierungspotenziale nicht ausschöpfen, betreffen zunächst den *Anteil der aus dem Pool entnommenen Aufgaben* sowie die *entsprechende Anpassung der Aufgaben* selbst. In Bezug auf die Quote der tatsächlich in der Prüfung zur Anwendung kommenden Poolaufgaben, attestiert man vonseiten der Amtschefs den Aufgabenpools eine bis dato verbesserungswürdige Standardisierungsleistung, da, mit der Ausnahme von zwei Bundesländern mit einer jeweils höheren Quote, *„die Länder [im Prüfungsjahr 2019] einen Anteil von Aufgaben gezogen [haben], der unter 20 Prozent lag [AC1: 68]*. Das bedeutet im Umkehrschluss, dass die Mehrheit der in den Ländern eingesetzten Aufgaben die gemeinsame Qualitätsschleife am IQB gar nicht durchlaufen haben, was die ent-

sprechenden Heterogenitäten in den Anforderungen, trotz der neu existierenden Institution der Aufgabenpools, weiter stabilisiert und nicht zu einer Homogenisierung selbiger beiträgt.

Eine ähnlich gelagerte Entkopplung von Struktur und Handlung zeigt sich, wenn auf inhaltlicher Ebene die Bundesländer die fertigen Poolaufgaben, bevor sie diese einsetzen, noch einmal auf Basis eigener landesinterner Logiken verändern; hier wird konkret die ländergemeinsam geschaffene Struktur zur Qualitätssicherung am IQB durch weiterhin zur Anwendung gebrachte landesinterne Mechanismen der Rekontextualisierung zumindest teilweise überlagert. Ein konkretes Praxisbeispiel einer solchen institutionellen Vorgehensweise zeigt sich nach Aussage der Amtschefs dann, wenn ein Land sich zwar am erläuterten Prozess am IQB beteiligt, also eigene Aufgaben in den Qualitätssicherungsprozess einbringt und auch die entsprechenden gemeinsamen Modifikationen mitträgt, d. h. normativ legitimiert. Wenn es dann jedoch um den tatsächlichen Einsatz dieser entstandenen Poolaufgabe geht, wird gleichfalls nicht diese in der eigenen Prüfung eingesetzt, sondern die ursprünglich *intern* entwickelte *(„[…] wenn einer dann eine Aufgabe reinbringt, die dann so verändert wird, dass sie diesen Standards auch genügt, und anschließend die Aufgabe gezogen wird und wieder zurück verändert wird." [AC1: 48])*. Diese Vorgehensweise stellt ein prototypisches Beispiel für den von Meyer und Rowan beschriebenen institutionellen Vorgang der aktiven Entkopplung von Struktur- („talk") und Handlungsebene („action") dar, denn die Pools werden so bewusst nicht regulativ, doch nachweisbar normativ de-legitimiert. Solche institutionellen Übersetzungs- und Editierungsprozesse als spezifische Form des „window dressing" und die so nachweisbare Policy-practice-Entkopplung wird als klar transintentional im Sinne der Zielerreichung eingeschätzt *(„Dann brauche ich sie auch nicht reingeben. Das kann ich auch lassen." [AC1: 47])*. Insgesamt sprechen diese länderinternen und in Bezug auf das ursprünglich gemeinsame formulierte Ziel der Herstellung vergleichbarer Anforderungen im Abitur kontraproduktiven Übersetzungsprozesse dafür, dass tatsächlich, auch wenn man auf Grundlage der Art und Weise der Aufgabenerarbeitung, der daran beteiligten Akteure und des Defacto-Einsatzes der Poolaufgaben in den Abiturprüfungen aller Länder das Gegenteil hätte vermuten können, nicht in jedem Fall eine ausreichend breite normative Legitimitätsbasis für diese vorhanden ist. Vielmehr kann angenommen werden, dass in den Bundesländern tradierte, kulturell-kognitive Sinnlogiken zur Herstellung qualitativ hochwertiger Bildung in Bezug auf das Abitur und die Verleihung der AHR vorliegen, welche eine übergreifende Homogenisierung von Anforderungen über das regulative Element hinaus aufgrund ihrer institutionellen Persistenz zumindest stark erschweren (vgl. für die hier konkret nachweisbaren Handlungslogiken auf Länderebene Groß, i.V.).

Auf Basis eines so beschriebenen Umgangs mit den Poolaufgaben könnte sich ggf. eine ähnliche Entwicklung abzeichnen, wie sie bereits in Bezug auf die Bildungsstandards der Fall ist; hier wurde eine ländergemeinsam eingeführte Institution, welche auf Handlungsebene jedoch nicht die Standardisierungswirkung zeitigt, die man sich auf Basis des vorhandenen Potenzials erhofft hatte, institutionell um ein weiteres

Instrument zur Zielerreichung, hier gemeinsame Poolaufgaben, ergänzt. Um einem solchen Kreislauf immer aufs Neue notwendiger Institutionalisierungen mit dem Ziel der Herstellung bundesweit vergleichbarer Anforderungen entgegenzuwirken, ist es aus Sicht der Amtschefs wesentlich, dass die institutionellen Standardisierungspotenziale der Pools auch entsprechend ausgeschöpft werden. Das bedeutet konkret, dass die Länder auf die aktuell noch gemeinsam vereinbarte Möglichkeit der Modifikation, die neben der Aufgabe selbst auch den mitentwickelten Bewertungsmaßstab betrifft, in Zukunft verzichten sollen. Hier hält man konkret fest, dass *„die Aufgaben, die gezogen werden, auch eingesetzt werden, wie sie da drinliegen, und da wird weder der Bewertungsmaßstab noch die Aufgabe inhaltlich verändert" [AC1: 48]*. Diese wiederholt angeführte Intention *(„Das muss das Ziel sein: ein unverändertes Ziehen aus dem Aufgabenpool." [AC2: 85])* zeigt deutlich, dass sowohl der Prozess der Zusammenarbeit am IQB als auch die daraus entstehenden gemeinsamen Poolaufgaben als Output dessen in hohem Maße normativ legitimiert sind.

Dass die Herstellung vergleichbarer Anforderungen also keineswegs nur auf rein formaler Ebene („talk") stattfinden soll, zeigt sich des Weiteren darin, dass man sich auf KMK-Ebene auch auf eine feste bundesweite Quote der zu ziehenden Aufgaben einigen möchte, wobei an dieser Stelle die entsprechenden Potenziale jedoch nicht vollumfänglich ausgeschöpft werden sollen *(„[...] dass wir die Hälfte mindestens ziehen und die andere Hälfte durch ländereigene Aufgaben ergänzt werden kann." [AC1: 86])*.

Während man so also auf der einen Seite auf Makroebene insgesamt die Wichtigkeit einer *tatsächlich* geteilten Norm betont, auf deren Grundlage es erst zu Isomorphieprozessen auf Handlungsebene und somit der intendierten Standardisierung kommen kann, wird ebenso deutlich, dass auch transintentionale Übersetzungsprozesse auf Länderebene entstehen, welche auf der erläuterten Art und Weise der Institutionalisierung von Verfügungsrechten beruhen; ein entsprechendes Umsetzen eines ländergemeinsamen Beschlusses wie vereinbart, ist aufgrund dessen also nicht von Vornherein gewährleistet *(„Wir beschließen, wir haben Beschlüsse unserer Kultusministerkonferenz und jetzt kommt es: Wie verbindlich ist das?" [AC2: 16])*. Aufgrund der fehlenden regulativ-institutionellen Bindewirkung der Beschlüsse auf Makroebene wird vonseiten der Amtschefs sogar auf die nicht in der Form beobachtete, aber gleichwohl als theoretische Möglichkeit beschriebene Gefahr einer vollumfänglichen Entkopplung auf Länderebene hingewiesen *(„Ein Land sagt: Ihr macht den Pool, wir machen gar nichts." [AC1: 92])*.[6] Hiermit wird das wesentliche Strukturprinzip der Kulturhoheit der Länder als grundsätzliche institutionelle Herausforderung begriffen, und zwar gerade auch deswegen, da die konkrete Zusammenarbeit der Länder in der KMK ohnehin bereits von der schwierigen institutionellen Voraussetzung geprägt ist, unterschiedliche Handlungsroutinen und Fächerlogiken der Länder bei der Erstellung der Aufgaben zu berücksichtigen. Diese persistenten Logiken aufzulösen wird

6 Ein solcher Umgang war z. B. beobachtbar im Fall Niedersachsens, wo seit dem Schuljahr 2019/20 keine landesweiten Vergleichsarbeiten (VERA-3 und VERA-8) mehr durchgeführt werden, obwohl diese einen wesentlichen Teil der gemeinsam vereinbarten KMK-Gesamtstrategie zum Bildungsmonitoring darstellen. Vgl. hierzu https://www.mk.niedersachsen.de/startseite/schule/schulqualitat/externe_evaluation/vergleichsarbeiten_vera/vergleichsarbeiten-vera-135419.html [zuletzt abgerufen am 01.06.2022].

als *„eine riesengroße kommunikative Aufgabe" [AC1: 28]* eingeschätzt, da z. T. auf dieser Grundlage entsprechende *„Glaubenskriege" [AC1: 28]* zwischen den Bundesländern beobachtet werden, wenn es darum geht, wie qualitativ hochwertige Abiturprüfungsaufgaben auszusehen haben *(„Unsere Aufgaben [...] sind besser geeignet, das zu überprüfen als eure Aufgaben." [AC1: 28])*; so entstehen die Poolaufgaben am Ende weniger auf Basis eines gemeinsamen inhaltlichen Konsenses *(„Nein, Nein" [AC1: 30])*, sondern stellen vielmehr einen *„Kompromiss" [AC1: 30]* dar, welcher sich im Idealfall als *„tragfähig" [AC1: 30]* erweist und dann entsprechend ebenso legitimiert ist.

Auf Basis solcher erwartbaren inhaltlichen Divergenzen kann aus Sicht der Amtschefs auch immer konkreter *„Widerstand gegen diese Gemeinsamkeiten" [AC1: 40]* erwachsen, was wiederum auf entsprechende De-Legitimierungsvorgänge auf Länderebene hindeutet, welche die Erarbeitung des intendierten tragfähigen inhaltlichen Kompromisses erschwert. Insgesamt kann auf dieser Grundlage zusammenfassend festgehalten werden, dass aus der Existenz persistenter Länderlogiken, gepaart mit den auf Intermediärer Ebene verorteten Verfügungsrechten und der entsprechenden Möglichkeit, diese auch zur Anwendung zu bringen, ein Implementationsrisiko für die Aufgabenpools und das hiermit verbundene Ziel der Herstellung vergleichbarer Anforderungen erwächst.

4.3 Die Kulturhoheit der Länder als umfassend legitimierte Standardisierungsherausforderung

In den bisherigen Ausführungen kam sowohl aus theoretischer Sicht als auch auf Basis entsprechender Praxisbeobachtungen der Amtschefs zur Einführung ländergemeinsamer Abituraufgabenpools klar zum Ausdruck, dass Bildungsreformen per se und auch dieser bildungspolitischen Neuerung im Speziellen, ein entsprechendes Risiko des Scheiterns innewohnt. Im konkreten Fall stellen auf der einen Seite die ins Werk gesetzten regulativen Übersetzungen der normativ geprägten Reformaufforderung durch die Kultusministerien bereits die entsprechenden Weichen für eine erfolgreiche Standardisierung von Anforderungen, gleichzeitig werden vonseiten der Amtschefs konkret entsprechende Rekontextualisierungen auf Länderebene beobachtet, die dieser Standardisierungslogik entgegenlaufen. Während dergestalte Übersetzungsprozesse wie erläutert de-legitimiert werden, ist die Kulturhoheit der Länder als hier wesentlich zum Tragen kommendes Strukturprinzip gleichzeitig trotzdem umfassend und im institutionentheoretischen Sinne sogar als unhinterfragte Selbstverständlichkeit legitimiert. Seitens der Amtschefs hält man hier entsprechend pointiert fest: *„Wir sind Bildungsföderalisten." [AC2: 119].* Die hier zugrunde gelegte Überzeugung, dass *„aufgrund der Hoheit auch niemand beim anderen mitzuschnabeln hat" [AC2: 14]*, legt eine aktive Legitimation *(„ist auch gut so" [AC2: 14])* für fest institutionalisierte Unterschiedlichkeiten *(„historisch gewordene Eigenheiten" [AC1: 10]; „Gebräuche" [AC1: 10]; „regionale Spezifika" [AC2: 5])* auf Länderebene offen. Insgesamt wird hierbei die grundsätzliche Notwendigkeit der länderübergreifenden Zusammenarbeit wiederum betont *(„Wer zum Bildungsföderalismus steht, muss zur KMK stehen." [AC2: 18])*, trotzdem wird gleichermaßen die Tatsache legitimiert, dass lediglich ein Handlungs-*„Rahmen für alle Länder [existiert], der [...] ganz unterschiedlich ausgefüllt werden [kann]" [AC2:*

18]. Im Kontext der Organisation des Abiturs manifestiert sich dann nach eigener Einschätzung die entsprechende *„Stärke des Föderalismus" [AC1: 10]* konkret insofern als man so in den einzelnen Bundesländern auf die Menschen vor Ort und deren Lebenswelten und Bedürfnisse, die sich innerhalb Deutschlands klar voneinander unterscheiden, inhaltlich adäquat eingehen kann *(„[…] das Fach Französisch im Saarland hat eben eine herausragende Bedeutung, in Schleswig-Holstein hat es das so nicht." [AC1: 10])*. Insgesamt bemerkt man auf Amtschefs-Ebene selbst, dass man mit dem vorgegebenen Ziel der bundesweiten Standardisierung von Anforderungen und dem gleichzeitig immer noch voll institutionalisierten Strukturprinzip der Kulturhoheit der Länder, letztlich von zwei grundsätzlich *„konkurrierenden Pole[n]" [AC1: 10]* sprechen muss; jene jeweils legitimierten, nichtsdestoweniger kontradiktorischen Intentionen der *externen Standardisierung* auf der einen Seite bei gleichzeitiger Möglichkeit der *internen Individualisierung* auf der anderen, zeigen das sich hier manifestierende institutionelle Spannungsfeld auf, in welchem die Einführung der Aufgabenpools bereits zu Beginn des Institutionalisierungsprozesses auf Makroebene einzuordnen ist. In diesem institutionellen Gefüge sind die auf KMK-Ebene beschlossenen Standardisierungsbemühungen aus Sicht der Amtschefs dann so zu verstehen, dass hier ein entsprechendes institutionelles Ungleichgewicht in Richtung einer Überbetonung individueller Länderlogiken bestand, welches nun entsprechend durch die Institutionalisierung weiterer Maßnahmen zur Herstellung von Vergleichbarkeit, wie in dem Fall den Aufgabenpools, ausgeglichen werden soll *(„Das tun wir im Moment gerade mehr." [AC1: 10])*. Um vergleichbare Anforderungen zwischen den Bundesländern zu erreichen, muss aus Sicht der Amtschefs externe Standardisierung also nicht notwendigerweise mit der Ablösung interner Individualisierungsspielräume einhergehen; es geht vielmehr um die gleichberechtigte Anwendung und somit institutionelle Reproduktion und Stabilisierung beider Prinzipien.

5 Zusammenfassung und Ausblick

Insgesamt kommt auf Basis der bisherigen Ergebnisse zur Implementation der Pools zum Ausdruck, dass auf Amtschefs-Ebene der KMK ein klar nachweisbarer Standardisierungswille hin zur Herstellung bundesweit vergleichbarer Anforderungen im Abitur besteht; hierbei kommt die grundlegende Handlungslogik zum Tragen, dass *„wir ja wegkommen wollen davon, dass jeder macht, was er will" [AC1: 70]*. Diese De-Legitimation umfassender individueller Sonderwege der Länder beruht auf der Einschätzung, dass man seitens der KMK als bildungspolitisches Organ nur so auch nach außen eine entsprechende *„Handlungsfähigkeit unter Beweis stellen" [AC1: 56]* kann, wenn es um die adäquate und gemeinsame Bewältigung der erläuterten Vergleichbarkeitsherausforderung geht. Die aus diesem Grund beschlossene *„überfällig[e]" [AC1: 56]* Standardisierung über gemeinsame Poolaufgaben soll dann auch über ein reines „window dressing" einer lediglich vermeintlichen Homogenisierung hinausgehen und konkret auch auf inhaltlicher Ebene im Sinne eines Einsatzes der Aufgaben wie

in den Pool eingegeben und einer fest vereinbarten Aufgabenquote wirken. Gleichzeitig haben diese Bemühungen ihre institutionellen und auch legitimatorischen Grenzen dort, wo die regulative Kompetenz der Bundesländer ins Spiel kommt, deren umfassende Handlungspotenziale als legitimierte Selbstverständlichkeit angesehen werden und demgemäß erhalten werden sollen. Aus dieser aktiven institutionellen Stabilisierung der Möglichkeit, weiterhin individuelle Sonderwege zu beschreiten, resultiert, dass eine erfolgreiche gemeinsame Standardisierung von Anforderungen immer notwendigerweise auf *tatsächlich* geteilten Sinn- und Handlungslogiken der Bundesländer beruhen muss, wenn diese über die Sichtebene der rein formalen Angleichung hinausweisen soll. Von solch geteilten Vorstellungen und einem entsprechenden Standardisierungserfolg der Pools ist man aufgrund der von Amtschefsseite beobachteten transintentionalen Rekontextualisierungen und der Anwendung entsprechender Individuallogiken auf Intermediärer Ebene insgesamt noch entfernt; gleichwohl sieht man sich mit den als notwendig erachteten Maßnahmen, welche dann, wie bereits angedeutet, im Oktober 2020 auch von den Ländern gemeinsam in der Form beschlossen wurden, aktuell und für die Zukunft auf einem guten und Erfolg versprechenden Weg *(„Alle sind an Bord, wir schreiten gemeinsam voran." [AC2: 117])* in Richtung der Herstellung einer *„annähernden Vergleichbarkeit" [AC1: 44]* der Anforderungen im Abitur und letztlich einem *„gemeinsamen Verständnis davon, wie jemand am Ende nachweisen soll, dass er etwas kann" [AC1: 40]*. Aufgrund des auch hier genuin normativen Charakters eines solchen Beschlusses, besteht weiterhin die Möglichkeit der institutionellen Entkopplung von „talk" und „action" und einer Priorisierung von Individuallogiken. Ob diese Vereinbarungen letztlich tatsächlich den praktischen Erfolg zeitigen, den man sich hier verspricht, oder ob sich ggf. auch hier transintentionale Rekontextualisierungen ergeben, bleibt hier noch offen und kann erst auf Basis der tatsächlichen Art und Weise der Übersetzung dieser Norm in Länderhandeln überprüft werden.

An dieser Stelle kann noch festgehalten werden, dass man auf Makroebene explizit und ganz allgemein den Wunsch formuliert, als Länder *„in der Bildung wieder zusammenzufinden" [AC2: 18]*. Auch wenn ein solches Governance-Vorhaben, wie die Standardisierung über gemeinsam entwickelte und eingesetzte Abiturprüfungsaufgaben aufgrund der dargelegten legitimatorischen Rahmenbedingungen und institutionellen Steuerungsrealitäten als per se komplex und herausfordernd einzuschätzen ist, sind kontinuierliche Standardisierungsbemühungen, wie z. B. die aktuell stattfindende Entwicklung von Poolaufgaben auch für die Naturwissenschaften (vgl. auch Beitrag 2 von Hoffmann et al. in diesem Band) und der Versuch zu ländergemeinsamen Lösungen zu gelangen, klar erkennbar.

Abschließend bleibt, im Sinne eines entsprechenden Ausblicks noch festzuhalten, dass es nun auf Grundlage dieser Analyse der frühesten Phase der Institutionalisierung ländergemeinsamer Poolaufgaben auf Makroebene der KMK im weiteren Verlauf dieses Forschungsprojekts von besonderem Interesse ist, wie auf den nächsten Systemebenen aktuell mit dieser Reformaufforderung tatsächlich handelnd umgegangen wird. Hierbei stehen insbesondere wieder die entsprechend handlungslei-

tenden Sinngebungsprozesse der einzelnen Akteure im Fokus. Im nächsten analytischen Schritt sollen, dem Implementationsprozess weiter folgend, die auf Intermediärer Ebene der Kultusministerien vorgenommenen regulativen Institutionalisierungen ebenso beschrieben und auf Basis des neo-institutionalistischen Theoriegerüsts auch erklärt werden. Diese bilden wiederum den regulativen Handlungsrahmen für den im weiteren Implementationsverlauf durch das Mehrebenensystem nachweisbaren Umgang mit den Poolaufgaben und entsprechend stattfindende (De-) Legitimationsprozesse durch zentrale Akteure auf Schulebene (siehe Abbildung 1).

Insgesamt ergibt sich letztlich so und auf Grundlage des entsprechenden Fallstudiendesigns ein umfassender, multidimensionaler Blick auf den Prozess der Einführung gemeinsamer Poolaufgaben, wobei insbesondere die vorhandenen Implementationsherausforderungen und institutionellen Reibungsverluste für alle Systemebenen und Akteure so herauskristallisiert werden können, wie das in diesem Beitrag und mit Bezug auf die Makroebene der KMK angedeutet wurde (vgl. Groß, i.V.).

Literatur

Aljets, E. (2015). *Der Aufstieg der Empirischen Bildungsforschung. Ein Beitrag zur institutionalistischen Wissenschaftssoziologie*. Wiesbaden: Springer VS.

Altrichter, H. (2008). Veränderungen der Systemsteuerung im Schulwesen durch die Implementation einer Politik der Bildungsstandards. In T. Brüsemeister & K.-D. Eubel (Hrsg.), *Evaluation, Wissen und Nichtwissen* (S. 75–115). Wiesbaden: VS Verlag für Sozialwissenschaften.

Altrichter, H. & Heinrich, M. (2007). Kategorien der Governance-Analyse und Transformationen der Systemsteuerung in Österreich. In H. Altrichter, T. Brüsemeister & J. Wissinger (Hrsg.), *Educational Governance. Handlungskoordination und Steuerung im Bildungssystem* (S. 55–103). Wiesbaden: VS Verlag für Sozialwissenschaften.

Altrichter, H., Heinrich, M. & Soukup-Altrichter, K. (2011). Schulprofilierung – Annäherungen an ein Phänomen. In H. Altrichter, M. Heinrich & K. Soukup-Altrichter (Hrsg.), *Schulentwicklung durch Schulprofilierung? Zur Veränderung von Koordinationsmechanismen im Schulsystem* (S. 11–45). Wiesbaden: VS Verlag für Sozialwissenschaften.

Altrichter, H. & Maag Merki, K. (2016). Steuerung der Entwicklung des Schulwesens. In H. Altrichter & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (S. 1–27). Wiesbaden: Springer VS.

Benz, A. (2007). Multilevel Governance. In A. Benz, S. Lütz, U. Schimank & G. Simonis (Hrsg.), *Handbuch Governance. Theoretische Grundlagen und empirische Anwendungsfelder* (S. 297–310). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bogner, A., Littig, B. & Menz, W. (2014). *Interviews mit Experten. Eine praxisorientierte Einführung*. Wiesbaden: Springer VS.

Bormann, I. (2014). Diskursanalyse als Verfahren einer wissensorientierten Governance-Forschung. In K. Maag Merki, R. Langer & H. Altrichter (Hrsg.), *Educational Governance als Forschungsperspektive. Strategien. Methoden. Ansätze* (S. 157–181). Wiesbaden: Springer VS.

Bosche, A. (2013). *Schulreformen steuern. Die Einführung neuer Lehrmittel und Schulfächer an der Volksschule (Kanton Zürich, 1960er- bis 1980er-Jahre)* (Bildungsgeschichte und Bildungspolitik, Band 4). Bern: hep.

Bosche, A. & Lehmann, L. (2014). Governance und die Suche nach Regelungsmechanismen. In K. Maag Merki, R. Langer & H. Altrichter (Hrsg.), *Educational Governance als Forschungsperspektive. Strategien. Methoden. Ansätze* (S. 237–257). Wiesbaden: Springer VS.

Bromley, P. & Powell, W. W. (2012). From Smoke and Mirrors to Walking the Talk: Decoupling in the Contemporary World. *Academy of Management Annals, 6*(1), 1–48.

Brüsemeister, T. (2008). *Qualitative Forschung. Ein Überblick* (2., überarb. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bundesverfassungsgericht (19.12.2017) 1 BvL 3/14, Absatz 1–253. Zugriff am 30.07.2021. Verfügbar unter: http://www.bverfg.de/e/ls20171219_1bvl000314.html

Dietrich, F. (2018). Konturen einer Rekonstruktiven Governanceforschung. In M. Heinrich & A. Wernet (Hrsg.), *Rekonstruktive Bildungsforschung. Zugänge und Methoden* (S. 73–94). Wiesbaden: Springer VS.

DiMaggio, P. J. & Powell, W. W. (1983). The Iron Cage Revisited. Institutional Isomorphism and Collective Rationality in Organizational Fields. *American Sociological Review, 48*(2), 147.

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Berlin: Springer Fachmedien.

Dresing, T. & Pehl, T. (2018). *Praxisbuch Interview, Transkription & Analyse. Anleitungen und Regelsysteme für qualitativ Forschende*. Zugriff am 23.07.2021. Verfügbar unter: https://www.audiotranskription.de/wp-content/uploads/2020/11/Praxisbuch_08_01_web.pdf

Eldh, A. C., Årestedt, L. & Berterö, C. (2020). Quotations in Qualitative Studies: Reflections on Constituents, Custom, and Purpose. *International Journal of Qualitative Methods, 19*, 1–6.

Fend, H. (2008). *Schule gestalten: Systemsteuerung, Schulentwicklung und Unterrichtsqualität*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Fuchs, H.-W. (2004). Schulentwicklung und Organisationstheorie: Welche Erklärungskraft besitzt die Bürokratietheorie heute? In W. Böttcher & E. Terhart (Hrsg.), *Organisationstheorie in pädagogischen Feldern. Analyse und Gestaltung* (Organisation und Pädagogik, Bd. 2, S. 206–220). Wiesbaden: VS Verlag für Sozialwissenschaften.

Gasterstädt, J. (2019). *Der Komplexität begegnen und Inklusion steuern: Eine Situationsanalyse zur Umsetzung von Artikel 24 der UN-BRK in Deutschland*. Wiesbaden: Springer Fachmedien.

GG– Grundgesetz der Bundesrepublik Deutschland, Art. 30. Zugriff am 01.06.2022. Verfügbar unter: https://www.gesetze-im-internet.de/gg/art_30.html

GG – Grundgesetz der Bundesrepublik Deutschland, Art. 70. Zugriff am 01.06.2022. Verfügbar unter: https://www.gesetze-im-internet.de/gg/art_30.html

Giddens, A. (1979). *Central Problems in Social Theory. Action, Structure, and Contradiction in Social Analysis*. Los Angeles: University of California Press.

Gräsel, C. & Parchmann, I. (2004). Implementationsforschung – oder: Der steinige Weg, Unterricht zu verändern. *Unterrichtswissenschaft, 32*(3), 196–214.

Graß, D. (2015). Legitimation neuer Steuerung: Eine neo-institutionalistische Erweiterung der Governance-Perspektive auf Schule und Bildungsarbeit. In J. Schrader (Hrsg.), *Governance von Bildung im Wandel. Interdisziplinäre Zugänge* (S. 65–93). Wiesbaden: Springer VS.

Hepp, G. F. (2011). *Bildungspolitik in Deutschland. Eine Einführung*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Herbrechter, D. & Schemmann, M. (2019). Educational Governance und Neo-Institutionalismus in der Weiterbildungsforschung. In R. Langer & T. Brüsemeister (Hrsg.), *Handbuch Educational Governance Theorien* (S. 181–199). Wiesbaden: Springer Fachmedien Wiesbaden.

Hoffmann, L., Schröter, P. & Stanat, P. (2022a). Jüngere Entwicklungen bei Abitur und Abiturprüfungen in Deutschland. In L. Hoffmann, P. Schröter, A. Groß, S. M. Schmid-Kühn & P. Stanat (Hrsg.), *Das unvergleichliche Abitur. Entwicklungen – Herausforderungen – Empirische Analysen* (S. 27–50). Bielefeld: wbv Publikation.

Hoffmann, L., Schröter, P. & Stanat, P. (2022b). Evaluation der Gemeinsamen Abituraufgabenpools der Länder. In L. Hoffmann, P. Schröter, A. Groß, S. M. Schmid-Kühn & P. Stanat (Hrsg.), *Das unvergleichliche Abitur. Entwicklungen – Herausforderungen – Empirische Analysen* (S. 109–128). Bielefeld: wbv Publikation.

Jansen, T. (2019). Gütekriterien in der qualitativen Sozialforschung als Form der Reflexion und Kommunikation. Eine Replik auf die Beiträge von Strübing et al. und Eisewicht & Grenz. *Zeitschrift für Soziologie, 48*(4), 321–325.

Jepperson, R. L. (1991). Institutions, Institutional Effects, and Institutionalism. In W. W. Powell & P. J. DiMaggio (Hrsg.), *The New Institutionalism in Organizational Analysis* (S. 143–163). Chicago/London: University of Chicago Press.

Kalthoff, H. (2019). Einleitung: Zur Dialektik von qualitativer Forschung und soziologischer Theoriebildung. In H. Kalthoff, S. Hirschauer & G. Lindemann (Hrsg.), *Theoretische Empirie. Zur Relevanz qualitativer Forschung* (3. Aufl., S. 8–34). Frankfurt a. M.: Suhrkamp.

Kelle, U. (2019). Mixed Methods. In N. Baur & J. Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung* (2. Aufl., S. 159–172). Wiesbaden: Springer VS.

Klatetzki, T. (2012). Professionelle Organisationen. In M. Apelt (Hrsg.), *Handbuch Organisationstypen* (S. 165–183). Wiesbaden: Springer VS.

Klemm, K. (2022). Die Geschichte der Allgemeinen Hochschulreife in Deutschland. In L. Hoffmann, P. Schröter, A. Groß, S. M. Schmid-Kühn & P. Stanat (Hrsg.), *Das unvergleichliche Abitur. Entwicklungen – Herausforderungen – Empirische Analysen* (S. 7–26). Bielefeld: wbv Publikation.

KMK- Ständige Konferenz der Kultusminister in der Bundesrepublik Deutschland (2017). *FAQs Gemeinsamer Abituraufgabenpool der Länder*. Zugriff am 01.06.2022. Verfügbar unter: https://www.kmk.org/themen/qualitaetssicherung-in-schulen/bildungsstandards/bildungsstandards-und-allgemeine-hochschulreife.html

KMK- Ständige Konferenz der Kultusminister der Bundesrepublik Deutschland (2020). *Ländervereinbarung über die gemeinsame Grundstruktur des Schulwesens und die gesamtstaatliche Verantwortung der Länder in zentralen bildungspolitischen Fragen (Beschluss der Kultusministerkonferenz vom 15.10.2020)*. Zugriff am 01.06.2022. Verfügbar unter: https://www.kmk.org/aktuelles/artikelansicht/kmk-verabschiedet-zukunftsweisende-laendervereinbarung-und-richtet-staendige-wissenschaftliche-kommiss.html

Kruse, J. (2015). *Qualitative Interviewforschung. Ein integrativer Ansatz*. Weinheim: Beltz Juventa.

Kuckartz, U. (2016). *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung*. Weinheim: Beltz Juventa.

Kühn, S. M. (2010). *Steuerung und Innovation durch Abschlussprüfungen?* Wiesbaden: VS Verlag für Sozialwissenschaften.

Lamnek, S. & Krell, C. (2016). *Qualitative Sozialforschung*. Weinheim, Basel: Beltz.

Langer, R. (2015). Educational Governance und Theoriebildung. In J. Schrader (Hrsg.), *Governance von Bildung im Wandel. Interdisziplinäre Zugänge* (S. 45–64). Springer VS.

Langer, R. & Brüsemeister, T. (2019). Ein Fazit aus der Theoriediskussion. In R. Langer & T. Brüsemeister (Hrsg.), *Handbuch Educational Governance Theorien* (S. 771–783). Wiesbaden: Springer Fachmedien.

Maag Merki, K. (2018). Reformen im Bildungswesen. In F. Imlig, L. Lehmann & K. Manz (Hrsg.), *Schule und Reform. Veränderungsabsichten, Wandel und Folgeprobleme* (S. 243–254). Wiesbaden: Vieweg.

Mannheim, K. (1980). *Strukturen des Denkens*. Frankfurt a. M.: Suhrkamp.

Meyer, J. W. & Rowan, B. (1977). Institutionalized Organizations. Formal Structure as Myth and Ceremony. *American Journal of Sociology, 83*(2), 340–363.

Muslic, B. (2017). *Kopplungen und Entscheidungen in der Organisation Schule: Organisationsbezogenes Schulleitungshandeln im Kontext von Lernstandserhebungen*. Wiesbaden: Springer VS.

Niedersächsisches Kultusministerium (Hrsg.). *Vergleichsarbeiten (VERA)*. Zugriff am 01.06.2022. Verfügbar unter: https://www.mk.niedersachsen.de/startseite/schule/schulqualitat/externe_evaluation/vergleichsarbeiten_vera/vergleichsarbeiten-vera-135419.html

Niedlich, S. (2019). Was kann die Strukturationstheorie zur Educational Governance-Forschung beisteuern? In R. Langer & T. Brüsemeister (Hrsg.), *Handbuch Educational Governance Theorien* (S. 351–376). Wiesbaden: Springer Fachmedien.

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. Newbury Park: Sage.

Pflüger, J., Pongratz, H. J. & Trinczek, R. (2017). Fallstudien in der Organisationsforschung. In S. Liebig, W. Matiaske & S. Rosenbohm (Hrsg.), *Handbuch empirische Organisationsforschung* (S. 389–413). Wiesbaden: Springer Gabler.

Przyborski, A. & Wohlrab-Sahr, M. (2014). *Qualitative Sozialforschung. Ein Arbeitsbuch* (4., erw. Aufl.). München: Oldenbourg.

Sahlin, K. & Wedlin, L. (2008). Circulating Ideas: Imitation, Translation and Editing. In R. Greenwood, C. Oliver, R. Suddaby & K. Sahlin (Hrsg.), *The SAGE Handbook of Organizational Institutionalism* (S. 218–242). London: Sage.

Schimank, U. (2007). Die Governance-Perspektive: Analytisches Potenzial und anstehende konzeptionelle Fragen. In H. Altrichter, T. Brüsemeister & J. Wissinger (Hrsg.), *Educational Governance. Handlungskoordination und Steuerung im Bildungssystem* (S. 231–260). Wiesbaden: VS Verlag für Sozialwissenschaften.

Schmid-Kühn, S. M. & Groß, A. (2022). Struktur der gymnasialen Oberstufe und Rahmenbedingungen für die Abiturprüfung im Ländervergleich. In L. Hoffmann, P. Schröter, A. Groß, S. M. Schmid-Kühn & P. Stanat (Hrsg.). *Das unvergleichliche Abitur. Entwicklungen – Herausforderungen – Empirische Analysen* (S. 51–76). Bielefeld: wbv Publikation.

Schütz, A. (1971). Über die Mannigfaltigen Wirklichkeiten. In A. Schütz (Ed.), *Gesammelte Aufsätze. I Das Problem der sozialen Wirklichkeit* (S. 237–298). Dordrecht: Springer Netherlands.

Scott, W. R. (2014). *Institutions and organizations. Ideas, Interests, and Identities.* London: Sage.

Sekretariat der Kultusministerkonferenz (Hrsg.). (2016). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring.* Köln: Carl Link.

Senge, K. (2011). *Das Neue am Neo-Institutionalismus.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Seo, M.-G. & Creed, W. E. D. (2002). Institutional Contradictions, Praxis, and Institutional Change: A Dialectical Perspective. *Academy of Management Review, 27*(2), 222.

Stanat, P. & Pant, H. A. (2013). *Konzeption für die Entwicklung und Nutzung eines Pools von Abiturprüfungsaufgaben. Arbeitsfassung.*

Strübing, J. (2018). *Qualitative Sozialforschung. Eine komprimierte Einführung.* Berlin: de Gruyter.

Suchman, M. C. (1995). Managing Legitimacy: Strategic and Institutional Approaches. *The Academy of Management Review, 20*(3), 571–610.

Süß, S. (2009). *Die Institutionalisierung von Managementkonzepten. Diversity-Management in Deutschland.* Mering: Rainer Hampp Verlag.

Tarkian, J. & Thiel, F. (2016). Steuerung im Bildungssystem (SteBis). Bilanz der Befunde aus der ersten Förderphase. In Bundesministerium für Bildung und Forschung (BMBF) (Hrsg.), *Steuerung im Bildungssystem. Implementation und Wirkung neuer Steuerungsinstrumente im Schulwesen* (S. 3–6). Berlin: BMBF.

Terhart, E. (2000). Zwischen Aufsicht und Autonomie. Geplanter und ungeplanter Wandel im Bildungsbereich. *Neue Sammlung. Vierteljahres-Zeitschrift für Erziehung und Gesellschaft, 40*(1), 123–140.

Türk, K. (1997). Organisation als Institution der kapitalistischen Gesellschaftsformation. In G. Ortmann, J. Sydow & K. Türk (Hrsg.), *Theorien der Organisation. Die Rückkehr der Gesellschaft* (S. 124–176). Wiesbaden: VS Verlag für Sozialwissenschaften.

Wacker, A. (2008). *Bildungsstandards als Steuerungsinstrumente der Bildungsplanung. Eine empirische Studie zur Realschule in Baden-Württemberg*. Bad Heilbrunn: Klinkhardt.
Walgenbach, P. (2019). Neoinstitutionalistische Ansätze in der Organisationstheorie. In A. Kieser & M. Ebers (Hrsg.), *Organisationstheorien* (S. 300–350). Stuttgart: Kohlhammer.
Walgenbach, P. & Meyer, R. (2008). *Neoinstitutionalistische Organisationstheorie*. Stuttgart: Kohlhammer.
Weber, M. (1980). *Wirtschaft und Gesellschaft. Grundriss der verstehenden Soziologie*. Tübingen: J. C. B. Mohr.
Weick, K. E. (1976). Educational Organizations as Loosely Coupled Systems. *Administrative Science Quarterly, 21*(1), 1–19.
Weick, K. E. (1995). *Sensemaking in organizations*. Thousand Oaks: Sage.
Wilson, T. P. (1980). Theorien der Interaktion und Modelle Soziologischer Erklärung. In Arbeitsgruppe Bielefelder Soziologen (Hrsg.), *Alltagswissen, Interaktion und Gesellschaftliche Wirklichkeit* (S. 54–79). Wiesbaden: VS Verlag für Sozialwissenschaften.
Yin, R. K. (2018). *Case study research and applications. Design and methods*. Los Angeles: Sage.
Zucker, L. G. (1977). The Role of Institutionalization in Cultural Persistence. *American Sociological Review, 42*(5), 726–743.

Abbildungsverzeichnis

7 Abiturprüfungspraxis und Abituraufsatz 1882 bis 1972

Michael Kämper-van den Boogaart & Sabine Reh

Zusammenfassung

Im Rahmen eines von der Leibniz-Gemeinschaft geförderten Forschungsprojekts zur Geschichte des deutschen Abituraufsatzes zwischen 1872 und 1972, dem ein umfangreiches Aufsatzkorpus zugrunde lag, wurden verschiedene Determinanten erfasst, die auf Praktiken der Aufgaben- und Texterstellung wie der Korrektur und Bewertung einwirkten. Dabei handelt es sich namentlich um rechtliche Vorgaben, um politische Erwartungen und didaktische Programme, die ihrerseits Verständnisse von Reife oder von Studierfähigkeit implizieren. Im vorliegenden Beitrag wird an Schlaglichtern aus der Geschichte dieser Prüfungsform deutlich, dass sich nicht von einem linearen Prozess sprechen lässt, sondern vielfach von Effekten, die aus einem Wechselspiel interdependenter Normierungsversuche und Diskurslagen resultieren. Im Ergebnis zeigen sich bei veränderten Vorzeichen frappierend anmutende Kontinuitäten, etwa in den Klagen über eine Überforderung der Kandidaten und Kandidatinnen oder der Frage, inwieweit fachliches und lernbares Wissen den Horizont des schriftlich Darzubietenden abstecken sollte. Zu vermuten ist mithin, dass nicht wenige der historischen Diskurskonstellationen die Beobachter:innen der gegenwärtigen Auseinandersetzungen über diesen Prüfungsteil vertraut anmuten werden.

1 Einleitung

Die Geschichte der deutschen Abiturprüfung prägen von Beginn an – so könnte man naiv sagen – ein erstaunlich lockerer Bezug zwischen einem der zentralen Prüfungsbestandteile, nämlich dem deutschen Abituraufsatz, und einem disziplinären Fachwissen sowie unkonkrete Bestimmungen zu Prüfungsinhalten und Bewertungsmaßstäben für eben jenen Aufsatz. Vor dem Hintergrund gegenwärtiger Debatten um die Vereinheitlichung von Prüfungsbedingungen im Abitur einerseits und der immer wieder zu hörenden Klage über fehlschlagende Versuche der Steuerung des Unterrichts- und Prüfungssystems dürfte es daher interessant sein zu untersuchen, wie genau und in welchen kontextuierenden Konstellationen sich das Verhältnis zwischen Abituraufsatz, Curriculum und Unterrichtspraxis zueinander im Laufe der Zeit entwickelte, und zu fragen, ob der Abituraufsatz möglicherweise aufgrund dieses besonderen Verhältnisses in höherem Maße der Herkunftsprivilegierung in einem meritokratisch legitimierten schulischen Prüfungsverfahren diente.

Während die bislang vorliegenden historischen Studien zur Geschichte des deutschen Abituraufsatzes zumeist die Themenstellungen der Aufsätze unter ideologiegeschichtlichen Aspekten in den Fokus nahmen oder zu lokalgeschichtlichen Illustrationen nutzten[1], rückt ein solches Interesse Praktiken des Schreibens und Beurteilens von Abituraufsätzen und den Modus des Abituraufsatzes als bedeutenden Bestandteil der Reifeprüfung beziehungsweise als Indikator von Studierfähigkeit in den Mittelpunkt. Wir haben daher in einem Forschungsprojekt[2] das Abitur und den deutschen Abituraufsatz als Prüfungspraxis in Preußen (beziehungsweise ehemals preußischen Territorien), in Bayern, Baden und Württemberg über knapp 100 Jahre rekonstruiert und Abituraufsätze aus dieser Zeit einer genauen bildungshistorischen und fachdidaktischen Analyse unterzogen.[3]

Im Zuge unserer bildungs- und fachhistorischen Studien bestätigte sich die Annahme, dass die Einflüsse auf den deutschen Aufsatz als Prüfungsinstrument weit vielschichtiger, widersprüchlicher und in der Summe kontingenter waren, als man aus der Perspektive einer aktuellen Bildungsforschung vermuten mag, die, an der Rationalität wissenschaftlicher Leistungsmessung orientiert, eine Komplexität der Entwicklungen im Schul- und Prüfungswesen gern übersieht. Erkennbar wird diese im Folgenden an Beispielen zu skizzierende Komplexität zum einen in der Phase eines hohen Bedeutungsgewinns für den deutschen Abituraufsatz gegen Ende des 19. Jahrhunderts und zum anderen in dem sich eben nur gebrochen umsetzenden didaktischen Normenwandel in der Zeit nach 1945.

So spielten in der zweiten Hälfte des 19. Jahrhunderts politische Auseinandersetzungen – namentlich nationalpädagogische Vorbehalte gegen den Fächer- und Prüfungskanon des neuhumanistischen Gymnasiums – eine gravierende Rolle für die

1 Zum Beispiel: Lütgemeier (2008). Überblick bei Eiben-Zach (2021) sowie Löwe, Eiben-Zach und Reh (2020).

2 Das Projekt unter dem Titel „Abiturprüfungspraxis und Abituraufsatz 1882 bis 1972. Wissens(re)präsentation in einem historisch-praxeologischen Pilotprojekt“ (05/2016–12/2021) wurde von der Leibniz-Gemeinschaft gefördert und basiert auf einer Kooperation zwischen der Bibliothek für Bildungsgeschichtliche Forschung (BBF) in Berlin sowie dem Informationszentrum Bildung, beides Abteilungen des DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation, und dem Institut für deutsche Literatur der Humboldt-Universität zu Berlin.

3 Dafür wurde ein Korpus von etwa 2400 Abituraufsätzen für den Untersuchungszeitraum erstellt. Der Zeitraum reichte von der Vereinheitlichung der Lehrpläne und Prüfungsordnungen sämtlicher höherer Schulen im Jahr 1882 in Preußen bis zur Reform der Oberstufe 1972 in der Bundesrepublik Deutschland, Grundlage des Korpus ist ein im Archiv der BBF vorhandener Bestand an Prüfungsakten und Abituraufsätzen aus Berlin. Weitere Quellenbestände wurden in Bayern und Württemberg ermittelt und erfasst. Hier existierten, im Gegensatz zu Preußen, seit unterschiedlichen Zeitpunkten Formen zentraler Prüfungen. Sämtliche Abituraufsätze wurden retrodigitalisiert und ca. 350 mit Kommentaren und Beurteilungen der Lehrkräfte transkribiert und mit Metadaten versehen in einer virtuellen Forschungsumgebung (VFU) aufgenommen. Sie bündelt einerseits die Arbeitsergebnisse des laufenden Projektes und legt andererseits die Grundlage für eine Nachnutzung des Korpus durch andere Wissenschaftler:innen. Sie enthält eine Annotationsfunktion beziehungsweise Exportmöglichkeit in externe Software; zum Schutz der Personen- und Urheberrechte sind die digitalisierten Abituraufsätze und Transkripte nur über einen moderierten Zugang zugänglich. Hierfür ist ein Antrag auf Verkürzung der Schutzfristen und die Einrichtung eines Nutzungsaccounts vonseiten der BBF notwendig. Gleichzeitig werden in einer zweiten Plattform Metadaten und diejenigen Aufsätze, die keiner Schutzfrist mehr unterliegen, ohne Zugangsbeschränkung dargestellt – d. h. alle Transkripte, deren Vorlagen älter als 1870 sind, sind Public Domain und werden nicht unter eine Lizenz gestellt. Zudem ist eine Gesetzessammlung mit ca. 900 Titeln zugänglich, vgl. Löwe 2020. Abituraufsätze wurden aus einer fachdidaktisch-historischen Perspektive untersucht in: Zach & Reh (2018); Born und Eiben-Zach (2020); Reh und Eiben-Zach (2021) und in der noch unveröffentlichten Dissertation Eiben-Zach (2021). Eine bildungshistorisch-praxeologische Analyse speziell zu den Lehrerkommentaren findet sich bei Scholz (2021). Aus einer wissens- und materialgeschichtlichen Sicht widmete sich Kerrin v. Engelhardt (geb. Klinger) den Abituraufsätzen und ihren administrativen Praktiken des Abiturs, z. B. in Klinger (2018, 2021), für eine bildungshistorische Analyse mit Fokus auf das Deutsche Kaiserreich und die Weimarer Republik als Dissertationsprojekt vgl. Löwe (i. Druck), zudem eine Auswertung von Schülerbögen in Abiturakten im Kontext von Verwaltungsgeschichte bei Löwe und Töpper (i. V. 2022). Außerdem ist ein Abschlussband für das Projekt geplant: Kämper-van den Boogaart, Reh, Schindler und Scholz (2022).

Priorisierung des deutschen Aufsatzes. Gleichzeitig war der deutsche Aufsatz gerade wegen seiner zunehmenden Identifikation als Medium der Geistes- und Charakterbildung ein prominenter Gegenstand grassierender Überforderungsdebatten, die immer wieder in publizistisch verbreitete Vorschläge zur Abschaffung der Abiturprüfung kulminierten. Im Gegensatz dazu standen wiederum Stimmen aus den Universitäten, die vor einer Studierendenschwemme und vor einer zu laxen Selektionspraxis der höheren Schulen warnten. Wiederum anders argumentierten die Protagonisten einer Pädagogik, die die persönlichkeitsbildenden Erträge einer Aufsatzerziehung durch das dem Aufsatz beigemessene Gewicht in der Prüfung konterkariert sahen und die zum Beispiel auf das Unwesen kommerzieller Aufsatzfabriken[4] hinweisen konnten. In bildungstheoretischer Perspektive nahm der Aufsatz zudem einen paradigmatischen Status für den stets kontrovers diskutierten Zusammenhang formaler und materialer Bildung ein. Die Abiturverordnungen reagierten auf diese Gemengelage, indem sie zu Kompromissformulierungen tendierten, die den divergenten Stimmungen Rechnung trugen. In der Konsequenz bedeutete dies, dass es sowohl die Schüler:innen als auch die korrigierenden Lehrkräfte den gesamten untersuchten Zeitraum über mit Textgenres und holistischen Leistungserwartungen zu tun hatten, die Unsicherheiten auslösten.

An den Aufgabenstellungen und über die in den Schreibprodukten und Korrekturen aufscheinenden Richtigkeitsvorstellungen lässt sich für die Jahre nach 1945 ein Wandlungsprozess nicht übersehen, der sich in gebrochener Form auch als Reaktion auf einen didaktischen Normenwandel erklären lässt. Allerdings wird man auch für diese Zeit keineswegs sagen können, dass sich die Praktiken des Abituraufsatzes normativ am *State of the Art* der Schreib- oder Deutschdidaktik ausrichteten; vielmehr muss umgekehrt konstatiert werden, dass sich die didaktische Publizistik in den 1970er-Jahren am Genre des Prüfungsaufsatzes fast ostentativ desinteressiert zeigt.[5] Ein Beispiel dafür, dass veränderte Positionen in der Fachdidaktik sich nur bedingt in Praktiken schulischen Aufsatzschreibens niederschlagen, stellt die Persistenz des viel gescholtenen dialektischen Besinnungsaufsatzes dar. Er ist im Unterricht allen Vorbehalten zum Trotz als Pro-und-Contra-Genre bis heute wohl deswegen bedeutsam geblieben[6], weil seine Zuspitzung auf eine Entscheidungsfrage die Orientierung an einem Schema erlaubt. Diesem traut man einiges für die Entwicklung des Argumentationsvermögens zu und es scheint mit der Erwartung auf eine finale Synthese für eine normativ begrüßte Haltung der Ausgewogenheit einzustehen – beides Attribute, die kurrenten Vorstellungen einer geistigen Reife zuträglich sind.

Auch die Kritik am Schematismus der Textsorte, die das Einschnürende der Form moniert und größere Spielräume für die Entfaltung von Selbstständigkeit for-

4 Eine auf Rechercheaktivitäten basierende Dokumentation lieferte der Herausgeber der Lehrerzeitschrift Gymnasium: Meyer (1899, 1905a, 1905b, 1905c). Über entsprechende Betrugsfälle informiert Wettberg (1914). Der Beitrag reagiert auf eine vorherige Warnung im Philologen-Blatt von Baar (1913). Charakteristisch hier der Rat an die Lehrkräfte: „Während der Klassenarbeit achte man auf die Hände und Augen der Schüler und merke sich solche, die schon bald einen gewissen Ort aufsuchen wollen, – um sich ihr Material zurechtzulegen. Hat ein Aufsatz das Thema verfehlt oder enthält er eine ganz andere Disposition oder einen andern Inhalt, als man nach der Besprechung und nach der Individualität des Schülers erwarten mußte, so lasse man sich den Entwurf vorlegen“ (ebd., S. 343).

5 Dies gilt insbesondere unter den Vorzeichen von Kommunikationsorientierung und Projektorientierung.

6 Vgl. Feilke (2017) sowie die weiteren Hinweise in Kämper-van den Boogaart (2022a).

dert, reflektiert einen gewissen Normenwandel. Zwischen den Befürwortern eines freien und denen eines eher gebundenen Schreibens zeigt sich aber nicht nur ein Disput, der um divergente Vorstellungen aufsatzpädagogischer Ziele und Menschenbilder kreist, wie sie etwa mit dem Akzent auf eine Sprachzucht bis in die 1950er-Jahre einerseits[7] und mit einer eher reformpädagogischen Wertschätzung freier Sprachentfaltung andererseits als Antagonismus beschrieben werden können. Und von Bedeutung ist schon seit der Zeit der Weimarer Republik immer wieder die Frage der Bewertung, etwa die Streitfrage, inwiefern vom Aufsatz mehr Kenntnisse erwartet werden dürfen, als im Unterricht erworben werden konnten, und vor allem der Wunsch nach objektiven Beurteilungskriterien und damit – vermeintlich oder tatsächlich – verbundenen gerechten Urteilen: „[...] daß es bei der Beurteilung von Prüfungsaufsätzen oft ungerecht zugeht, obgleich die Beurteiler selbst, bona fide natürlich, von Ungerechtigkeit nichts wissen wollen. Der einzige Trost, um den man hier nicht verlegen zu sein braucht, scheint der zu sein, daß diese Tatsache allgemein bekannt ist“ (Bobertag, 1934, S. 132).[8] Nicht von einem didaktischen und pädagogischen Normenwandel zu isolieren ist zudem die relativ spät, erst Ende der 1950er- bzw. Anfang der 1960er-Jahre einsetzende Karriere des Interpretationsaufsatzes, der sich zudem der germanistischen Verpflichtung zur werkimmanenten Interpretation und der Abwertung einer geistesgeschichtlichen Literaturgeschichte verdankt (Reh & Eiben-Zach, 2021; Eiben-Zach, 2021). Was hier als entschlossene Absage gegen die Verführung zum Baukastenpathos der älteren literarischen Aufsätze ins Spiel gebracht wird, zeitigt dann, wie die Bewertungen erweisen, ein neues Unbehagen am alten Reproduktionsverdikt. Konnte zuvor mit besten Gründen vermutet werden, dass die gravitätische Betonung des Menschheitswerts der klassischen Werke ohne Kenntnis derselben, aber unter Zuhilfenahme von Lektürehilfen und mit Nutzung einschlägig distribuierter Sprachformeln auf den Weg gebracht wurde, galten die Monita gegenüber den Interpretationsaufsätzen dem Faktum, dass nun die Inhalte der zu interpretierenden Texte paraphrasiert wurden. Für die Autorinnen und Autoren der Abituraufsätze zeigt sich ein vergleichbares Dilemma: Einerseits galt es, die soziale Erwünschtheit als sichere Bank im Auge zu behalten, andererseits musste gezeigt werden, dass die jeweils Schreibenden persönlich vom „Erlebnis der Dichtung“ durchdrungen waren und demgemäß eine authentisch anmutende Sprache fanden (Eiben-Zach, 2021).

Neben diesen Herausforderungen eines je richtigen Schreibens und Bewertens wirkte auch in der Phase nach 1970 Politisches auf die Praktiken des Abituraufsatzes ein. Mit den späten 1950er-Jahren hatte sich ein Fachlichkeitsdiskurs entwickelt, der konträr zu den stofflichen Expansionen stand, wie sie insbesondere die Deutschkunde zwischen den Weltkriegen verantwortet und den Unterricht und seine Produkte auf die umfassende Frage nach dem deutschen Wesen in allen Dingen ausge-

7 Vgl. Arbeitsgemeinschaft Deutsche Höhere Schule (1958) sowie Kämper-van den Boogaart (2022a).

8 Otto Bobertag, einer der führenden Theoretiker und Praktiker in der Testentwicklung und Teilnehmer in den Debatten um die Auslese-Problematik in der Weimarer Republik, hat schon in der Weimarer Republik Untersuchungen zur Aufsatzbeurteilung durch Lehrkräfte gemacht (Bobertag, 1934). In diesem Aufsatz findet sich in den wörtlich angeführten Lehrerurteilen immer wieder auch der negative Hinweis auf „Angelesenes“, „Hochklingendes“ und die „Phrasenhaftigkeit“ (ebd., S. 132).

richtet hatte. Mit der Besinnung auf einen an der (Mutter-)Sprache ausgerichteten Deutschunterricht versuchten Didaktiker:innen und Lehrerbildner:innen wie Erika Essen den Deutschunterricht für die fällige Oberstufenreform zu profilieren. Mit besagter Oberstufenreform am Beginn der 1970er-Jahre und der hier zunächst freizügigen Umsetzung durch die Bundesländer koinzidierte eine intendierte Expansion der Besucher:innen höherer Schulen. Das führte unter den gegebenen politischen Umständen zu einer empfindlichen Verknappung der Studienplätze und dank der föderalen Unterschiede hinsichtlich der Organisation der Oberstufe und der Abiturprüfung zu haarigen Gerechtigkeitsdebatten, in die sich dann erstmals das Bundesverfassungsgericht einschaltete. Das Resultat: Einerseits gewann die Notengebung mit Blick auf die hochschulischen Zulassungsbeschränkungen (NC) eminent an Bedeutung, andererseits war die Kultusministerkonferenz (KMK) gefordert, den Weg zum Abitur zu vereinheitlichen, um die durch föderale Strukturen bedingten Vor- und Nachteile für Landeskinder zu minimieren (Gass-Bolm, 2005).

Wir werden im Folgenden, um diese widersprüchlichen und vielschichtigen Einflüsse veranschaulichen zu können, ausgewählte Einblicke in Debatten um den Abituraufsatz und Vorstellungen über Reife und – wie es dann später heißt – Studierfähigkeit in verschiedenen Etappen der Geschichte des Abituraufsatzes geben und einige ihrer wechselhaften Trends darstellen. Ausgehend von Prüfungsvorschriften und anhand verschiedener Veröffentlichungen von „Schulmännern", also von pädagogisch ausgebildeten Fachleuten aus der Schulverwaltung, zeigen wir die Selektionserwartungen auf, wie sie sich schon im Laufe des 19. Jahrhunderts herausbildeten. Im Mittelpunkt dieses ersten Teils steht ein publizistischer Vorfall aus dem Jahr 1930, der die Erwartungen der Universität gegenüber den höheren Schulen und dem Abitur betrifft. Hier kann deutlich werden, als wie strittig sich die Markierung von Reife durch das Abitur im Grenzbereich von Wissenschafts- und Erziehungssystem erwies (1). In einem nächsten Abschnitt wird dann problemorientiert thematisiert, inwiefern ausgerechnet der deutsche Aufsatz zur Markierung der gewünschten Reife dienen soll (2), bevor in einem dritten Teil die Normierung der schriftlichen Abiturprüfung im Kontext der BVG-Rechtsprechung zu Zulassungsbeschränkungen in Augenschein genommen wird (3) und abschließend die komplexe historische Situiertheit des deutschen Abituraufsatzes als Rahmenbedingung gegenwärtiger Reformversuche markiert wird (4).

2 Selektionserwartungen an das Abitur

In Deutschland wird der Impuls zu den staatlichen Versuchen, im Übergang von der höheren Schule an die Universität eine mehr oder weniger einheitliche Prüfung zu implementieren, aus der Perspektive der aufnehmenden Institutionen, der Universitäten, gesetzt. Geschaffen wird damit eine Lösung, die sich unterscheidet von der in anderen Ländern, wie etwa zur gleichen Zeit in Frankreich, wo das Baccalauréat als Aufnahmeprüfung verankert wird, die wesentlich in der Verantwortung der universi-

tären Fakultäten liegt – das eröffnet, wie im Vergleich zwischen Deutschland und Frankreich bis heute zu erkennen ist, dann jeweils unterschiedliche Entwicklungspfade (Marchand, 2010).

2.1 Das Abitur als schulische Abschlussprüfung – zum Verhältnis von Gymnasium und Universität

> „Es ist bisher vielfältig bemerkt worden, daß so viele zum Studieren bestimmte Jünglinge ohne gründliche Vorbereitung unreif und unwissend zur Universität eilen, wodurch selbige nicht nur sich selbst schaden, [...] sondern auch zu gleich verursachen, daß viele Aemter, zu denen gründliche Kenntnisse erforderlich sind, wo nicht mit unwissenden doch mit seichten und unzweckmäßigen Subjecten besetzt werden. [...] Es ist daher beschlossen worden, daß künftig alle von öffentlichen Schulen zur Universität abgehende Jünglinge [...] öffentlich geprüft werden, und nachher ein detaillirtes Zeugnis über ihre bey der Prüfung befundene Reife oder Unreife zur Universität erhalten sollen."[9]

> „Der Zweck, einem nicht genugsam vorbereiteten Besuch der Universität bei der studirenden Jugend vorzubeugen, hat die Prüfungen der Schüler vor ihrer Entlassung zur Universität herbeigeführt, welche durch das Circulär vom 23. December 1788 angeordnet sind. Die seitdem darüber gesammelten Erfahrungen [...] machen neue und vollständigere Bestimmungen über diese Prüfungen nothwendig, welche durch gegenwärtige Instruction gegeben werden. [...] § 7. Um nun den Besitz oder Mangel der zum fruchtbaren Besuch der Universität nöthigen Ausbildung zu erforschen, wird die Prüfung angestellt, wobei theils die Kenntnisse selbst dargelegt, theils Uebungen vorgenommen werden müssen, woraus sich auf die erworbenen Fertigkeiten schließen läßt."[10]

> „Jeder Schüler, welcher sich einem Berufe widmen will, für den ein drei- oder vierjähriges Universitäts-Studium vorgeschriben, muß sich vor seinem Abgange zur Universität [...] einer Maturitäts-Prüfung unterwerfen. [...] Der Zweck dieser Prüfung ist, auszumitteln, ob der Abiturient den Grad der Schulbildung erlangt hat, welcher erforderlich ist, um sich mit Nutzen und Erfolg dem Studium eines besonderen wissenschaftl. Faches widmen zu können."[11]

Ersichtlich wird an diesen kleinen Auszügen aus preußischen Abitur-Reglements des 18. und 19. Jahrhunderts, dass der preußische Staat für seine modernisierte Administration Personen verlangt, die ein anspruchsvolles Universitätsstudium absolviert haben. Die Universitäten ihrerseits verlangen für diese Aufgabe Studienanfänger, die den universitär gesetzten Ansprüchen eines engagierten Studiums genügen. So hatte der Kanzler der Universität Halle, von Hoffmann, festgehalten:

> „Die Erfahrung lehrt [...], daß sich unter den jungen Leuten welche die Universitäten beziehn, beständig eine nicht geringe Anzahl von solchen Subjecten befindet, die nicht allein in den beyden sogenannten gelehrten Sprachen, sondern auch in den übrigen noch wichtigern Vorkenntnissen, die sie von den Schulen mitbringen sollten, so unwissend

9 Reglement für die Prüfung an den Gelehrten Schulen vom 23.12.1788, http://resolver.staatsbibliothek-berlin.de/SBB0002BEE700000000

10 Edict wegen Prüfung der zu den Universitäten übergehenden Schüler vom 12.10.1812/Instruction vom 26.06.1812, https://ghdi.ghi-dc.org/docpage.cfm?docpage_id=4220&language=german

11 Reglement vom 04.06.1834 für die Prüfung der zu den Universitäten übergehenden Schüler, https://play.google.com/store/books/details?id=xuoDDj3gp8cC&rdid=book-xuoDDj3gp8cC&rdot=1

> sind, daß ihre Unwissenheit bald Mitleiden, und bald Widerwillen erzeugen muß: und es ist nicht unerhört, daß zum Beyspiel einer, der sich dem Cameralwissen widmen will, seine Studien mit Dogmatik anfängt. Es ist natürlich, daß diese Leute, die keinen Grund gelegt haben, auf keinen Grund bauen können, und daß sie von der Universität eben so unwissend, ja mit verworrnern Begriffen wieder weggehen, als sie dahin gekommen sind, und am Ende sich und dem Staate lästig, und, wenn sie eine Versorgung finden, letzterem sogar nachteilig werden müssen." (Zit. n. Schwartz, 1910, S. 67 f.)

Zwischen 1788 und 1834 wird dieser Regelungsprozess in Preußen mehr oder weniger abgeschlossen, was die Form und den Status der Allgemeinen Hochschulzugangsberechtigung, wie sie uns bis heute vertraut ist, anbelangt. In diesem Zeitraum wird der erfolgreiche Abschluss der Prüfung für die Immatrikulation an einer Universität verbindlich, setzt der Staat seine Zugangsregelungen gegenüber väterlichen Ansprüchen durch, werden Sonderregelungen für privat unterrichtete Kandidaten abgeschliffen und es bildet sich ein Kanon schulischer Prüfungsgegenstände heraus, der Konsequenzen für die curriculare Organisation des gymnasialen Unterrichts hat. Das Ergebnis von 1834 bedeutet aber natürlich nicht ein Ende der Regelungsdynamik: So gehen von der Diversifikation im höheren Schulwesen, von Überforderungsdebatten im 19. und 20. Jahrhundert sowie von einer Neugewichtung ausdifferenzierter Schulfächer ebenso Veränderungsimpulse aus wie von den veränderten politischen und pädagogischen Konstellationen, etwa einer seit dem Ende des 19. Jahrhunderts zunehmenden Schul- und Prüfungskritik, einer Kritik am „Berechtigungswesen" und darin auch einer Kritik am Abitur, das eben nicht die Persönlichkeit, die Reife des Abiturienten prüfe, sondern ob er „voll" sei von dem, was in seinen Kopf getrichtert wurde, so Schwartz (1910, S. 66).[12]

Veränderungen zeigen sich deutlich, wenn man die Abiturordnung von 1926 mit ihren Vorgängern vergleicht. Konstant in dieser Dynamik bleibt zwar das zentrale Verständnis, dass die Abiturprüfung jene Reife zu examinieren hat, die als Fähigkeit zu einem erfolgreichen Studium als vonnöten erachtet wird.[13] Dies gilt auch dann, wenn die dominante Rolle des Staates für den Arbeitsmarkt von Universitätsabsolventinnen und -absolventen an Bedeutung einbüßt und die Universitäten in Gestalt der Professorenschaft mehr oder weniger eigenständig für ihre Wissenschaftskulturen diagnostizieren, welche Voraussetzungen aufseiten der jungen Studierenden für eine aussichtsreiche Enkulturation gegeben sein müssen. In diesem Verständnis wird das

12 Auf den Versuch, eine Theorie der schulischen Prüfung zu bieten, geht Münch (1932) ein. Vgl. zu diesem Zusammenhang auch Reh, Bühler, Hofmann und Moser (2021). Vgl. zu den Hintergründen der skizzierten Zusammenhänge je umfassender insbesondere die folgenden Arbeiten: Paulsen (1902); Flitner (1959); Blättner (1960); Herrlitz (1973); Kraul (1984); Wolter (1989); Jeismann (1996); Gass-Bolm (2005); Bölling (2010). Eine kompakte Darstellung mit einem Akzent auf der Diversifizierung der zum Abitur führenden Schulen gibt Wolter (2016).

13 Eine ziemlich klare und kompakte Funktionsbeschreibung der Abiturprüfung findet sich 1921 in dem prüfungskritischen Beitrag „Die Reifeprüfung" von Felix Behrend: „Die Reifeprüfung soll die Hochschulen vor Überflutung mit ungeeigneten Studenten schützen, also eine gewisse Schülerauslese verbürgen; sie soll eine Kontrolle der Zielleistungen der Schule, also eine Prüfung von Anstalt und Lehrern sein; sie soll die Gleichmäßigkeit der Leistungen der einzelnen Schulen herstellen oder aufrechterhalten; sie soll endlich Eltern und Schüler gegen unrechte Behandlung durch einzelne Lehrer oder zu hohe Anforderungen ganzer Kollegien schützen" (Behrend, 1921, S. 441).

Gymnasium klar zu einer das Studium vorbereitenden Schule[14] und es liegt seitens der Universitäten nahe, Probleme ihres Studienbetriebs auf mangelnde Leistungen der höheren Schulen zu projizieren. Spannungsreich ist diese Konstellation bis heute, da es de facto durch die Lösung, eine schulische Abschlussprüfung mit dem Zugangsrecht zu einem beliebigen Universitätsstudium zu verschmelzen und mithin die Alternative einer universitären Aufnahmeprüfung (mit Ausnahmen) auszuschließen, zu einer in spezifischer Weise konfliktträchtigen Rollenverteilung der verantwortlichen Akteure kommt. Die Lehrer:innen der höheren Schulen entscheiden als für die Abiturprüfung Verantwortliche über die formelle Eignung ihrer Schüler:innen, ein Studium aufzunehmen, und agieren so als Expertinnen und Experten in Sachen einer Reife für wissenschaftliches Arbeiten, während ebendiese Expertise begründet von den an den Hochschulen Lehrenden und für das wissenschaftliche Studium Verantwortlichen reklamiert wird. Solange der Zusammenhang zwischen Universitäts- und Gymnasialkulturen eng ist[15], sorgt dieser Widerspruch nur für moderate Konflikte. Dies änderte sich aber aus naheliegenden Gründen, sobald sich das höhere Schulwesen und das Hochschulsystem weiter ausdifferenzierten und sich Gymnasiallehrkräfte über den Kreis der erklärten Reformer hinaus immer weniger als Dienstleistende für die Universitäten, sondern in eigener pädagogischer Rolle sahen (Kraul, 2021).

Betrachtet man die an die Schulen und die Schulpolitik gerichteten Reklamationen der Hochschulseite, zeigt sich zum einen, dass beanstandet wird, dass die höheren Schulen mehr berechtigende Abschlusszeugnisse ausstellen, als für die Universitäten verkraftbar sind.[16] Zum anderen, und oft genug mit dem ersten Monitum einhergehend, wird der qualitative Vorbehalt nachhaltig, dass die mit dem erfolgreich absolvierten Abitur zertifizierten Qualifikationen universitär gebotene Eingangsstandards unterschreiten. Dies ist mit Blick auf Praktiken der Abiturprüfung, auf Korrektur und Bewertung der interessantere Punkt, da solche Standards in den Abiturregle-

14 So auch Anhalts Einschätzung: „Die Aussage von O. Jäger, 1894 auf der zweiten Versammlung deutscher Historiker vorgetragen, um den ‚Hauptzweck' des Gymnasiums zu bestimmen, ist daher auch heute noch zustimmungsfähig: ‚Das Gymnasium soll studieren lehren – soll seinen Schülern den Weg zeigen und die Mittel reichen, Wahrheit selbstthätig zu finden' […]. Das Gymnasium als Ort der ‚höheren' Schulbildung unterscheidet sich in diesem Punkt von den übrigen Schultypen: Es ist ausgerichtet auf den ‚Hauptzweck', Studierfähigkeit zu vermitteln. Gemeint ist damit, dass Absolventen des Gymnasiums in der Lage sein sollten, an einer Hochschule mit Aussicht auf Erfolg zu studieren" (Anhalt, 2014, S. 134 f.). Engagierter Widerspruch zur Instrumentalisierung der Gymnasien als Vorbereitungsanstalten eines Universitätsstudiums zeigt sich allerdings bereits 1851 in der lesenswerten Abhandlung des Parchimer Oberlehrers Adolph Steffenhagen (1851). Steffenhagen differenziert mit spitzer Feder zwischen Fach- und Humanitätsanstalten und akzeptiert einen engen Vorbereitungscharakter nur für Fachanstalten, während das Gymnasium eben keine Vorbereitungsschule auf einen besonderen „Lebensberuf" darstelle (ebd., S. 124) und postuliert: „So zweckmäßig wir also es finden, daß die Universität und andere Fachanstalten gewisse Kenntnisse und eine gewisse Bildungsstufe bei den zu rezipirenden Zöglingen voraussetzen, so folgt aus dieser Einrichtung derselben ganz und gar keine Verbindlichkeit für die Humanitätsanstalten, sich wer weiß welchen Anforderungen der Fachanstalten zu accomodiren, sich zu bloßen Vorbereitungsanstalten derselben herzugeben. Wenn sie das thun, so vergessen sie, daß sie eine selbständige Aufgabe zu lösen, daß sie an der Verwirklichung des rein menschlichen Bildungsideals zu arbeiten haben" (ebd., S. 125). Dass sich das Verständnis des Gymnasiums als Vorbereitungsschule seit dem 19. Jahrhundert verändert habe, betont ein Beiträger des Philologen-Blattes ausgerechnet 1930, wenn er gestrige Verständnisse geißelt: „Denn immer noch gibt es vielfach laienhafte Gemüter, welche in dem ‚Abitur' nicht das Siegel für ein Weggehen, sondern den Ausweis für ein Eintreten erblicken, denen das Reifezeugnis nicht die Urkunde für eine erreichte Höhe, sondern der Zulaßschein für einen neu zu erklimmenden Pfad ist. […] die Tage sind längst dahin, da nur die Universität die ‚Reife' ihrer jungen Besucher voraussetzt […]". (Schott, 1930, S. 166.).

15 Vgl. hierzu etwa die zum Teil sehr konkreten Ratschläge in Meyer (1899).

16 Zur „neurotische[n] Überfüllungsdiskussion der 1880/90er-Jahre" vgl. etwa Kraul (1984).

ments oft nur abstrakt, vage oder tautologisch expliziert werden können und da sich aus der Perspektive der Schulleute die Frage stellt, was in der vieljährigen Vorbereitung auf das Abitur durch Unterricht machbar ist. Bereits eine Sichtung der staatlichen Ordnungen gibt zudem zu erkennen, dass unter dem Begriff der Reife sowohl das Verfügen über Kenntnisse im Sinne eines deklarativen und prozeduralen Wissens subsumiert wird als auch unterschiedliche Persönlichkeitsmerkmale gefasst werden. Welche Auswirkungen die damit verbundenen Uneindeutigkeiten des Reifeurteils im Kontext moderner Transitionsmodelle[17] haben, sei an einem historischen Beispiel universitärer Kritik einer zu nachlässigen Selektionspraxis der Gymnasien illustriert, bevor unter diesen Vorzeichen die Hybridität des deutschen Abituraufsatzes in den Blick geraten soll.

2.2 1930 im Philologen-Blatt: Die überfüllte Universität und das Abitur

Am 30. April 1930 erscheint im *Deutschen Philologen-Blatt*, dem zentralen Organ der Schulphilologen, ein umfangreicherer Beitrag des Berliner Anglistikprofessors Wilhelm Dibelius[18] – während sich ansonsten meist Schulmänner, also Verwaltungsbeamte, aber auch Lehrkräfte äußerten. Dass er seine Ausführungen „Die Überfüllung der Universität" betitelt, nimmt unfreiwillig vorweg, was am 25.04.1933 im NS-„Gesetz gegen die Überfüllung deutscher Schulen und Hochschulen" als Tarntitel zu einer Verdrängung jüdischer Schüler:innen und Studierender firmiert. Die Ausrichtung des Beitrags veranlasst die Schriftleitung der Zeitschrift zu dem einleitend distanzierenden Hinweis, dass Dibelius' Schilderungen und Vorschläge als Diskussionsanregung zu begreifen seien, denen die Verantwortlichen ausdrücklich nicht vollumfänglich beipflichten könnten. Insbesondere wird Wert darauf gelegt, dass die „soziale Funktion der höheren Schule" „sich grundsätzlich dadurch geändert hat, daß sie nicht mehr allein für die Hochschule vorbereitet" (Dibelius 1930, S. 265). Diese Anmerkung ist verständlich, da der Universitätsanglist tatsächlich immer noch ganz aus der Perspektive der Universität argumentiert. Obgleich sein Beitrag auf eine empirische Absicherung verzichtet, spricht zunächst viel für seine These, dass der aktuelle

17 Gemeint sind damit jene Grenzen und Übergänge zwischen den Institutionen höherer Bildung beziehungsweise wissenschaftlichen Anstalten, die im Kontext der Berliner Universitätsgründung auf Wilhelm von Humboldt und seine Mitstreiter zurückgeführt werden. Die Tendenz zum Konflikt zeigt sich auch in dessen mit hundertjähriger Verspätung vielzitierten Darlegungen *Über die innere und äußere Organisation der höheren wissenschaftlichen Anstalten in Berlin*. Humboldt spricht hier einerseits davon, dass es die Pflicht des Staates sei, „seine Schulen so anzuordnen, dass sie den höheren wissenschaftlichen Anstalten gehörig in die Hände arbeiten" (Humboldt, 2010, S. 260). Dies will er andererseits aber so verstanden wissen, dass ebendiese Schulen nicht dazu berufen seien, „schon den Unterricht der Universitäten zu anticipieren" (ebd.). Im Ergebnis sollten die Schulen Absolventen heranbilden, die „physisch, sittlich und intellectuell der Freiheit und Selbsthätigkeit über lassen werden" könnten. Diese strebten dann nicht „zu Müssiggang oder zum praktischen Leben", sondern trügen eine „Sehnsucht" in sich, „sich zur Wissenschaft zu erheben" (ebd., S. 261). Dass der „Weg dahin zu gelangen, […] einfach und sicher" sei, wie Humboldt optimistisch erklärte (ebd.), dürfte sich vielen Akteuren nicht so ganz erschlossen haben. Für Spranger (1910, S. 251 f.) hingegen bot die sich an Humboldts Überlegungen anschließende Diskussion über die für formale Bildung zuträglichen Unterrichtstoffe die Möglichkeit, die Vorstellung einer organischen Denkkraft zu lancieren – ein folgenreicher Gedanke, der hier nicht weiter verfolgt werden kann.. Im Kern ähnlich wie Spranger argumentierte bereits 1829 ein Beiträger unter dem Kürzel „teuber" in der Allgemeinen Schulzeitung „Ueber Abiturienten-Prüfungen": „Hat das Gymnasium auch zunächst die Aufgabe der formellen Bildung der studirenden Jugend zu lösen, d. h. hat dasselbe es dahin zu bringen, dass der Jüngling durch Anregung und Entwickelung seiner geistigen Kräfte befähigt werde, auf der Akademie jedes beliebige Fachstudium mit günstigem und glücklichem Erfolge zu betreiben, so lässt sich doch jene formelle Bildung ohne materielle Grundlage gar nicht denken. Es muss demnach durchaus ein gewisser Grad materieller Kenntnisse schon auf der Schule erworben werden […]" (ebd., S. 1148).

18 Vgl. zu diesem Beitrag Bölling (2010).

„Andrang" zur Universität mit der Misere des Beschäftigungsmarktes zu tun hat: „Die Universität ist die große Wartehalle der Unentschiedenen geworden. Es drängt alles zur Universität – auch die Frauen" (ebd., S. 265). Zwar geben die Daten des Statistischen Reichsamtes für die Jahre 1921 bis 1928 nur zu erkennen, dass man sich 1928 nach einer deutlichen Delle Mitte der 1920er-Jahre wieder dem Immatrikulationsniveau von 1921 annähert. Zählte man 1919 53.108 Studierende an preußischen Universitäten, waren es 1930 57.130 (Titze, 1987). [19] Aber zutreffend ist trotz aller Übertreibung, dass mit der Weltwirtschaftskrise 1929 eine markante temporäre Erhöhung der Studentenzahlen einhergeht; das dürfte auch im Sinne der Wartehallenmetapher mit einer längeren Verweildauer zusammenhängen (Müller-Benedict, 2016).[20] In einer Statistik, die das *Deutsche Philologen-Blatt* im Juni 1930 abdruckt, werden die Semesterkohorten der Philologiestudierenden verglichen. Angesichts des Umstands, dass sich 1928/29 6.246 Studierende im 1. und 2. Semester befinden, während 2.339 für die Semester 9 und 10 registriert sind, lautet die Schlussfolgerung – analog zu Dibelius –, dass man es mit einem außerordentlichen „Zudrang zum Studium der Philologie in den letzten Jahren" (Simon, 1930, S. 355) zu tun habe.[21]

An dieser Stelle entscheidender als die Überfüllungsklagen sind indes die Diagnosen zu einer mangelnden Studierfähigkeit, wie es dann seit den 1960er-Jahren im Kontext eines wissenspropädeutisch ausgelegten Oberstufenunterrichts heißen wird[22]: Hier findet sich nach Hinweisen zur materiellen Situation, die vielen ein umfängliches Studium nicht erlaube, zunächst eine Reminiszenz: Die wenigen Menschen, die die alte Universität belebt hätten, seien dadurch geprägt gewesen, dass sie „Interesse haben für intellektuelle Probleme" (Dibelius, 1930, 266). Unmissverständlich klar macht Dibelius, dass er von der Gründungsidee dieser alten Universität nicht zugunsten eines Verständnisses im Sinne einer professionsbezogenen „Fachschule" abrücken will: „Die Universität will vielmehr nicht ein Wissen vermitteln, sondern

19 Die Abiturientenzahlen in Preußen zeigen allerdings einen deutlichen Anstieg: 1920: 11.746; 1930: 17.440; Daten für die Abiturquote liegen seit 1925 vor: Von dort bis 1930 entwickelt sich die Quote bei den Abiturienten von 2,78 auf 4,59, bei den Abiturientinnen von 0,28 auf 1,19. (Titze 1987, S. 171), schließlich die Studienanfänger: 5.179 (1919) gegenüber 8.052 (1930). 1925 betrug die Zahl der Neueingeschriebenen aber lediglich 4.807 (ebd., S. 188).

20 Mit Blick auf die betreffenden Jahre spricht Titze (1981) von einer vierten Überfüllungswelle akademischer Karrieren, die er u. a. so charakterisiert: „In der Zwangslage eines schwächer wachsenden Stellenmarkts für Akademiker zeigte sich das immer wiederkehrende Überfüllungsproblem in den 20er und 30er Jahren unseres Jahrhunderts in seiner bisher schärfsten Ausprägung in der deutschen Geschichte. Im Vorfeld der nationalsozialistischen Gewaltlösungen stieg das Reservoir der ‚Überzähligen', die von der ‚akademischen Berufsnot' (Schairer 1932) betroffen waren, weit in die Zehntausende. Als historisch neuartiger Faktor trat in dieser Krise das Frauenstudium hinzu, das in der zweiten Hälfte der 20er Jahre eine starke Ausweitung erfahren hatte" (Titze, 1981, S. 189).

21 Dabei gilt die Einschätzung, dass das Studium in der Regel im 10. und 11. Semester beendet werde, von Schwundquoten ist keine Rede (vgl. Simon, 1930, S. 355). Die Hochschulstatistik bei Titze weist für die Sprach- und Kulturwissenschaften in Preußen folgende Zahlen aus: 1919/20: 8.885; 1930: 12.671 (vgl. Titze, 1981, S. 97).

22 Attestierte Reife in der Weimarer Republik war zunächst als Reife der Persönlichkeit verstanden, die in einem „Bildungsgang", den die Abiturienten schriftlich darzulegen hatten, aufgezeigt werden musste (vgl. Stelmaszyk, 2002; Dausien & Kluchert, 2016; Klinger, 2019). In den Formulierungen Dibelius' wird eher abgehoben auf das, was später als Studierfähigkeit im Sinne einer generellen Leistungsbereitschaft konzipiert wird: „Als zentrales Merkmal gilt die diagnostizierte Leistung, d. h. die Leistungsbereitschaft und -fähigkeit einer Person im Kontext von Erwartungen im Bildungssystem. [...] Indem Studierfähigkeit an Leistung gebunden und Leistung gegen Beliebigkeit sowie Willkür in der Zuteilung von Chancen an der Teilhabe am gesellschaftlichen Verkehr gestellt wurde, war die Verbindung mit dem Wissenschaftssystem hergestellt. In diesem System wird, zumindest der Idee nach, Leistung als objektiv zurechenbares Produkt individuellen Verhaltens behandelt und konsequent von Statusattributen, Bekanntschaftsbeziehungen, Besitz, Nationalität, Geschlecht, Alter etc. freigehalten. Studierfähigkeit wurde im Sinne dieser Idee als Hochschulreife interpretiert und an die sachlichen Erfordernisse eines Hochschulstudiums geknüpft [...]." (Anhalt, 2014, S. 122 f.).

eine Geisteshaltung; sie will anleiten zum ruhigen, vorurteilslosen Durchdenken von Problemen; der Wissensstoff ist für sie nur das Material, an dem intellektuell gearbeitet werden soll" (ebd.). Im Studium ginge es demnach um ein „‚sich wundern' an den Dingen" und mithin um die Heranziehung junger Leute, „die daran gewöhnt sind, alles, was an sie herantritt, nach höheren Gesichtspunkten zu betrachten, es unabhängig von Volksmeinungen und augenblicklichen Modeströmungen ruhig zu beurteilen", von „Menschen mit intellektuellen Anlagen", die dazu disponiert seien, Probleme zuerst zu durchdenken, „ehe man daran geht, an Ethos und Willen zu appellieren" (ebd.). Diese intellektuell disponierten Studierenden diskriminiert er von denen, die er für eine berechtigte zeitgenössische Kritik an den Universitäten verantwortlich macht, die Majorität der „praktischen Naturen", die die Welt zwar nicht entbehren könne, die aber an der Universität nichts zu suchen hätten (ebd.). Mit diesen Ausführungen orientiert sich Dibelius durchaus an der Semantik der vor allem seit der Spranger-Edition 1910 zirkulierenden Konzepte Humboldts. Dies gilt in gewisser Weise auch für die auf den ersten Blick überraschende Wendung, die seine Schelte von Oberschulen und Abiturauslese nimmt.

Könnte man wegen der Betonung fehlender Dispositionen intellektueller Selbstständigkeit nämlich erwarten, dass der Schulkritiker eben auf diesem Feld dem universitätsvorbereitenden Unterricht Versäumnisse vorhält, sieht man sich getäuscht. Was stattdessen beklagt wird, sind fehlende stoffliche und sprachliche Kenntnisse der Abiturientinnen und Abiturienten, so der zu schwache „Unterbau von Tatsachen" und ein „erschreckender Mangel an Wissen", etwa auch an fremdsprachlichen Vokabeln und orthographischen Prinzipien (ebd., 267), sodass erwogen werde, an der Universität ein „Vorbereitungsjahr" einzuführen, „in dem der Student zu lernen gezwungen wird, was er auf der Schule lernen sollte" (ebd.). Diese Akzentuierung einer vorakademischen Lernanstalt scheint zwar dem Humboldtschen Denken zu widersprechen, wie wir es aus den Beschwörungen seiner vermeintlichen Bildungsideale zu kennen glauben. De facto entspricht es aber seiner Distinktion von wissenschaftlichen Bildungsanstalten, derzufolge einerseits „die Schule es nur mit fertigen und abgemachten Kenntnissen zu thun hat und lernt" (Humboldt, 2010, S. 256), während die „Vorübung des Kopfes zur reinen Wissenschaft" an die schulische Idealerfahrung formaler Bildung gebunden wird, dass das Lernen am sorgsam ausgewählten Gegenstand nicht durch „äussere Umstände, sondern durch seine innere Präzision, Harmonie und Schönheit Reiz gewinnt" (ebd., S. 261). Unabhängig von der Frage, ob diese Bestimmungen letztlich Praxisantinomien implizieren mögen, wird man im gegebenen Zusammenhang konstatieren dürfen, dass Dibelius' Schulkritik der Grundunterscheidung Humboldts nicht zuwiderläuft, wenn er moniert, dass Absolventinnen und Absolventen im Schulunterricht zu wenige Kenntnisse erworben haben. Um den Zustrom der nach dieser Auffassung ungeeigneten Studierenden abzuwehren, schlägt der Anglist eine Reihe von Maßnahmen vor, die letztlich auf eine verschärfte Selektion hinauslaufen. Dazu zählt unter anderem die Einführung eines fachspezifischen Numerus Clausus, aber ebenso eine Umgestaltung der Reifeprüfung, deren Neuord-

nung in Preußen erst Ostern 1927 in Kraft getreten war.[23] Dibelius' schroffer Befund lautet nämlich: „Als Mittel der Siebung, wie sie ursprünglich gedacht war, hat sie versagt" (ebd., S. 267).[24] Konkret schlägt Dibelius die Einführung eines Zentralabiturs auf der Ebene der Provinzialschulkollegien vor, wobei als bewertende Lehrkräfte agieren sollen, die die Prüflinge *nicht* unterrichtet haben. Hier solle man „wieder einmal den Mut haben, von jungen Leuten *Leistungen* zu verlangen" (ebd., S. 268, H. i. O.), wozu auch eine schärfer differenzierte Notenskala beitragen soll. Zwar konzediert er, dass fachliche Differenzierungen möglich sein müssten, sofern die „menschliche Gesamtleistung sich als gediegen herausstellt" (ebd.), relativiert dies aber sodann, indem er für einzelne Studienfächer spezifische Fachleistungen zur Voraussetzung erklären will. Diesen konkreten Vorschlägen folgen allgemeine Überlegungen zum Zusammenhang sozialer Eliten und Universitätsstudien, die zwar mentalitäts- oder ideologiegeschichtlich aufschlussreich sind, aber das engere Thema des Abiturs als Hochschulzugang transzendieren. Interessanter ist hier, wie seitens der philologischen Lehrkräfte in Folgenummern des „Philologen-Blatts" auf die Bestandsaufnahme und Rezeptideen des Hochschullehrers reagiert wird. Während einzelne Beiträger, so Hans Schlemmer, die Beurteilung der Schulabsolventinnen und -absolventen durch Dibelius skeptisch beurteilen und sogar den Verdacht hegen, dass solche Klagen „sehr häufig auf Unfähigkeit oder Unlust beruhen, der heutigen Art der Studenten wirklich Rechnung zu tragen" (Schlemmer, 1930, S. 449), akzeptiert das Gros der Diskutanten doch, dass Dibelius' Darstellung der Probleme der Universität einen realistischen Kern besitzt. Auch der Vorbehalt unzureichender Selektion beziehungsweise einer falschen Auslese und zu großzügiger Titelvergabe wird nicht selten eingeräumt (Weßner, 1930, S. 450 f.). Interessanterweise steht einer diskursiven Befürwortung der Leistungsselektion häufig allerdings eine andere Praxis in der Abiturprüfung gegenüber – Selektion wird an dieser Stelle der Schulkarriere offenbar nicht unbedingt gern mehr geltend gemacht (Scholz et al., 2021), dafür werden anscheinend strukturelle Fragen diskutiert, etwa über Unschärfen in der Versetzungspraxis vor der Oberstufe (ebd.).

Studiert man die Kommentare der Gymnasiallehrer, wird man nicht sagen können, dass das Verständnis, die Hochschulreife habe sich im Abitur durch das Ausmaß dargebotener Faktenkenntnisse oder eines inkorporierten Fachwissens zu zeigen, den Bildungsvorstellungen der Schulleute wirklich entspricht. Dies zeigt sich insbesondere in der Diskussion über die Kriterien der von den Abiturkommissionen zu fällenden Reifeurteile, in der immer wieder auf die Notwendigkeit verwiesen wird, die Gesamtpersönlichkeit und den ganzen Menschen mit seinen Anlagen zu berücksichtigen, was gleichzeitig Flexibilität wie psychologische Expertise und genaue Beobachtung notwendig mache. Explizit gegen die von Dibelius unterstellte Arbeitsteilung, Stoffvermittlung durch die Schule, Geistesveredelung durch die Universität, wendet

23 Ordnung der Reifeprüfung an den höheren Schulen Preußens (1926), http://goobiweb.bbf.dipf.de/viewer/image/985843438_0068/1/.

24 Zwar trifft es zu, dass bereits frühe Verordnungen zum Abitur dessen Bedeutung für die Auswahl geeigneter Studierender betonten, aber richtig ist auch, dass nach der Festlegung des Abiturs als exklusiver Studieneingangsberechtigung entscheidende Selektionsprozesse durch Schulabgänge vor dem Abitur stattfanden (vgl. Scholz et al., 2021).

sich der Pasewalker Gymnasiallehrer Hermann Knust[25], wenn er einwirft: „Man sollte doch eigentlich erwarten, daß die Verlegung des Schwerpunkts der Arbeit in der höheren Schule vom Wissensstoff auf Entwicklung selbständigen Denkens die beste Universitätsvorbereitung ist“ (Knust, 1930, S. 424). „Das geringere Tatsachenwissen der heutigen Studenten ist zuzugeben“, räumt Knust ein, um allerdings darauf hinzuweisen, dass etwa „passiver Aufnahme einer Stoffülle“ nicht mehr die alte Bedeutung zugesprochen werde (ebd., S. 423). Obgleich er Dibelius daran erinnert, dass Nietzsche, Schopenhauer und Bismarck an der Universität nicht glücklich wurden und dass wohl auch die alte Universität nicht durchweg auf ein sicheres Tatsachenwissen der Studienanfänger setzen konnte, bleibt er doch vorsichtig, was den Ertrag des schulreformerischen Akzents auf „aktiver Entfaltung eigenen Wesens“ und „stärkerer Fühlungnahme mit dem Leben und der Gegenwart“ angehe (ebd.): „Sind sie nicht nur in den Tatsachen unsicherer, sondern auch geistig unbeweglicher, unselbständiger und kraftloser, so wären die höhere Schule und alle Anstrengungen der Reform gerichtet. Dann wäre das Opfer der bis dahin für unentbehrlich geltenden Wissensstoffe umsonst gewesen; denn gerade, was sie dafür gewinnen sollte, erlangte sie in geringerem Maße als früher: Steigerung der produktiven Kraft. Aber hier ist wohl noch nicht das letzte Wort gesprochen“ (ebd., S. 424). Vergleichbar ambivalent fällt auch Knusts Sicht auf das reformierte Abitur aus. Während er dem alten Abitur vorhält, ein „Schreckgespenst auch für tüchtige Leute“ gewesen zu sein[26] und „eine Überschätzung abfragbaren Stoffs“ forciert zu haben, konzediert er, dass man es nach Dibelius nun mit „bedenkliche[r] Weichheit“ (ebd.) zu tun habe. Auf eine veränderte Didaktik geht auch Leonhard ein, wenn er das Ziel der „Urteilsbildung“ (Leonhard, 1930, S. 484) fokussiert. Dass die Jugend „von Natur“ dazu neige, „Urteile zu fällen“, bedeute nicht, dass es keiner pädagogischen Interventionen bedürfe, um die Jugend „durch Kritik fortschreitend zu befähigen, ihre Urteile auf Erfahrung und Denknotwendigkeit aufzubauen“ (ebd.). Der Gegenstand, an dem sich diese Fähigkeit ausbilden lasse, sei traditionell der deutsche Aufsatz: „Als das wichtigste Mittel hierzu wurde ehedem der vom Lehrer zu kritisierende deutsche Aufsatz angesehen. Umso mehr ist zu bedauern, daß neuerdings dem sogenannten Erlebnisaufsatz gegenüber die abhandelnde Form mehr und mehr zurücktritt und daß insbesondere das literarische Gebiet fast verschlossen ist“ (ebd.)[27]

Was zeigt nun die kleine Debatte aus dem Jahr 1930? Deutlich wird durch den Initialbeitrag Dibelius’ zunächst, dass seitens der nun in die Kritik geratenen Hochschulen selbstbewusst die Position vertreten wird, dass die höheren Schulen für die notwendige Qualität der Studienanfänger:innen zu sorgen haben. Wegen der Synthese von Abitur und Allgemeiner Hochschulzugangsberechtigung sehen sich Professoren als mandatiert an, die Anforderungen an die mit dem Abitur selektiv zertifizierte Reife zu definieren. Dies gilt in quantitativer Perspektive mit Blick auf die Zahl

25 Hermann Knust (1878–?) war Volksschullehrer und Studiendirektor am Lyzeum Pasewalk für die Fächer Geschichte, Französisch, Englisch. 1921 promovierte er zum Dr. phil. in Greifswald (vgl. Ortmeyer, 2018). Im NSLB war er ab 1933, in der NSDAP ab 1937 Mitglied. Knust veröffentlichte 1933 „Volkstum und Rasse im Geschichtsunterricht“.

26 Vgl. hierzu etwa die Abhandlung Pischels (1912, S. 165 ff.).

27 Vgl. zum Hintergrund Thies (1928a, 1928b).

der Ausgewählten als auch in qualitativer Hinsicht. Im letzten Punkt wird auf Humboldts Arbeitsteilung rekurriert, wonach die Schulen gesichertes Wissen zu vermitteln haben, während den Universitäten die Aufgabe einer Bildung durch Wissenschaft[28] zufällt. Seitens der höheren Schulen wird bei allem Verständnis für die Nöte der Professoren einem Dienstleistungsverständnis widersprochen, das die Schulbildung auf Stofffülle und passive Aneignung verpflichten will. Auch hierbei kann man sich auf Humboldt und die Vorstellung formaler Bildung berufen, dass es darauf ankomme, an „einer möglichst geringen Anzahl von Gegenständen" ein der Wissenschaft zugewandtes „Gemüth" vorzubereiten (Humboldt 2010, S. 261). Im Spannungsfeld dieser diskrepanten Normsetzungen stehen die Aufgaben der Abiturprüfung, wenn es hierzu in der Ordnung von 1926 allgemein und anspruchsvoll heißt: „Die Aufgaben sollen die geistige Reife des Prüflings ermitteln, nicht Einzelkenntnisse feststellen. Sie müssen [...] so gestaltet sein, daß die Prüflinge bei ihrer Lösung ihre Befähigung zu wissenschaftlicher Arbeit nachweisen können. Keine Aufgabe darf daher bereits gelösten Aufgaben so nahe stehen oder im Unterricht so weit vorbereitet sein, daß ihre Bearbeitung aufhört, eine selbständige Leistung zu sein" (Ordnung der Reifeprüfung an den höheren Schulen Preußens, 1926, §13 [1]). Mit der Akzentuierung von Selbstständigkeit und wissenschaftlicher Leistung wird hier „geistige Reife" erkennbar weiter ausgelegt als ein Besitz von „fertigen und abgemachten Kenntnissen" (Humboldt, 2010, S. 256). Die „geistige Reife" zeigt sich speziell im deutschen Aufsatz daran, dass die Aufgabe den Prüflingen mehr abverlangt „als die bloße Wiedergabe geläufiger Zusammenhänge" (Ordnung der Reifeprüfung an den höheren Schulen Preußens, 1926, §13 [2]). Wird hier also eine Ausrichtung auf Produktion statt auf Reproduktion dekretiert, um geistige Reife demonstrieren zu lassen, so werden im selben Paragraphen auch Risiken eines solchen Anspruchs adressiert: Der Gedankengang darf nicht zu schwierig sein, die Aufgaben müssen in den Gesichtskreis der Prüflinge fallen, die „Gefahr eines ziellosen Umherschweifens ausschließen" und auf „keinen Fall darf zu ihrer Bearbeitung eine größere Erfahrung und ein reiferes Urteil erforderlich sein, als man bei dem Alter der Prüflinge voraussetzen kann" (ebd.). Augenfällig ist, dass diese Vorgaben die Aufgabensteller:innen vor delikate Probleme stellen. Einerseits soll sich Reife in einer Selbstständigkeit erweisen, die dem Ideal wissenschaftlicher Praxis („Einsamkeit und Freiheit", Humboldt, 2010, S. 255) entsprechen, andererseits gilt es, den bekannten Risiken von Überforderung und Verleitung zu spontaner Borniertheit („Geschwätz", ebd.) auszuweichen. Tendenziell erweist sich dieser Strukturkonflikt als Basis eines dauerhaften Klagemusters, wenn es um das Schreiben und Bewerten von Abituraufsätzen geht. Die Schreibprodukte sollen klar strukturiert, aber keineswegs schematisch heruntergeschrieben, sondern originell sein, die bemühten Stoffkenntnisse sollen stimmen, aber nicht abgedroschen wirken, das gewählte Textformat soll Ausschweifungen verhindern, aber die persönliche Gestaltungskraft nicht einengen, Urteile sollen individuell authentisch wirken, aber nicht an den Haaren herbeigezogen – die Aufzählung könnte fortgesetzt werden.

28 Zu den Ambivalenzen dieser gern genutzten Formel vgl. Tenorth (2018, insb. S. 203–218).

3 Der deutsche Aufsatz als Zeugnis der Reife: Eine schwierige Karriere

Die Karriere des deutschen Aufsatzes im Prüfungskanon des Abiturs ist schwerlich von den heftigen schulpolitischen Debatten zu isolieren, die über das 19. Jahrhundert hinaus das höhere Schulwesen tangierten. Hierbei scheinen uns zwei gegenläufige Linien beachtenswert: Auf der einen Seite hängt dessen Bedeutung selbstverständlich mit dem wachsenden Status des muttersprachlichen Unterrichts im Fächerkanon höherer Schulen zusammen (Kämper-van den Boogaart, 2013, 2019; Reh, Kämper-van den Boogaart & Scholz, 2017). Entsprechende Bewegungen, die die Rolle des Faches in der Stundentafel berühren und die zum Teil heftige Konflikte mit den habituellen Anhängern des neuhumanistischen Gymnasiums und des altphilologischen Bildungskapitals implizieren, sind ihrerseits verwoben mit politischen und kulturellen Transformationen, die im Vormärz ihren Anfang nahmen und die mit einer massiven Aufwertung der deutschsprachigen Literatur mit dem Fixpunkt der Weimarer Klassik einhergingen.[29] Nach 1848 steht der Kampf um die Bedeutung des Deutschunterrichts, und hier nicht losgelöst von den heftigen Debatten über die Alternativen zum Typus des neuhumanistischen Gymnasiums, erheblich im Zeichen einer nationalpädagogischen Programmatik, für die in der Fachdidaktik zweifellos Rudolf Hildebrand und Otto Lyon mit ihrer Zeitschrift für den Deutschen Unterricht maßgebliche Protagonisten waren (Kämper-van den Boogaart, 2017b). Noch vor der umfassenden Transformation eines entsprechend nationalistischen Verständnisses des Fachs zur in der Weimarer Republik verbreiteten „Deutschkunde“ war es der berühmt-berüchtigte Auftritt Wilhelms II. auf der die Zukunft des Abiturs traktierenden Schulkonferenz von 1890, der Zeichen setzte. In seinen vielzitierten Ausführungen monierte der junge Kaiser, dass es dem Gymnasium an einer „nationalen Basis“ mangele, und postulierte, dass „das Deutsche“ zur Grundlage gemacht werden müsse: „wir sollen nationale junge Deutsche erziehen und nicht junge Griechen und Römer“. Gegen die Priorisierung des lateinischen Aufsatzes im Abitur hieß es kategorisch: „Der deutsche Aufsatz muß der Mittelpunkt sein, um den sich alles dreht. Wenn einer im Abiturientenexamen einen tadellosen deutschen Aufsatz liefert, so kann man daraus das Maß der Geistesbildung des jungen Mannes erkennen und beurteilen, ob er etwas taugt oder nicht“ (zit. n. Giese, 1961, S. 197).

Während nun einerseits die zitierte kaiserliche Einschätzung als Hinweis auf eine prioritäre diagnostische Funktion der Leistungen im deutschen Aufsatz für die Feststellung, ob ein Schüler etwas tauge, gelesen werden kann, gibt sich andererseits eine gegenläufige Linie im 19. Jahrhundert immer wieder zu erkennen. So heißt es 1851 in einem anonym publizierten Beitrag mit dem Titel „Der deutsche Aufsatz als Kennzeichen der gewonnenen Bildung“ unter anderem:

29 Zu verweisen wäre hier auf die Literaturdidaktik des Merseburger Gymnasiallehrers Rudolf Heinrich Hiecke (1842) und auf die für den Deutschunterricht gedachte und in Überarbeitungen vielfach aufgelegte Literaturgeschichte des Gymnasiallehrers Karl August Koberstein (1827).

> „Der Schüler soll sich frei bewegen in einer nicht zu definirenden Form und dadurch soll er Form gewinnen: das ist der Widerspruch, den er fühlt, und der ihm diese Art der formalen Beschäftigung so widerwärtig macht. Danach spricht denn der deutsche Abiturientenaufsatz keineswegs die Bildung des Schülers, sondern nur seine Gewandtheit in der Beherrschung der Form aus. […] Es ist ein schlimmes und sehr übles Zeichen, daß die Primaner der Schulen schon gerne deutsche Aufsätze anfertigen über Papsttum und Deutschtum, über Nationalität und Griechen- und Römertum und wie die Sachen sonst noch heißen mögen; denn das beweist nur, daß sie am Ende ihrer Schullaufbahn glücklich dahin gebracht sind, über Dinge zu schreiben, die sie nicht verstehen, oder was dasselbe ist, sich in Phrasen zu ergehen." (Anonym, 1851, S. 283, 285 f.)

Dieses sehr skeptische Urteil über den Bildungswert des deutschen Aufsatzes enthält wichtige Vorbehalte, die sich bis weit in das 20. Jahrhundert immer wieder registrieren lassen. Beginnen wir mit dem zweiten Punkt, der einen wesentlichen Aspekt der steten Überforderungsklagen[30] berührt: Mit seinen Themenstellungen verführe der deutsche Aufsatz die Schüler dazu, ihre Urteilskraft auf Fragen zu richten, die sie keineswegs kundig erfassen könnten.[31] Die Aufsatzerziehung bewirke unter diesen Bedingungen eben nicht, die tatsächliche „Geistesbildung" der jungen Menschen zu fördern, sondern sie fördere stattdessen das Vermögen zu sprachlichen Produktionen, die ohne probate Sachkenntnis zu mehr oder weniger eloquenten Urteilen gelangten. Was das für die Entwicklung der Gymnasiasten bedeute, charakterisierte Nietzsche 1872 in psychologischer Manier:

> „Man muß nur denken, was in einem solchen Alter, bei der Produktion einer solchen Arbeit, vor sich geht. Es ist die erste eigne Produktion; die noch unentwickelten Kräfte schießen zum ersten Male zu einer Krystallisation zusammen; das taumelnde Gefühl der geforderten Selbständigkeit umkleidet diese Erzeugnisse mit einem allerersten, nie wiederkehrenden berückenden Zauber. Alle Verwegenheiten der Natur sind aus ihrer Tiefe hervorgerufen, alle Eitelkeiten, durch keine mächtigere Schranke zurückgehalten, dürfen zum ersten Male eine litterarische Form annehmen: der junge Mensch empfindet sich von jetzt ab als fertig geworden, als ein zum Sprechen, zum Mitsprechen befähigtes, ja aufgefordertes Wesen. Jene Themata nämlich verpflichten ihn, sein Votum über Dichterwerke abzugeben oder historische Personen in die Form einer Charakterschilderung zusammenzudrängen oder ernsthafte ethische Probleme selbständig darzustellen oder gar, mit umgekehrter Leuchte, sein eignes Werden sich aufzuhellen und über sich selbst einen kritischen Bericht abzugeben: kurz, eine ganze Welt der nachdenklichsten Aufgaben breitet sich vor dem überraschten, bis jetzt fast unbewußten jungen Menschen aus und ist seiner Entscheidung preisgegeben." (Nietzsche, 1988, S. 679 f.)[32]

Nietzsches Verweis auf die „Komödie der deutschen Arbeit" hat nicht nur diese Hybris-Effekte für die Schreibenden im Blick, sondern auch die Intervention durch die korrigierenden Lehrer. Würden die Schüler durch die Gelegenheiten einer „allzufrühzeitigen Erregung" dazu verführt, sich stilistisch wie Originalgenies zu gerieren, sei es an den Lehrern, via Korrektur „zu Gunsten einer unoriginalen Durchschnitts-

30 Vgl. Kämper-van den Boogaart (2013). Hier finden sich auch Hinweise darauf, wie die frühen Abiturverordnungen mit dem Risiko der Überforderung umgehen.
31 Beispiele finden sich bei Selbmann (1988).
32 Vgl. Kämper-van den Boogaart & Hamelmann (2013) und Reh (2017, S. 117 f.).

anständigkeit“ zu intervenieren und die „uniformierte Mittelmäßigkeit“ „verdrossen“ zu loben (ebd., S. 680). Auch wenn man geneigt sein mag, die Anstaltskritik Nietzsches wegen dessen unzweifelhaft antidemokratischen Tenors beiseitezulegen, scheint hier doch in der Sache ein Dilemma zur Sprache gebracht zu werden, das sich in der Geschichte des Abituraufsatzes immer wieder zeigt und das, unseren Beobachtungen folgend, Korrektur- und Bewertungspraktiken prägt. So soll der Aufsatz in entscheidender Hinsicht authentisch sein, nämlich persönliche Bildung, darunter sprachliche und solche der Urteilskraft, dokumentieren, was eine reine Reproduktion intersubjektiv geteilten Wissens ausschließt. Ebendiese Erwartung forciert die Tendenz zu performativen Schreibpraktiken, also solchen, die im Idealfall eine gewisse persönliche Kühnheit dokumentieren, sowie zu Themen, deren Bearbeitung Praktiken des Fingierens nahelegen: Die Schreibenden äußern sich abwägend und urteilend, als ob sie Kundige wären und als ob ihnen die vorgelegte Menschheits- oder Literaturfrage eminent wichtig wäre. Und auch beim Korrektor ist Nietzsches Wink zu einer verdrossenen Haltung keineswegs unmotiviert: Einerseits liest er die Schreibprodukte seiner Schüler, als ob diese literarische Werke darstellten, und will unterhalten und angenehm überrascht werden wie ein Romanleser. In dieser Hinsicht moniert er auch das Epigonale, als das sich das bloß artig Gelernte oder gar Auswendiggelernte zeigt. Andererseits muss er aber auf die Einhaltung der Regeln und auf Mäßigung dringen, „Excesse der Form“ (ebd.) oder der Urteile ausmerzen und sich, wenn man so will, über sein Bedürfnis, sich nicht zu langweilen, professionell schämen (Reh, Kämper-van den Boogaart & Scholz, 2017).

Besagte „Excesse der Form“ dürften mit jenen antinomischen Erwartungen zu tun haben, die der Anonymus 1851 beklagt, wenn er darauf hinweist, dass den Kandidaten abverlangt werde, sich „in einer nicht zu definirenden Form“ zu bewegen und eben „dadurch […] Form [zu] gewinnen“ (s. o.). Selbst in der hochgradig reflektierten und sehr einflussreichen Aufsatzdidaktik, die Ernst Laas mit der zweiten Auflage 1877 in zwei Bänden vorlegt (Ludwig, 1988, S. 199 ff., S. 221 ff.), wird deutlich, wie bedingt das von den Primanern Erwartete letztlich als erlernbar gilt. Laas misst zwar dem Aufsatzschreiben eine hochgradige Bedeutung bei und unterstellt, dass der Aufsatz erprobe, ob sein Autor sich „tüchtige Bildung und verständige Methoden“ angeeignet habe, warnt aber vor der enthusiastischen Vorstellung, dass der „Abiturient in seinem deutschen Probeaufsatz ein Bild seiner ganzen geistigen Gabe abgeben“ könne (Laas, 1877, S. 67). Entscheidender noch dürfte sein, dass Laas trotz seiner vielfältigen Hinweise nicht der Auffassung anhängt, dass der Plan[33] zu einem gelungenen Aufsatz „a priori *beigebracht* werden kann, sondern aus der *Eigenthümlichkeit* der Sache *organisch* hervorwachsen muss“: „Jedes von aussen herangebrachte Schema vergewaltigt sie, zerstört die *Eigenthümlichkeit*“ (ebd., S. 203, H. i. O.). Um dies dem Schüler zu verdeutlichen, konstatiert er gar: „Der deutsche Aufsatz muss sein, man erschrecke nicht, wie ein Kunstwerk, wie ein Organismus“ (ebd., S. 204). Die Hintergründe dieses auch für die Korrektur und Bewertung folgenreichen Verständnisses können hier nicht en detail nachgezeichnet werden; deutlich werden sollte allerdings, dass ihm

33 Bei Laas liegt der Akzent auf der „Inventio“, auch eine Konsequenz seiner Ablehnung schematisierender Aufsatzlektionen.

nicht allein ein ästhetisches Primat der Totalität und ein auch politisch später durchaus fragwürdig werdender Holismus natürlicher Entwicklung zugrunde liegt, sondern zudem ein Respekt vor der Individualität der Sache, der Notwendigkeit, die Planung und Ausgestaltung des Aufsatzes dem konkreten Gegenstand und der auf ihn eingenommenen thematischen Perspektive anzupassen. Dass das Schreiben eines solchen Aufsatzes neben der Erfahrung und dem gebührenden Weltwissen eine meditative Besinnung der Autorinnen und Autoren voraussetzen soll, verwundert nicht. Dass solche Meditation als Gelingensvoraussetzung wiederum Muße voraussetze, ist dann ein Punkt, den ein anderer Fachdidaktiker, nämlich der Karlsruher Gymnasiallehrer und Schulleiter Gustav Wendt, einige Jahre später macht:

> „Es sind in der Regel nicht die schlechtesten Köpfe, die etwas längerer [sic!] Zeit zum Nachdenken und zur Wahl der bezeichnendsten Ausdrücke brauchen; die Leichtigkeit, womit manche ihre Worte aufs Papier werfen, ist gar nicht selten nichts als ein Zeichen grosser Oberflächlichkeit. Eben deshalb ist es in nicht wenigen Fällen sehr gewagt, die geistige Reife eines jungen Menschen nach dem Ausfall einer deutschen Klausurarbeit zu messen." (Wendt, 1896, S. 124)

Wendt stellt zudem noch die labilen Nerven der Examinierten in Rechnung und den äußeren Schreibdruck, um mindestens für Ausgleichmöglichkeiten zu plädieren. Denn: „Dazu aber, dass man den deutschen Aufsatz aus der Reifeprüfung ganz fortlässt, werden sich die deutschen Schulverwaltungen schwerlich sobald entschließen" (ebd.).

4 Abituraufsatz und Deutschunterricht 1975 – Streit um die Normierung

Von den hier skizzierten Problemen des deutschen Abituraufsatzes ist seine politisch gesetzte Bedeutung, die mit der Schulkonferenz 1890 deutlichen Niederschlag fand, schwer zu trennen. Ein nicht geringer Teil dieser historisch zu registrierenden Probleme dürfte gerade darauf zurückgehen, dass der Aufsatz sowohl in seiner thematischen Ausrichtung als auch in seiner Funktion als Indikator für die geistige und politische Reife über den curricularen Kontext des Deutsch- und muttersprachlichen Schreibunterrichts hinausging. Unter der Prämisse, dass dem Deutschunterricht eine hegemoniale Rolle im Fächerkanon zukommen sollte, so die Apologeten von Deutschkunde und einer Deutschwissenschaft bereits deutlich vor 1933, ergab sich daraus wohl das konkrete und durchaus beklagte Problem der Überforderung und – seitens der Lehrkräfte – der stofflichen Überfrachtung ihres Unterrichts. Nach 1945 war es dann insbesondere die Marburger Seminarleiterin und Didaktikerin Erika Essen, die Anstrengungen unternahm, um dem sachlichen Dilettieren im Deutschunterricht den Garaus zu machen und auch den muttersprachlichen Unterricht stärker als einen Fachunterricht wie andere zu profilieren. Eine ähnliche, aber radikalere Stoßrichtung zeigten nach 1968 Reformvorschläge, die am Veto des Wissenschaftsrates scheiterten und daran, das Fach Deutsch beziehungsweise die Nationalphilologien an Schulen

und Universitäten zugunsten eines literaturwissenschaftlichen und eines linguistischen Unterrichtsfachs aufzulösen, in dem die in Fachstudien entsprechend ausgebildete Lehrkraft sich je auf die professionelle Kompetenz der von ihr studierten Domäne beschränken sollte (Kolbe, 1969). Mindestens ein Problembewusstsein dafür, „daß im Deutschunterricht über schlechthin alles geredet und geschrieben werden kann" (Lämmert, 1991, S. 79), zeigen bereits die „Empfehlungen für die Neuordnung der Höheren Schule", die der Deutsche Ausschuss für das Erziehungs- und Bildungswesen im Oktober 1964 vorstellte. Hier unterscheiden die Autorinnen und Autoren zwischen zwei Grundbildungsfunktionen des Deutschunterrichts, einer sprachlichen und einer literarischen. Hierzu heißt es: „Die literarische Grundbildung steht herkömmlicherweise im Vordergrund. Für sie vor allem bringt der Deutschlehrer die wissenschaftlichen Voraussetzungen von der Universität mit. Im Blick auf sie ist Deutsch ein Fach neben anderen Fächern"[34] Dass das in Hinblick auf die sprachliche Grundbildung nicht gilt, wird damit begründet, dass dies Sache auch anderer Fächer sei, in denen fachsprachlich gesprochen und geschrieben werde. Dies führt zwar nicht zur Missachtung einer deutschunterrichtlichen Sprachbetrachtung (im Gegenteil werden hier curriculare Vorschläge für die Oberstufe gemacht, die auch heute noch als modern gelten können), es führt aber zu der Feststellung: „Um der Sachtreue willen sind bei [...] schriftlichen Übungen facheigene Aufgaben des Deutschunterrichts vorzuziehen".[35] Als facheigene Aufgaben gelten den Gutachterinnen und Gutachtern solche Arbeiten, die den Bereichen literarischer Bildung zuzuschlagen sind. In gewisser Weise lässt sich die Aufteilung in einen das Fachliche transzendierenden Bereich sprachlicher Kompetenzen und in einen fachlich umrissenen Kernbereich auch mit dem „Tutzinger Maturitätskatalog" von 1958 in Verbindung bringen.[36]

Als einschneidender müssen zweifellos die Modifikationen gesehen werden, die für den deutschen Abituraufsatz mit den „Normenbüchern Deutsch" beziehungsweise den „Einheitlichen Prüfungsanforderungen in der Abiturprüfung Deutsch" (KMK-Sekretariat, 1975) verbunden sind, die die KMK unter vehementen Protesten[37] als Reaktion auf die hier schon erwähnten Entscheidungen des Bundesverfassungsgerichts über Zulassungsbeschränkungen zum Hochschulstudium herausbrachten (Bormann, 1978, S. 27 f.). Die erste Entscheidung des BVG fiel 1972, ging als „Numerus-Clausus-Urteil"[38] in den interessierten Sprachgebrauch über und führte unter anderem zur Einrichtung einer zentralen Studienplatzvergabe in zulassungsbeschränkten Studienfächern (ZVS). Moniert wurde in diesem Zusammenhang durch das BVG auch, dass die NC-Regelungen der Bundesländer sich vornehmlich der Abiturnote als Kriterium bedienten, diese aber bundesweit nach uneinheitlichen Maßstäben vergeben würde. Auch wenn das BVG-Urteil sich gegenüber alternativen Zulassungskriterien mindestens aufgeschlossen zeigt, setzte die KMK eine Arbeitsgruppe „Rahmen-

34 Empfehlungen und Gutachten des Deutschen Ausschusses für das Erziehungs- und Bildungswesen 1953–1965 (1966, S. 527, 597, 735).
35 Ebd., S. 598.
36 Ebd. S. 1030 f.
37 In aktualisierender Perspektive siehe Benner (2012).
38 Einsehbar u. a. hier: https://www.hrk.de/fileadmin/redaktion/hrk/02-Dokumente/02-03-Studium/02-03-04-Hochschulzulassung/bverfg_nc-urteil_18071972.pdf.

abiturprüfung“ ein, die ihrerseits Fachkommissionen mandatierte, fachspezifische Vorgaben zu entwickeln, um die Leistungsbewertung im Abitur national besser vergleichbar zu gestalten. Im Ergebnis wurden ab 1975 die ersten von fünfzehn „Normenbüchern“ publiziert, darunter auch die Anforderungen für das Fach Deutsch (KMK-Sekretariat, 1975). In Hinsicht auf die schriftliche Prüfung beziehungsweise für den Abituraufsatz[39] werden hier die Aufgabenarten kodifiziert: „Analyse nach Texten“ (fiktional oder nichtfiktional), die „Problemerörterung anhand von Texten oder Materialien“, „die Problemerörterung mit fachspezifischem Thema ohne Textgrundlage“. Beispielhaft werden hierfür die folgenden Aufgaben angeführt:

> „Problemerörterung anhand von Texten: Gottfried Benn, Können Dichter die Welt ändern? Rundfunkinterview 1925 [...] Aufgabe: 1. Stellen Sie das Problem dar, um das es in dem vorliegenden Text geht, und arbeiten Sie den Standpunkt von B. heraus! 2. Beschreiben Sie den Aufbau des Gesprächs! 3. Nehmen Sie zu der Position von B. Stellung, und gehen Sie dabei auf andere Auffassungen zu dem Problem ein!“
>
> „Problemerörterung ohne Textgrundlage: *Thema*: Gibt es ein gesellschaftliches Interesse an Lyrik?“
>
> „Analyse von Texten aus: Alexander Solschenizyn, Der Archipel Gulag. Aufgabe: Analysieren Sie den Anfang dieses Werkes vornehmlich unter Berücksichtigung der Intention des Autors, der Erzählstruktur und der Interdependenz von Intention und Erzählhaltung. Zeigen Sie dabei die Bedeutung auf, die den Faktoren Situation und Vorwissen bei der Textrezeption von Lesern zukommt, die sich innerhalb des dargestellten Machtbereiches befinden, und solcher, die diesem Machtbereich nicht angehören.“
>
> „Analyse von Texten: Typischer Bericht eines Massenblattes z. B. über einen sensationellen Kriminalfall. Aufgabe: 1. Verfassen Sie einen Bericht, in dem alle sachdienlichen Informationen enthalten sind, die Sie dem Zeitungstext entnehmen können! 2. Welche Intentionen verfolgen die Verfasser mit ihrem Text, und welcher sprachlichen Mittel bedienen sie sich dazu? 3. Welchen Lesererwartungen wird der Text gerecht?“ (KMK-Sekretariat, 1975)

Im Vergleich zu früheren Kodifizierungen folgt durch diese Festlegungen eine deutliche Verfachlichung der Erwartungen an die Textproduktion. Dies gilt insbesondere für den Bereich, der zuvor durch thematische Abhandlungen, freie Erörterungen beziehungsweise Besinnungsaufsätze abgedeckt wurde und der die notorische Tendenz hatte, weltanschauliche Fragen in den Fokus zu nehmen und für das böse, aber nicht unpassende Wort „Gesinnungsaufsatz“ zu sorgen. Für das Format der Problemerörterung mit fachspezifischem Thema ohne Textgrundlage heißt es sogar in auffälliger Redundanz: „Das Thema der Problemerörterung ohne Textgrundlage wird den Gegenständen des Deutschunterrichts entnommen, orientiert sich an seinen Lernzielen und fordert Auswahl und selbständige Verarbeitung der im Unterricht erworbenen Kenntnisse“ (ebd.). Dass diese Form der Themenentwicklung indessen nicht leicht-

39 Der „Spiegel“ vermeldet in dieser Hinsicht für 1975 eine ruhige Lage: „Wie eh und je wurde auch im Jahre 1975 versucht, bei den Themen für den Abitur-Aufsatz den ewigen Werten und den Zeitläuften gleichermaßen gerecht zu werden. So durften in Baden-Württemberg die Kandidaten unter anderem wählen, ob sie fünf Stunden lang ein Stückchen ‚Iphigenie‘ betrachten, Gedichte von Eichendorff und Heym ‚interpretieren und vergleichen‘ oder sich über das Schlagwort ‚Kommunikation‘ auslassen wollten. Im Saarland lieferten Zitate von Francois La Rochefoucauld (gestorben 1680), Alfred Döblin (gestorben 1957), Alexander Mitscherlich und Heinrich Böll vier der fünf Themen“ (DER SPIEGEL, 1975).

zufallen schien, zeigt das Aufgabenbeispiel, das sich unter dem Strich doch nicht allzu sehr von früheren literarischen Abhandlungen zu unterscheiden scheint. Wie Eiben-Zach (2021) in ihrer Analyse ausgewählter Aufsätze des Projektkorpus sehr detailliert zu zeigen vermochte, verführt dieser Aufsatztypus dazu, unter dem Aspekt sozialer Erwünschtheit auf formelhafte Bekenntnisse zurückzugreifen. Das Anforderungsraster der EPA 1975 beugt dem zwar insofern vor, als erwartet wird, dass die Schreibenden, wie annonciert, konkrete Unterrichtskenntnisse zu gattungspoetischen Programmen mobilisieren sollten, das Raster lässt aber eine Bestimmung dessen aus, was als „gesellschaftliches Interesse" an Lyrik gelten und ob mit der Frage die empirische Gesellschaft erfasst sein soll. Offensichtlich nicht im Rahmen der Erwartungen liegt zudem die Variante einer globalen Verneinung der Entscheidungsfrage („Gibt es …?"). Nicht vorgesehen ist mithin offenbar eine Problemerörterung, die zu dem Befund gelangt, dass kein gesellschaftliches Interesse an Lyrik registriert werden kann. Ohne dies hier detailliert ableiten zu können, zeigt sich an diesem Aufgabentypus, dass trotz der stärkeren Akzentuierung der curricularen Relevanz alte Probleme, ob als Überforderung oder als durchsichtige Evokation von Klischeeaussagen registriert, Bestand haben. Dies berührt bedingt auch die Kritik, die von den pädagogischen Kritikern des „Normenbuchs" vorgetragen wird. In dem Paperback mit dem deutlichen Titel „Abitur-Normen gefährden die Schule" sind es – neben dem Teilabdruck eines prominenten Essays von Hans Magnus Enzensberger – der Erziehungswissenschaftler Dieter Lenzen und der renommierte Linguist Dieter Wunderlich, die sich die EPA für das Fach Deutsch vornehmen (Lenzen & Wunderlich, 1976). Ihr Befund zu den schriftlichen Aufgabenformaten liest sich so:

> „Die Autoren des Normenbuchs [optieren] für die Vergangenheit: So standen – summarisch gesprochen – im Vordergrund der schriftlichen Aufgaben im Deutschunterricht immer schon die reflexive Problemerörterung und die mehr oder weniger werkimmanente Erörterung literarischer Texte. Derartige Textproduktionen dienten weniger außerunterrichtlichen Zwecken (die einzige Verwendung der Aufsätze erschöpfte sich in ihrer Verwendung für die Leistungsmessung), sondern der ziemlich vagen Kontrolle relativ allgemeiner Lernziele" […]." (ebd., S. 198 f.).

Zwar deckt sich die vermeintliche Kontinuitätslinie nur in Umrissen mit (weniger summarischen) Befunden der Fachgeschichte[40], doch richtig ist sicher, dass die EPA insbesondere mit der Problemerörterung ohne Textgrundlage keine vollständige Zäsur darstellen.[41] Studiert man die vielen Monita, die die fachliche Kritik an den EPA seitens Lenzen und Wunderlich munitionieren, fällt noch ein anderer und umfassenderer Aspekt auf: Indem die EPA im Vergleich zu vorherigen Abiturverordnungen die Prüfungsinstrumente relativ detailliert kodifizieren, werden sie deutlich anfälliger für

40 Insbesondere die Rede von einer textimmanenten Erörterung muss verwundern. Sicher kann man von Impulsen zu werkimmanenten Interpretationen sprechen. Als Abituraufgaben tauchen diese aber erst relativ spät als Reaktion auf negative Erfahrungen mit Themen auf, wie sie etwa das Aufgabenbeispiel zur gesellschaftlichen Relevanz zur Lyrik repräsentiert. Vgl. zu den Projektbefunden hierzu insbesondere die Dissertation von Eiben-Zach (2021) sowie Reh & Eiben-Zach (2021).

41 Unter Implementationsaspekten wäre das wohl auch kaum probat gewesen. Zu den konkreten Implementationsbedingungen der EPA und der Herausforderung, ein neues Abitur mit der laufenden Oberstufenreform zu koordinieren, vgl. den Erfahrungsbericht aus der Praxis von Kutzschbach (1980).

fachdidaktische Einwände. Von solchen[42] gibt es bei Lenzen und Wunderlich (1976), sich durchaus widersprechend, einige. So wird zum Beispiel konstatiert, dass die EPA nicht die Pluralität der fachdidaktischen Normbildungen – vom konservativen Ulshöfer bis zum linken Bremer Kollektiv – spiegeln. Attackiert werden Beurteilungskriterien, die sich auf die Verwendung bildungssprachlicher Mittel richten, da diese Festlegung die (seinerzeit) aktuellen Befunde zu schichtenspezifischen Sprachbarrieren konterkarierten.[43] Interessanterweise wird ebenfalls die Wissenschaftsorientierung der Abiturfestlegungen an den Pranger gestellt, da mit diesen ein Bias zugunsten der Mainstream-Literaturwissenschaft und der Textlinguistik einherginge, wobei unter der Hand von den Autoren das ältere Thema reformuliert wird, dass der höhere Schulabschluss den Erwerb allgemeinerer Kenntnisse und Fähigkeiten und nicht diejenigen einer einzelnen Wissenschaftsdisziplin zu erfassen habe (s. o.). Neu ist allerdings die Argumentationsweise: In Anschlag gebracht werden nunmehr Positionen, wie sie in den 1970er-Jahren im Kontext einer sogenannten kommunikativen Wende der Deutschdidaktik vertreten wurden – einer Programmatik, die in der Konsequenz einerseits das Unterrichtsfach aufwertet, andererseits disziplinäre Grenzen porös werden lässt (vgl. Kämper-van den Boogaart, 2022b) Zielt die Kritik der Autoren auf dieser Argumentationsbasis eher darauf, die Abiturprüfung und den Unterricht stärker auf „wichtige Verwendungssituationen der antizipierbaren Lebens- und Berufspraxis des Schülers“ (Lenzen & Wunderlich, 1976, S. 204) auszurichten, lässt sich in einer anderen Darstellung dezent ein anderer Ton vernehmen. Für einen Jubiläumsband zeichnet der Deutschlehrer Hans Dieckhöfer 1992 nach, zu welchen Themen an seinem Dorstener Gymnasium seit 1877 Abituraufsätze geschrieben wurden. Seine durchweg interessanten Kommentare sind dabei bemüht, die lokalen Ereignisse in bildungsgeschichtlich relevante Kontexte einzurücken. Die Entwicklung nach 1972 wird hierbei von ihm durchaus positiv als eine der stärkeren fachlichen Fundierung des Abiturs gewertet und auch das Bemühen, Leistungen objektiver und vergleichbarer zu bewerten, explizit gewürdigt. Interessanterweise aber findet sich am Ende des fundierten Beitrags die folgende Bemerkung:

> „Deswegen – und weil wir Jugendlichen Antworten schuldig sind, wann immer sie uns fragen – müssen wir Deutschlehrer uns heutzutage nicht in ängstlicher Selbstbeschränkung angesichts eines durch die Vergangenheit des Faches di[s]kreditierten Gesinnungsbegriffs im Schweigen verkriechen, sondern können und müssen auch unsere persönliche Antwort geben, wo wir nach Haltungen und Werten gefragt werden, und haben – gerade auch in unseren Tagen, wo die Notwendigkeit dazu besteht – die Pflicht zu sagen, was in unserer Gesellschaft und darüberhinaus angesichts der oben genannten Werte vertretbar ist.“ (Dieckhöfer, 1992, S. 143)

42 Ein gravierender weiterer Einwand bezieht sich auf die als naiv eingestufte Vorstellung, mit den EPA-Aufgabenformaten zuverlässige Messinstrumente vorgelegt zu haben.

43 Die Fokussierung unterrichtsrelevanter Sprachbarrieren rekurrierte sehr stark auf Befunde und Erklärungsmodelle des britischen Soziolinguisten und Bildungssoziologen Basil Bernstein (1970), die in deutscher Übersetzung zunächst als Raubdruck kursierten, bevor sie dann von verschiedenen deutschen Verlagen in unterschiedlicher Zusammenstellung publiziert wurden. Eine Übersicht vermitteln die vier Bände „Class, Codes and Control“ (London: Routledge, 1971 ff.: Volume I: Theoretical Studies Towards a Sociology of Language; Volume II: Applied Studies Towards a Sociology of Language; Volume III: Towards a Theory of Educational Transmissions; Volume IV: The Structuring of Pedagogic Discourse). Für die Diskussion in der Bundesrepublik zudem sehr wirkungsmächtig ist Oevermann (1972).

In diesem Beharren auf einen Deutschunterricht, der im Dialog mit den Lernenden Haltungen und Werte thematisiert, die erzieherisch über die Grenzen des Fachlichen hinausweisen, zeigt sich wohl, dass trotz der affirmativen Haltung gegenüber einer Abrechnung mit der Gesinnungspädagogik früherer Jahrzehnte die mit den EPA auf den Weg gebrachte Abstinenz letztlich nicht geteilt wird. Ähnliches lässt sich historiografischen Darstellungen entnehmen, die explizit als Forschungsbeiträge firmieren. So nehmen Jasper und Müller-Michaels zugunsten einer „Vernunft der Praxis" Stellung, wenn sie in ihrer Darstellung zum „Abituraufsatz in Westdeutschland von 1945–1989" konstatieren:

> „In dem Augenblick allerdings, in dem, seit Anfang der siebziger Jahre, das Selbstbewusstsein der jungen Generation gewachsen ist, wird die freie Argumentation (deklariert als ‚Besinnungsaufsatz') abgelöst durch sachgerechte Analysen von expositorischen und fiktionalen Texten. Das Ziel, Probleme erörtern zu lernen, die eigene Position mit Argumenten zu stützen, fachübergreifend zu denken und Ergebnisse zu sichern, wird aufgegeben. Fachliche soll allgemeine Bildung auch im Deutschunterricht ersetzen. Die Vernunft der Praxis gegenüber politischer Ideologie zeigt sich darin, dass heute die beiden Grundformen der Erörterung (aufbauend und dialektisch) als Aufsatzformen wiederentdeckt werden." (Jasper & Müller-Michaels, 2011, S. 387 f.)

Inwieweit es legitim ist, die Prüfungsanforderungen als Reflexe „politischer Ideologie" zu klassifizieren und für die eigene Position „Vernunft" zu reklamieren, sei dahingestellt. Angesichts der empirisch zu beobachtenden Praxis scheint die Einschätzung motiviert, dass die strenge Beschränkung auf die fachlichen Ziele des Deutschunterrichts nach 1975 nur eingeschränkt verfängt. Verantwortlich dafür könnten neben den pädagogischen Selbstkonzepten der Deutschlehrkräfte auch objektive Schwierigkeiten sein, sich im Umgang mit Sprachprodukten und -prozessen auf ein rein Fachliches zu beschränken. Dies gilt nicht nur für die Problemerörterung als relevant geltender Themen, in der sich die geforderten Urteile in der Regel komplementär zu lebensweltlichen Klärungen verhalten dürften,[44] vielmehr trifft dies gerade auch auf Fragen zu, die eine Ethik sprachlichen Handelns berühren. Ebenso sind Analysen expositorischer und poetischer Texte in der Schule kaum von jenen Aspekten zu isolieren, die die Textwelten in Relation zu Vorstellungen sozialer Wirklichkeit aufwerfen: Auch wenn man sich in Erzähltextanalysen zum Beispiel auf die Erörterung der applizierten Fokalisierungstechniken beschränken könnte, werden spätestens mit der Berücksichtigung der Textintentionen die Grenzen der Formanalysen überschritten, wie auch Lämmert einräumt, wenn er 1970 in der Auseinandersetzung mit dem Wissenschaftsrat zugesteht, dass „mit sprachlichen und literarischen Texten immer auch Kenntnisse und Erfahrungen über reelle Vorgänge, Fakten, Werte vermittelt werden" (Lämmert, 1991, S. 79).

44 Lebenswelt, hier im Sinne von Habermas (1981). Die Beispielaufgabe der EPA 1975 zum Beispiel basiert auf normativen Verständigungen über Gesellschaft.

5 Schlussfolgerungen

Die Schlaglichter, die hier auf die Problemgeschichte des deutschen Abituraufsatzes geworfen wurden, illuminieren eine Herausforderung, die in ganz anderer Perspektive auch aktuelle Auseinandersetzungen um die Bildungsstandards der allgemeinen Hochschulreife und die Vorgaben durch zentrale Aufgaben berühren. Versucht man sich nämlich an einer Geschichte des Abituraufsatzes, die über eine Ideologiegeschichte seiner Themen hinausreicht, stößt man auf die Schwierigkeit, keine determinierenden, sondern immer nur sich wechselseitig beeinflussende Kontexte identifizieren zu können, über die sich Entwicklungslinien motivieren lassen. Zunächst scheint der Gedanke naheliegend, Transformationen des Gegenstands, des Prüfungsaufsatzes, seiner Formen und Themen beziehungsweise der damit verbundenen Praktiken auf Entwicklungen pädagogischer und didaktischer Diskurse zurückzuführen, wie dies auch Jasper und Müller-Michaels (2011) anhand eines Phasenmodells praktizieren. Indes geht diese Rechnung nur mit erheblichen Einschränkungen auf, wie gerade am letzten Beispiel noch einmal zu sehen war. So sind die konkreten Formatierungen durch die EPA wohl nur zu erklären, begreift man sie als ein Amalgam politischer, juristischer und didaktischer Einflüsse. Wäre es beispielsweise so, dass die (als phasenprägend eingestufte) Kommunikationsdidaktik die Festlegungen diktiert hätte, wäre weder zu erklären, wieso die Zeichen auf Fachlichkeit gestellt wurden, noch, wieso überhaupt an einem Aufsatzformat festgehalten wurde.

Dass Letzteres der Fall war, kann zweifellos auch darauf zurückgeführt werden, dass die je aktuellen didaktischen Normen die Praxis und ihr Brauchtum nur sehr eingeschränkt prägen – ein Umstand, den realistische Strategien, neue Prüfungsformate und Textformen zu implementieren, kaum ignorieren können. Die Beharrlichkeit eines didaktischen Brauchtums dürfte sich auch aus generationsübergreifenden Vorstellungen speisen, zu denen der finale Aufsatz als performative Szene des Abschließens mit all seinen narrativ tradierten Erlebnismomenten zählt. Dass der Aufsatz in der Szenographie des Abiturs diese Rolle einnimmt und gegen alle Anfechtungen behauptet, ist nicht ohne Kenntnis einer Vorgeschichte zu begreifen, die der schriftlichen Selbstdarstellung der Personen eine so gewichtige Rolle für die Manifestation von Bildung, gebotener Maturität und Hochschulreife eingeräumt hat – eine Rolle, die sich kaum nur der Auslassungen einer Aufsatzerziehung oder der Schreibdidaktik verdankt –, und erst recht nicht der Kriterien von Messtheorien. Vielmehr zeigen sich hier, wie auch im Status des Deutschunterrichts im Fächerkanon höherer Bildung, klare politische Intentionen – und dies gegebenenfalls im Dissens mit den Universitäten in ihrer Rolle als Abnehmer examinierter Schulabgänger:innen.

Nicht zu erklären wäre ansonsten, wieso bis heute begründet von einer systematischen Überforderung der Prüflinge gesprochen wird (Steinmetz, 2013; Kämper-van den Boogaart, 2017a) und trotzdem an einem hybriden Prüfungsformat festgehalten wird, das sowohl einen gründlichen Kenntniserwerb dokumentieren wie die Kompetenz demonstrieren soll, die Ebene schlichter Kenntnisse intellektuell hinter sich zu lassen. Charakteristisch angesichts der frappanten Beharrungskräfte einer problema-

tischen Gattung und bezeichnend für die Hybridität normativer Erwartungen ist wohl bis heute, dass einerseits holistische Ansprüche an eine im Aufsatz zu formulierende Individualität ins Feld geführt werden, die bereits im 19. Jahrhundert zum Vergleich mit dem Kunstwerk motivierten, und dass andererseits diese hochfahrenden Ansprüche mit im lokalen didaktischen Brauchtum verankerten Regularien entschärft werden sollen. Ersichtlich ist dabei, dass diese Hybridität nicht nur die Praxis der Produzierenden prägt, sondern auch die Praktiken der Evaluation, die bis heute zwischen sehr allgemeinen und sehr konkreten Bestimmungen ihren Ort finden müssen.

Literaturverzeichnis

Abitur: Einbahnstraße wird zur Sackgasse (o. V.). *DER SPIEGEL* 22/1975. Zugriff am 21.11.2021 unter https://www.spiegel.de/politik/abitur-einbahnstrasse-wird-zur-sackgasse-a-48d8685a-0002-0001-0000-000041521031

Anhalt, E. (2014). Was bedeutet Studierfähigkeit gestern und heute? In S. Lin–Klitzing, D. Di Fuccia & R. Stengl–Jörns (Hrsg.), *Abitur und Studierfähigkeit. Ein interdisziplinärer Dialog*. (S. 117–141). Bad Heilbrunn: Klinkhardt.

Anonym (1851). Der deutsche Aufsatz als Kennzeichen der gewonnenen Bildung. *Pädagogische Rundschau 27*, 278–286.

Arbeitsgemeinschaft Deutsche Höhere Schule (1958). *Bildungsauftrag und Bildungspläne der Gymnasien*. Berlin: Springer.

Baar, J. (1913). „Kein Schüler liefert mehr einen ungenügenden Aufsatz". *Deutsches Philologen-Blatt, 21*(27), 341–343.

Behrend, F. (1921). Die Reifeprüfung. *Deutsches Philologen-Blatt, 29*(28), 439–443.

Benner, D. (2012). Schule im Spannungsfeld von Input- und Outputsteuerung. In ders.: *Bildung und Kompetenz. Studien zur Bildungstheorie, systematischen Didaktik und Bildungsforschung* (S. 95–109). Paderborn: Schöningh.

Bernstein, B. (1970). *Soziale Struktur, Sozialisation und Sprachverhalten: Aufsätze 1958–1970.* (Übers. aus d. Engl. von e. Pädagogen–Kollektiv an d. Univ. Frankfurt). Amsterdam: de Munter.

Blättner, F. (1960). *Das Gymnasium*. Heidelberg: Quelle & Meyer.

Bobertag, O. (1934). Die Beurteilung von Prüfungsaufsätzen am Berliner Abendgymnasium. In ders.: *Schülerauslese. Kritik und Erfolge*. (S. 117–139). Berlin: Berger.

Bölling, R. (2010). *Kleine Geschichte des Abiturs*. Paderborn: Schöningh.

Bormann, M. (1978). *Bildungsplanung in der Bundesrepublik Deutschland: System und Grundlagen*. Opladen: Westdeutscher Verlag. https://www.hrk.de/fileadmin/redaktion/hrk/02-Dokumente/02-03-Studium/02-03-04-Hochschulzulassung/bverfg_nc-urteil_18071972.pdf

Born, S. & Eiben–Zach, B. (2020). Erträge reduzierter Situierung. Überlegungen zur Adressaten- und Situationsorientierung in Abituraufsätzen der 1960er/1970er Jahre. In J. Heideklang & U. Stobbe (Hrsg.), *Kleine Formen für den Unterricht: Historische Kontexte, Analysen, Perspektiven* (S. 101–120). Göttingen: Vandenhoeck & Ruprecht.

Dausien, B. & Kluchert, G. (2016). „Mein Bildungsgang" – Biographische Muster der Selbstkonstruktion im historischen Vergleich. Beispiele und Argumente für eine historisch-empirische Forschungsperspektive. In *BIOS. Zeitschrift für Biographieforschung. Oral History und Lebensverlaufsanalysen, 29*(2), S. 220–240.

Dibelius, W. (1930). Die Überfüllung der Universität. *Deutsches Philologenblatt 38*, (Heft 18), 265–272.

Dieckhöfer, H. (1992). „Über die Vaterlandsliebe" oder „Dieter Wellershoff: Literatur und Veränderung". Deutsche Reifeprüfungsthemen am Gymnasium Petrinum zwischen 1877 und 1992. In *Festschrift des Gymnasium Petrinum zu Dorsten* (hrsg v. Gymnasium Petrinum Dorsten, S. 118–143). Dorsten.

Edict wegen Prüfung der zu den Universitäten übergehenden Schüler vom 12.10.1812/ Instruction vom 26.06.1812. In W. Demel & U. Puschner (Hrsg.) (1995), *Von der Französischen Revolution bis zum Wiener Kongreß 1789–1815, Deutsche Geschichte in Quellen und Darstellung* (hrsg. v. R. A. Müller, Band 6, S. 373–382). Stuttgart: P. Reclam.

Eiben–Zach, B. (2021). Das Bewerten von Literatur. Literarische Normen im fachdidaktischen Diskurs und in Abituraufsätzen der 1960er Jahre. In L. Brenz & T. Pflugmacher (Hrsg.), *Normativität literarischen Verstehens. Interdisziplinäre Beiträge zur Theorie und Praxis eines zentralen Problems* (S. 175–196). Frankfurt a. M.: Peter Lang.

Eiben–Zach, B. *Literarische Texte als Gegenstand von Abiturprüfungen (1946–1972)*. Unveröffentlichte Dissertation, Humboldt-Universität zu Berlin 2021.

Empfehlungen und Gutachten des Deutschen Ausschusses für das Erziehungs- und Bildungswesen 1953–1965. Gesamtausgabe hrsg. v. H. Bohnenkamp et al. (1966). Stuttgart: Klett.

Engelhardt, K. v. (2021). „Der papierene Drache". Der Reifeprüfungsaufsatz zwischen 1890 und 1930. In S. Reh, P. Bühler, M. Hofmann & V. Moser (Hrsg.), *Schülerauslese, schulische Beurteilung und Schülertests 1880–1980* (S. 171–190). Bad Heilbrunn: Klinkhardt.

Feilke, H. (2017). „Auf offener See" – Beobachtungen zum Gebrauch didaktischer Werkzeuge. *Didaktik Deutsch, 22*(42), 53–69.

Flitner, W. (1959). *Hochschulreife und Gymnasium. Von Sinn wissenschaftlicher Studien und von der Aufgabe der gymnasialen Oberstufe*. Heidelberg: Quelle & Meyer.

Gass–Bolm, T. (2005). *Das Gymnasium 1945–1980. Bildungsreform und gesellschaftlicher Wandel in Westdeutschland*. Göttingen: Wallstein.

Giese, G. (1961). *Quellen zur deutschen Schulgeschichte seit 1800*. Göttingen: Musterschmidt.

Habermas, J. (1981). *Theorie des kommunikativen Handelns. Band 1: Handlungsrationalität und gesellschaftliche Rationalisierung*. Frankfurt a. M.: Suhrkamp.

Herrlitz, H.–G. (1973). *Studium als Standesprivileg. Die Entstehung des Maturitätsproblems im 18. Jahrhundert*. Frankfurt a. M.: Fischer.

Hiecke, R. H. (1842). *Der deutsche Unterricht auf deutschen Gymnasien: ein pädagogischer Versuch*. Leipzig: Eduard Eisenach.

Humboldt, W. v. (2010): Über die innere und äußere Organisation der höheren wissenschaftlichen Anstalten in Berlin. In *Werke IV: Schriften zur Politik und zum Bildungswesen*. (S. 255–266). Darmstadt: WBG.

Jasper, R. & Müller-Michaels, H. (2011). Der Abituraufsatz im Fach Deutsch in Westdeutschland von 1945–1989. In: T. Roberg, S. Susteck & H. Müller-Michaels (Hrsg.), *Geschichte des Deutschunterrichts von 1945 bis 1989. Teil 2: Deutschunterricht im Widerstreit der Systeme* (S. 365–388). Frankfurt a. M.: Peter Lang.

Jeismann, K.–E. (1996). *Das preußische Gymnasium in Staat und Gesellschaft* (2 Bde.). Stuttgart: Klett Cotta.

Kämper-van den Boogaart, M. (2013). Der deutsche Aufsatz und das Abitur – was man vielleicht aus der Geschichte lernen könnte. In H. Feilke [u. a.] (Hrsg.), *Textkompetenzen in der Sekundarstufe II* (S. 41–62). Stuttgart: Fillibach bei Klett.

Kämper-van den Boogaart, M. (2017a). Nach PISA. Bildungsstandards und alledem: Klagen über die Kompetenzen deutscher Abiturienten. In Verein Schweizerischer Deutschlehrerinnen und Deutschlehrer (Hrsg.), *Verstand und Gefühl. Lesen im Spannungsfeld von Allgemeinbildung und Bildungsstandards. Deutschblätter 2016/17* (S. 191–201). Wil: vsdl.

Kämper-van den Boogaart, M. (2017b). Rudolf Hildebrand: Ein historisches Konzept von Sprachbildung im Deutschunterricht. In B. Jostes, D. Caspari & B. Lütke (Hrsg.), *Sprachen – Bilden – Chancen: Sprachbildung in Didaktik und Lehrkräftebildung* (S. 59–76). Münster: Waxmann.

Kämper-van den Boogaart, M. (2019). Geschichte des Lese- und Literaturunterrichts. (Neubearbeitung). In M. Kämper-van den Boogaart & K. H. Spinner (Hrsg.), *Lese- und Literaturunterricht. Teil 1: Geschichte und Entwicklung, Konzeptionelle und empirische Grundlagen* (DTP 11.1., 3., stark überarb. Auf., S. 3–88). Baltmannsweiler: Schneider Hohengehren.

Kämper-van den Boogaart, M. (2022a). *Aufsatzmethodik in der Diskussion: Das Genre des dialektischen Besinnungsaufsatzes in der didaktischen Publizistik der 1950er und 1960er Jahre*. In: Kämper-van den Boogaart, Reh, Schindler & Scholz 2022 (in Vorb.).

Kämper-van den Boogaart, M. (2022b). *Sachlichkeit und Fachlichkeit in der Oberstufe – Erika Essens Vorschläge zu einer Modernisierung des Deutschunterrichts der 1950er und 1960er Jahre und die kommunikative Modernisierung nach 1971*. In: Kämper-van den Boogaart, Reh, Schindler & Scholz 2022 (in Vorb.).

Kämper-van den Boogaart, M. & Hamelmann, M. (2013). „Kritik am Privileg wird zum Privileg: so dialektisch ist der Weltlauf": einige Anmerkungen zur Crux kritischen Urteilens im Deutschunterricht. In C. Dawidowski & D. Wrobel (Hrsg.), *Kritik und Kompetenz: die Praxis des Literaturunterrichts im gesellschaftlichen Kontext* (S. 41–59). Baltmannsweiler: Schneider Hohengehren.

Kämper-van den Boogaart, M., Reh, S., Schindler, Ch. & Scholz, A.: (Hrsg.). (2022): *Abitur und Abituraufsätze zwischen 1882 und 1972. Prüfungspraktiken, professionelle Debatten und Aufsatztexte* (in Vorbereitung). Bad Heilbrunn: Klinkhardt.

Klinger, K. (2018a). Aktenprozesse – Zur Dinglichkeit des Abiturs. *Zeitschrift für Museum und Bildung, 84–85*, 138–152.

Klinger, K. (2018b). Das Abitur – eine Akte: Zu einer historischen Praxeologie des Abiturs. *Jahrbuch für Historische Bildungsforschung, 23*, 172–204. http://nbn-resolving.org/urn:nbn:de:0111-pedocs-166104

Klinger, K. (2018c). „Quo vadis? Eine Selbstüberprüfung beim Abgang von der Schule" – Zu Abitur und Ritual. In K. Berdelmann et al. (Hrsg.), *Transformation von Schule, Unterricht und Profession* (S. 229–245). Wiesbaden: Springer.

Klinger, K. (2019). „Das Leben hat mich mit seinen Härten nicht verschont, und das ist gut so." Der Lebenslauf als Teil der Abiturprüfung im Preußen der Weimarer Republik. In H. Zaunstöck & C. Weiß (Hrsg.), *Moderne Jugend? Jungsein in den Franckeschen Stiftungen 1890–1933* (S. 221–231). Halle (Saale): Harrassowitz Verlag.

KMK–Sekretariat (Hrsg.). (1975). *Beschlüsse der Kultusministerkonferenz: Einheitliche Prüfungsanforderungen in der Abiturprüfung Deutsch (EPA Deutsch)*. Neuwied: Luchterhand.

Knust, H. (1930). Die Überfüllung der Universität. *Deutsches Philologen-Blatt*, 28, 423–425.

Knust, H. (1933). Volkstum und Rasse im Geschichtsunterricht. *Monatsschrift für höhere Schulen*, 32, 282–287.

Koberstein, K. A. (1827). *Grundriß der Geschichte der deutschen Nationalliteratur*. Leipzig: Vogel.

Kolbe, J. (Hrsg.). (1969). *Ansichten einer künftigen Germanistik*. München: Hanser.

Kraul, M. (1984). *Das deutsche Gymnasium 1780–1980*. Frankfurt a. M.: Suhrkamp.

Kraul, M. (2021). Gymnasiallehrer im Vormärz (1830–1848): Zwischen Wissenschaft und Beruf. In R. Casale, J. Windheuser, M. Ferrari & M. Morandi (Hrsg.), *Kulturen der Lehrerbildung in der Sekundarstufe in Italien und in Deutschland. Nationale Formate und ‚Cross Culture'* (S. 163–176). Bad Heilbrunn: Klinkhardt.

Kutzschbach, D. (1980). Tagebuchnotizen zum Schulalltag von Lehrern und Schülern in der reformierten Oberstufe. *Zeitschrift für Pädagogik, 26*, 271–277.

Laas, E. (1877). *Der deutsche Aufsatz in den oberen Gymnasialklassen. Theorie und Materialien*. Erste Abtheilung. Berlin: Weidmannsche Buchhandlung.

Lämmert, E. (1991). *Das überdachte Labyrinth. Ortsbestimmungen der Germanistik*. Stuttgart: Metzler.

Lenzen, D. & Wunderlich, D. (1976): Die Normierung des Sprechens in der Abiturprüfung. *Zeitschrift für Pädagogik, 22*, 113–128.

Leonhard, H. (1930). Höhere Schule und Hochschule. *Deutsches Philologen-Blatt*, 33, 484–485.

Löwe, D. (2020). Abituraufsätze des 19. und 20. Jahrhunderts als bildungshistorische Quellen. In *bildungsgeschichte.de*. Berlin. https://doi.org/10.25523/32552.1

Löwe, D. (i. Druck). Reife auf dem Prüfstand – Debatten und Deutungen über das Abitur und Abiturient*innen im Deutschen Kaiserreich und der Weimarer Republik. In J. Stiller, C. Laschke & L. Goecke (Hrsg.), *Berlin–Brandenburger Beiträge zur Bildungsforschung 2022* (S. 183–216). Berlin: Peter Lang.

Löwe, D., Eiben–Zach, B. & Reh, S. (2020). Moderne Heimat Wolfsburg. Gymnasium im Kontext der Stadt – Stadt im Kontext des Gymnasiums. In A. Kraus & S. Reh (Hrsg.), *Stadt macht Schule. Schulentwicklung im „Soziallabor" der Bundesrepublik, 1945 bis 1980* (S. 19–95). Göttingen: Wallstein.

Löwe, D. & Töpper, D. (i. V. 2022). Vereinfachende Schulverwaltung. Zur Entstehung und Wirksamkeit subjektbezogener Formulare in Volksschule und Gymnasium und zu ihrem Einfluss auf die „Normalität" der Schüler*innen im 19. und 20. Jahrhundert. In B. Moser & J. Garz (Hrsg.), *Das A(b)normale in der Pädagogik*. Bad Heilbrunn: Klinkhardt.

Ludwig, O. (1988). *Der Schulaufsatz. Seine Geschichte in Deutschland*. Berlin: De Gruyter.

Lütgemeier, G. (2008). *Deutsche Besinnungen 1911–1971. Hundert Reifeprüfungsaufsätze als Spiegel ihrer Zeit*. Frankfurt a. M.: Peter Lang.

Marchand, P. (Hrsg.). (2010). *La Baccalauréat, 1808–2008. Certification française ou pratique européenne?* Lyon: ENS Éditions.

Matthias, A. (1939). *Erlebtes und Zukunftsfragen aus Schulverwaltung Unterricht und Erziehung*. Berlin: Weidmannsche Buchhandlung.

Meyer, P. (1899). Oberlehrer und Wissenschaft. *Gymnasium, XVII. Jg.*, Nr. 14, 475–486.

Meyer, P. (1905a). Aufsatzfabriken. *Gymnasium, XXIII. Jg.*, Nr. 19, 678–686.

Meyer, P. (1905b). Aufsatzfabriken. *Gymnasium, XXIII. Jg.*, Nr. 20, 717–722.

Meyer, P. (1905c). Nachträgliches zu den Aufsatzfabriken. *Gymnasium, XXIII. Jg.*, Nr. 24, 863–868.

Müller-Benedict, V. (2016). *Das Wachstum des Hochschulbereichs*. https://www.bpb.de/nachschlagen/zahlen-und-fakten/deutschland-in-daten/220094/hochschulbereich

Nietzsche, F. (1988). Ueber die Zukunft unserer Bildungsanstalten. Vortrag II. In F. Nietzsche, *Die Geburt der Tragödie – Unzeitgemäße Betrachtungen* (Kritische Studienausgabe hrsg. v. G. Colli & M. Montinari, Band 1, S. 672–692). München: dtv/de Gruyter.

Oevermann, U. (1972). *Sprache und soziale Herkunft. Ein Beitrag zur Analyse schichtenspezifischer Sozialisationsprozesse und ihrer Bedeutung für den Schulerfolg*. Frankfurt a. M.: Suhrkamp.

Ordnung der Reifeprüfung an den höheren Schulen Preußens (1926). *Zentralblatt für die gesamte Unterrichtsverwaltung in Preußen*. http://goobiweb.bbf.dipf.de/viewer/image/985843438_0068/1/

Ortmeyer, B. (2018). *NS-Ideologie in der NSLB-Zeitschrift „Die Deutsche Höhere Schule" 1934–1944. Eine dokumentarische Analyse. Die schulspezifische ideologische Ausrichtung der Lehrkräfte in der NS-Zeit. Teil III (Forschungsstelle NS-Pädagogik)*. Frankfurt a. M.: Protagoras Academicus.

O.V. (1975). Abitur: Einbahnstraße wird zur Sackgasse. *Der Spiegel*, 25.05.1975. https://www.spiegel.de/politik/abitur-einbahnstrasse-wird-zur-sackgasse-a-48d8685a-0002-0001-0000-000041521031

Paulsen, F. (1902). *Die deutschen Universitäten und das Universitätsstudium*. Berlin: Asher.

Pischel, F. (1912). Die Frage einer Neuordnung der Reifeprüfung. In *Jahrbuch des Vereins für Wissenschaftliche Pädagogik*, S. 154–187.

Reglement vom 23.12.1788 für die Prüfung an den Gelehrten Schulen. http://resolver.staatsbibliothek-berlin.de/SBB0002BEE700000000

Reglement vom 04.06.1834 für die Prüfung der zu den Universitäten übergehenden Schüler. In *J.F.W. Koch, Die preussischen Universitäten, eine Sammlung der Verordnungen welche die Verfassung und Verwaltung dieser Anstalten betreffen* (2 Bde.). https://play.google.com/store/books/details?id=xuoDDj3gp8cC&rdid=book-xuoDDj3gp8cC&rdot=1

Reh, S. (2017). Die Ambivalenz der Rede über die „Sache" des Unterrichts. Beobachtungen zur Korrektur von deutschen Abituraufsätzen aus den 1950er Jahren. In C. Thompson, N. Ricken & R. Casale (Hrsg.), *Die Sache(n) der Bildung* (S. 107–125). Paderborn: Schöningh.

Reh, S., Bühler, P., Hofmann, M. & Moser, V. (2021). Einleitung. Prüfen, testen Auslesen und Zuweisen. Zum Inklusions-Paradox des Schulsystems. In dies. (Hrsg.), *Schülerauslese, Schulische Beurteilung und Schülertest 1880–1989* (S. 7–28). Bad Heilbrunn: Klinkhardt.

Reh, S. & Eiben–Zach, B. (2021). Das Bewerten von Literatur. Literarische Normen im fachdidaktischen Diskurs und in Abituraufsätzen der 1960er Jahre. In L. Brenz & T. Pflugmacher (Hrsg.), *Normativität literarischen Verstehens. Interdisziplinäre Beiträge zur Theorie und Praxis eines zentralen Problems* (S. 175–196). Frankfurt a. M.: Peter Lang.

Reh, S., Kämper-van den Boogaart, M. & Scholz, J. (2017). Eine lange Geschichte: Der deutsche Abituraufsatz als „Gesammtbildung der Examinanden". Prüfungspraxis und Lehrerkommentare von Abituraufsätzen in den 1950er Jahren. *Zeitschrift für Pädagogik, 63*, 280–298.

Schlemmer, H. (1930). Abschaffung der Reifeprüfung? *Deutsches Philologen-Blatt* 30/31, 449–450.

Scholz, J. (2021). „In zweifelhaften Fällen mag der Geist der Milde den Ausschlag geben" – Korrektur und Benotung des deutschen Aufsatzes in historischen Debatten und Praktiken. In S. Reh, P. Bühler, M. Hofmann & V. Moser (Hrsg.), *Schülerauslese, schulische Beurteilung und Schülertests 1880–1980* (S. 153–170). Bad Heilbrunn: Klinkhardt.

Scholz, J., Löwe, D., Engelhardt, K. v. & Reh, S. (2021). Normieren und Drohen. Der Exklusivitätscharakter der höheren Schulen im Deutschen Kaiserreich und in der Weimarer Republik. In M. Vogt, M.-A. Boger & P. Bühler (Hrsg), *Inklusion als Chiffre. Bildungshistorische Analysen und Reflexionen* (S. 205–217). Bad Heilbrunn: Klinkhardt.

Schott, E. (1930): Die neue württembergische Reifeprüfungsordnung. *Deutsches Philologen-Blatt, 38*(29), 166–167.

Schwartz, P. (1910). *Die Gelehrtenschulen Preußens unter dem Oberschulkollegium (1787–1806) und das Abiturientenexamen* (Erster Band). Berlin: Weidmannsche Buchhandlung.

Selbmann, R. (1988). „Deutsch sein heißt arbeiten". Zeitgeist in Aufsatzthemen zwischen Kaiserreich und Drittem Reich, zugleich ein Beitrag zur Geschichte des Wilhelmsgymnasiums. In *Wilhelmsgymnasium München: Jahresbericht 1987/88*, S. 137–151. https://epub.ub.uni-muenchen.de/5225/1/5225.pdf

Simon, E. (1930). Der philologische Nachwuchs Deutschlands und die Besuchszahlen der höheren Lehranstalten im Lichte der deutschen Hochschulstatistik. *Deutsches Philologen-Blatt, 38*(23/24), 353–356.

Spranger, E. (1910). *Wilhelm von Humboldt*. Berlin: Reuther & Reichard.

Steffenhagen, A. (1851). Sind Schulen Vorbereitungsanstalten? *Pädagogische Revue* 27, 116–126.

Steinmetz, M. (2013). *Der überforderte Abiturient im Fach Deutsch. Eine qualitativ–empirische Studie zur Realisierbarkeit von Bildungsstandards*. Wiesbaden: Springer.

Stelmaszyk, B. (2002). *Rekonstruktionen von Bildungsgängen preußischer Gymnasiasten sowie der zugehörigen Lehrergutachten aus Reifeprüfungsverfahren der Jahre 1926–1946*. Habilitationsschrift – Johannes Gutenberg-Universität Mainz.

Tenorth, H.–E. (2018). *Wilhelm von Humboldt: Bildungspolitik und Universitätsreform*. Paderborn: Schöningh.

Teuber (1829). Ueber Abiturienten-Prüfungen. *Allgemeine Schulzeitung* (25. November), 1145–1151.

Thies, B. (1928a). Das Stoffproblem des Erlebnisaufsatzes und die Aufsatzverfügung des Berliner Provinzialschulkollegiums. *Deutsches Philologen-Blatt, 36*(16), 226–229

Thies, B. (1928b). Das Stoffproblem des Erlebnisaufsatzes und die Aufsatzverfügung des Berliner Provinzialschulkollegiums. *Deutsches Philologen-Blatt, 36*(17), 241–243.

Titze, H. (1981). Überfüllungskrisen in akademischen Karrieren: eine Zyklustheorie. *Zeitschrift für Pädagogik, 27*, 187–224.

Titze, H. (1987). *Datenhandbuch zur deutschen Bildungsgeschichte. Das Hochschulstudium in Preußen und Deutschland 1820–1944* (unter Mitarbeit von H.–G. Herrlitz, V. Müller-Benedict & A. Nath). Göttingen: Vandenhoeck & Ruprecht.

Wendt, G. (1896). *Didaktik und Methodik des deutschen Unterrichts und der philosophischen Propädeutik*. München: Beck.

Weßner, P. (1930). Zur Frage der Auslese auf den höheren Schulen. *Deutsches Philologen-Blatt* 30/31, 450–451.

Wettberg, K. (1914). Aufsatzinstitut und Schule. *Deutsches Philologen-Blatt, 22*(22), 425–426.

Wolter, A. (1989). *Von der Elitenbildung zur Bildungsexpansion. Zweihundert Jahre Abitur.* (1788–1988). Oldenburger Universitätsreden Nr. 28. Bibliotheks- und Informationssystem der Universität Oldenburg.

Wolter, A. (2016). Gymnasium und Abitur als „Königsweg" des Hochschulzugangs: Historische Entwicklungslinien und institutionelle Transformationen. In J. Kramer, M. Neumann & U. Trautwein (Hrsg.), *Abitur und Matura im Wandel. Historische Entwicklungslinien, aktuelle Reformen und ihre Effekte* (S. 1–28). Wiesbaden: Springer.

Zach, B. & Reh, S. (2018). Abituraufgaben in der späten Weimarer Republik zwischen Normierung der Aufgabenbearbeitung und dem Anspruch nach „Selbstständigkeit". *Didaktik Deutsch, 23*(44), 44–60.

8 Wie vergleichbar sind die Bewertungen von Abiturarbeiten im Fach Deutsch? Empirische Studien zu verschiedenen Bewertungsmodellen

Pauline Schröter, Hannelore Söldner, Lars Hoffmann,
Anja Riemenschneider, Jörg Jost & Dorothee Wieser

Zusammenfassung

Die Vergleichbarkeit der Bewertungen von Abiturarbeiten und die Gestaltung der Erwartungshorizonte stehen seit Jahren im Fokus fachdidaktischer Diskussionen im Fach Deutsch. In den Ländern der Bundesrepublik Deutschland finden unterschiedliche Bewertungsmodelle Anwendung, die neben stark analytischen und holistischen Modellen auch viele Mischformen einschließen, die sich vornehmlich in der Differenzierung und Gewichtung von Bewertungskriterien unterscheiden. Der vorliegende Beitrag präsentiert zwei experimentelle Studien, die vom IQB durchgeführt wurden, um die Entwicklung einer ländergemeinsamen Regelung zur Bewertung von Abiturarbeiten im Fach Deutsch durch eine empirische Grundlage zu unterstützen. In der ersten Studie wurden Erkenntnisse über die Vor- und Nachteile des holistischen sowie des analytischen Bewertungsmodells gewonnen. In einer Folgestudie wurde untersucht, welche Auswirkungen die Vorgabe unterschiedlich differenzierter Kriterien und deren Gewichtung auf die Bewertung von Abiturarbeiten haben. Die Ergebnisse beider Studien deuten darauf hin, dass es unabhängig vom implementierten Bewertungsmodell dringend geboten ist, die Objektivität der Bewertungen von Abiturarbeiten durch geeignete Maßnahmen zu erhöhen.

1 Einleitung

Mit dem Beschluss der Kultusministerkonferenz (KMK) im Jahr 2012, für die Fächer Deutsch und Mathematik sowie die fortgeführten Fremdsprachen Englisch und Französisch auf Basis der Bildungsstandards für die allgemeine Hochschulreife ländergemeinsame Abituraufgabenpools zu entwickeln (vgl. Beitrag 2 von Hoffmann, Schröter & Stanat in diesem Band), hat ein Annäherungsprozess begonnen, der unter anderem zum Ziel hat, die mit den Abiturprüfungen der Länder der Bundesrepublik Deutschland (im Folgenden: Länder) verbundenen Anforderungen sukzessive anzugleichen. Um die Vergleichbarkeit von Prüfungsergebnissen sicherzustellen, sollen neben den Rahmenbedingungen auch die Gestaltung der Prüfungsaufgaben und die

Bewertung der Prüfungsarbeiten zwischen den Ländern vereinheitlicht werden. Für das Fach Deutsch gibt es bislang keine ländergemeinsame Regelung zur Bewertung. Stattdessen finden in den Ländern unterschiedliche Bewertungsmodelle Anwendung. In einigen Ländern erfolgt die Bewertung von Abiturarbeiten anhand eines holistischen Bewertungsverfahrens, das in einem globalen, synthetischen Urteil resultiert. In anderen Ländern ist hingegen eine analytische Bewertung von Prüfungsarbeiten vorgesehen, bei der verschiedene Teilleistungen jeweils separat beurteilt und gewichtet werden. Die in den Ländern angewendeten Bewertungsmodelle sind jedoch nicht immer eindeutig als analytisch oder holistisch zu klassifizieren. Sie lassen sich vielmehr auf einem Kontinuum abbilden, das neben den beiden Polen auch Mischformen einschließt. Alle Bewertungsmodelle geben Kriterien zur Bewertung der Abiturarbeiten vor, unterscheiden sich jedoch darin, wie differenziert diese Kriterien ausgewiesen werden sowie ob und in welcher Form Teilleistungen zu gewichten sind. Um im Rahmen des Annäherungsprozesses der Länder die Entwicklung einer gemeinsamen Regelung zur Bewertung von Abiturarbeiten im Fach Deutsch zu unterstützen, wurden die Vor- und Nachteile verschiedener Bewertungsmodelle empirisch untersucht. Die Ausgangssituation sowie die beiden experimentellen Studien, die dazu vom Institut zur Qualitätsentwicklung im Bildungswesen (IQB) durchgeführt wurden, werden im Folgenden beschrieben.

2 Ausgangssituation

2.1 Modelle zur Bewertung von Abiturarbeiten

Gemäß den Bildungsstandards im Fach Deutsch für die allgemeine Hochschulreife (KMK, 2012) wird jeder Prüfungsaufgabe eine Beschreibung der erwarteten Leistungen einschließlich der Angabe von Bewertungskriterien beigefügt. Allen Bewertungsmodellen, die in den Ländern zur Korrektur und Bewertung von Abiturarbeiten im Fach Deutsch implementiert sind, ist gemeinsam, dass sie für die Bearbeitung relevante Wissensbestände abbilden und Kriterien enthalten, anhand derer die Leistung der Prüflinge in Bezug auf den Inhalt (auch: Verstehensleistung) und die Sprache (auch: Darstellungsleistung) beurteilt werden soll. Dies wird in Form von Erwartungshorizonten und Bewertungshinweisen (im Folgenden: EWH) realisiert, die den Lehrkräften zur Verfügung gestellt werden. Die Bildungsstandards benennen allgemeine Kriterien, auf deren Grundlage die Bewertung von Abiturarbeiten erfolgen soll, machen aber keine Vorgaben zur konkreten Ausgestaltung der EWH.

Der wesentliche Unterschied zwischen analytischen und holistischen Bewertungsmodellen besteht in der Art und Weise, wie die in den EWH aufgeführten Kriterien bei der Gesamtbewertung zu berücksichtigen und zu gewichten sind. So sind die Lehrkräfte bei den analytischen EWH angehalten, jedes einzelne Kriterium zunächst separat zu bewerten. Dies erfolgt in der Regel durch die Erteilung von Punkten. Die Anzahl der maximal erreichbaren Punkte variiert dabei je nach Kriterium, wodurch vorgegeben ist, wie die einzelnen Kriterien zu gewichten sind. Im Rahmen der Ge-

samtbewertung werden die bei den einzelnen Kriterien erreichten Punkte addiert und anhand einer Notentabelle der entsprechenden Notenstufe zugeordnet. Im Unterschied dazu haben die Lehrkräfte in Ländern mit holistischen Bewertungsmodellen mehr Entscheidungsspielräume bei der Bewertung, da die entsprechenden EWH keine quantifizierten Vorgaben zur Gewichtung und Synthese von Kriterien enthalten. Vielmehr obliegt es der jeweiligen Lehrkraft, mittels ihrer Fachexpertise und auf der Grundlage der im EWH genannten Kriterien eine Gesamtbeurteilung der Prüfungsarbeit vorzunehmen. Die erteilte Notenpunktzahl wird in der Regel in Form eines Wortgutachtens begründet.

Zwischen den Polen *analytisch* und *holistisch* lassen sich Mischformen finden, die sich sowohl im Grad der Differenzierung der Bewertungskriterien als auch in Bezug auf deren Gewichtung unterscheiden. Eine solche Mischform bilden beispielsweise die EWH ab, die zu den Prüfungsaufgaben des gemeinsamen Abituraufgabenpools der Länder im Fach Deutsch entwickelt werden. Neben detaillierten Erwartungen an die inhaltliche Leistung werden hier fünf standardbasierte Kriterien für die Bewertung der sprachlichen Leistung vorgegeben: „Aufgabenbezug, Textsortenpassung und Textaufbau", „Fachsprache", „Umgang mit Bezugstexten und Materialien", „Ausdruck und Stil" sowie „standardsprachliche Normen". Für jede Aufgabe wird eine Gewichtung der Verstehensleistung und der Darstellungsleistung durch eine mit dem Zusatz „ca." versehene Prozentangabe festgelegt. Darüber hinaus ist eine Gewichtung der Teilaufgaben gleichfalls durch mit „ca." versehene Prozentangaben vorgesehen. Die Ermittlung der Notenpunkte erfolgt im Sinne einer holistischen Bewertung mit den hierbei üblichen Entscheidungsspielräumen.

2.2 Forschung zu Bewertungsmodellen

Historisch ist die Entwicklung analytischer Modelle zur Beurteilung von Schreibprodukten eng mit der Entstehung der Psychometrie zu Beginn des 20. Jahrhunderts verbunden, die sich insbesondere im angloamerikanischen Raum sehr schnell etablierte (Hamp-Lyons, 2016). Bereits in den 1920er-Jahren wurde die Komplexität von Wortschatz und Satzbau in den von Studienbewerberinnen und -bewerbern verfassten Essays als Indikator genutzt, um ihre Eignung für die Universität zu beurteilen. In der Folge etablierten sich analytische Bewertungsmodelle an den Universitäten, da angenommen wurde, dass sie angesichts stetig wachsender Studierendenzahlen eine transparente wie auch objektive, reliable und valide Auswahl von Bewerberinnen und Bewerbern ermöglichen. Parallel dazu fanden analytische Bewertungssystematiken in zunehmendem Maße auch in der schulischen Praxis in den USA Verbreitung (Elliott, 2005; Elliot & Haswell, 2019).

In der internationalen Forschungsliteratur wurde die Angemessenheit des analytischen Vorgehens für die Beurteilung von schulischen Schreibprodukten immer wieder kritisch diskutiert. Insbesondere wurde problematisiert, analytische Bewertungen würden der Komplexität von Schreibprodukten nicht gerecht, da die einzelnen Qualitätsaspekte nur separat und nicht in ihrem Zusammenspiel betrachtet würden (Sadler, 2009). Daher kam es in den 1970er-Jahren auch in den angloamerikanischen Län-

dern zu einer stärkeren Verbreitung von holistischen Bewertungsansätzen (Hamp-Lyons, 2016; Elliot & Haswell, 2019).

Eine mit den USA vergleichbare analytische Bewertungstradition und eine entsprechende Forschungslage finden sich für Deutschland nicht. In der hiesigen Forschungsliteratur werden analytische Bewertungsmodelle insbesondere im Kontext von Schulleistungsstudien thematisiert (Hartmann, 1989; Lehmann, 1988; Neumann, 2007a, 2007b). Als in der Deutschdidaktik einflussreich gelten hierbei vor allem die im Zuge der DESI-Studie entwickelten analytischen Kriterien zur Beurteilung der Textqualität von Briefen (Neumann, 2007a).

Holistische Urteile gelten häufig als weniger reliabel als analytische Bewertungen (z. B. Jonsson & Svingby, 2007; Kreuzer, 2018). Allerdings ist die empirische Befundlage hierzu nicht so eindeutig, wie oftmals angenommen wird (z. B. Barkaoui, 2011; Harsch & Martin, 2013; Klein et al., 1998; Lindauer & Sommer, 2018; Rezaei & Lovorn, 2010). Zwar wurde vielfach gezeigt, dass sich mit analytischen Kriterienrastern eine hohe Beurteilerübereinstimmung bei der Textbewertung erzielen lässt (z. B. East, 2009; Neumann, 2007a, 2007b; Jonsson & Svingby, 2007; Silvestri & Oescher, 2006; Zlatkin-Troitschanskaia et al., 2019), es wurden gleichzeitig aber auch Forschungsarbeiten vorgelegt, die hohe Reliabilitäten für holistische Bewertungssystematiken gefunden haben (z. B. Bramley, 2007; Grzesik & Fischer, 1984; Lindauer & Sommer, 2018). Die wenigen Studien, in denen die Reliabilität analytischer und holistischer Bewertungssystematiken anhand desselben Textkorpus ermittelt wurde, konnten bestenfalls geringe Vorteile zugunsten von analytischen Kriterienrastern nachweisen (z. B. Barkaoui, 2011; Van den Bergh et al., 2012; Bouwer et al., 2016). Insgesamt legt der aktuelle Forschungsstand den Schluss nahe, dass weniger die Wahl eines analytischen oder holistischen Bewertungsmodells als vielmehr andere Faktoren – wie etwa eine angemessene Schulung der Bewerter:innen oder die Verwendung von Benchmarktexten – maßgeblicher sein dürften, um reliable Bewertungen zu gewährleisten (z. B. Fahim & Bijani, 2011; Jonsson & Svingby, 2007; Rezaei & Lovorn, 2010).

Nicht selten wird außerdem angenommen, analytische Bewertungen von Schreibprodukten würden im Vergleich zu holistischen Urteilen im Mittel etwas strenger ausfallen. Als Grund hierfür wird insbesondere genannt, analytische Kriterienraster würden vor allem auf die Defizite der jeweils zu bewertenden Schreibprodukte fokussieren, wohingegen holistische Modelle wegen der größeren Entscheidungsspielräume ein kompensatorisches Vorgehen befördern würden, bei dem Minderleistungen bei einzelnen Kriterien durch sehr gute Leistungen bei anderen Kriterien ausgeglichen werden könnten. Die empirische Befundlage hierzu ist allerdings ebenfalls gemischt (z. B. Jonsson & Svingby, 2007; Barkaoui, 2011).

Im Vergleich zu holistischen Urteilen scheinen analytische Urteile jedoch in etwas geringerem Maße anfällig für Urteilsverzerrungen zu sein. Urteilsverzerrungen treten dann auf, wenn die Wahrnehmung und Interpretation eines zu beurteilenden Gegenstands unbewusst von für die Beurteilung irrelevanten Faktoren beeinflusst wird. Ein bekanntes Beispiel für Urteilsverzerrungen bei der Bewertung von Texten ist der sogenannte Halo-Effekt (Thorndike, 1920), bei dem sich auffällige Oberflächen-

merkmale (wie zum Beispiel die Handschrift oder der Vorname der Verfasser:innen, sprachliche Normverletzungen durch fehlerhafte Rechtschreibung, grammatikalische oder Zeichensetzungsfehler) auf die inhaltliche Bewertung des Textes auswirken (Cumming et al., 2002; Rezaei & Lovorn, 2010). Weitere Beispiele für Urteilsverzerrungen sind Positions- und Referenzgruppeneffekte, die daraus resultieren, dass die Bewertung des Textes eines Prüflings nicht vollkommen unabhängig von der Bewertung der Texte anderer Prüflinge erfolgt (Trautwein & Baeriswyl, 2007). Je nach Leistungsspektrum der jeweiligen Gruppe, die als soziale Bezugsnorm bei der Leistungsbeurteilung herangezogen wird, können selbst Texte von durchschnittlicher Qualität hervorstechen und damit potenziell unter- oder überschätzt werden. Nicht zuletzt treten Verzerrungen bei der Bewertung von Texten auch infolge individueller Faktoren seitens der Bewerter:innen auf (Lumley, 2002; Schoonen, 2005; Shabani & Panahi, 2020). Automatisierte Denkweisen, auf die Bewerter:innen bei der Bildung eines Urteils unbewusst zurückgreifen, werden als Urteilsheuristiken bezeichnet (Tversky & Kahneman, 1974). Diese basieren auf subjektiven Erfahrungen und treten im Bewertungsprozess in Gestalt von Faustregeln und als Bauchgefühl auf. In vergleichenden Studien zur Beurteilung von Schülerarbeiten wurde eine geringere Anfälligkeit für Urteilsverzerrungen im Rahmen der analytischen Bewertung festgestellt (z. B. Böhme et al., 2009). Dieses Befundmuster ist konsistent mit den Ergebnissen von Untersuchungen aus anderen Domänen (wie etwa der klinischen Diagnostik in der Medizin), die ebenfalls darauf schließen lassen, dass kriteriale Urteile im Vergleich zu ganzheitlichen Urteilen weniger anfällig für Urteilsverzerrungen sind (Meehl, 1954; Grove et al., 2000).

2.3 Übertragbarkeit der Befunde auf die Bewertung von Abiturarbeiten

Inwiefern sich die internationale Befundlage auf die Modelle zur Bewertung von Abiturarbeiten im Fach Deutsch übertragen lässt, ist unklar. Während in den USA die analytische Bewertungstradition über Jahrzehnte hinweg gewachsen ist, haben in Deutschland erst in den letzten 15 Jahren einzelne Länder analytische Kriterienraster für die Bewertung von Abiturarbeiten eingeführt. Diese sind jedoch in ihrer Konzeption nur bedingt vergleichbar mit den Bewertungsrastern, die in den vorab erwähnten Studien untersucht wurden. Während bei Letzteren die Bewertung von Schreibprodukten anhand einer Vielzahl recht differenzierter (nicht selten sogar dichotomer) Kriterien erfolgt, sind die Kriterien der von einzelnen Ländern für die Bewertung von Abiturarbeiten verwendeten Bewertungsraster deutlich weniger differenziert. In der Regel müssen Lehrkräfte hier keine dichotomen Einzelaspekte beurteilen. Vielmehr bündeln die Kriterien der Bewertungsraster mehrere Einzelanforderungen, die dann gemeinsam einzuschätzen sind.

Auch die Notenurteile, die in einer Vielzahl von Ländern auf einer holistischen Bewertung basieren, sind kaum mit den holistischen Urteilen vergleichbar, die in den oben erwähnten Studien erhoben wurden. Während bei Letzteren die Bewertung in der Regel anhand eines Gesamteindrucks erfolgt, bewerten Lehrkräfte in Ländern mit holistischem Bewertungsmodell die Abiturarbeiten auf der Grundlage vorab festgelegter Kriterien und begründen ihr Notenurteil in einem Verbalgutachten.

Weitere Zweifel an der Übertragbarkeit bisheriger Befunde erwachsen aus dem Umstand, dass sich die in den Abiturarbeiten im Fach Deutsch verwendeten Textformen hinsichtlich ihres Umfangs und ihrer Komplexität deutlich von den Textformen abheben, die in den erwähnten Studien zu beurteilen waren. Während in der schriftlichen Abiturprüfung im Fach Deutsch textbezogene und materialgestützte Aufgaben zu fachspezifischen Domänen bearbeitet werden, bezieht sich die Forschungsliteratur überwiegend auf Schreibprodukte zu Aufgabenstellungen, deren Fokus auf gestalterischem oder fiktionalem Schreiben liegt. Dementsprechend bezieht sich die Textbewertung hierbei vor allem auf Aspekte der Darstellungsleistung, wobei die eingesetzten Bewertungsmodelle häufig nicht spezifisch für den jeweiligen Schreibauftrag entwickelt wurden, sondern generischen Charakter haben. Bei der Beurteilung von Abiturarbeiten liegt der Schwerpunkt der Bewertung im Vergleich dazu auf der inhaltlichen Leistung. Da diese jedoch nicht gänzlich getrennt von der sprachlichen Leistung bewertet werden kann, ist eine Konfundierung von Verstehens- und Darstellungsleistung bei der Bewertung nicht auszuschließen (Disdorn-Liesen, 2016).

Eine weitere Herausforderung bei der Bewertung von Abiturarbeiten besteht darin, dass auch andere als im EWH ausgeführte Lösungen bei der Bewertung der Prüfungsleistung als gleichwertig gewürdigt werden sollen, wenn sie der Aufgabenstellung entsprechen, sachlich richtig und nachvollziehbar sind. Zum einen ist die große Spanne dessen zu nennen, was von den Bewerterinnen und Bewertern als sachlich richtig und nachvollziehbar eingeschätzt wird – vor allem bei der Aufgabenart „Interpretation literarischer Texte“. Die EWH können hier auch nur sehr bedingt eine Orientierung bieten, da z. B. die Darstellung aller potenziell durch den Text zu stützenden Lesarten und Deutungen nicht möglich ist. Zum anderen ist zu berücksichtigen, dass in der Tradition des Abituraufsatzes im Fach Deutsch dem Kriterium der Eigenständigkeit und Kreativität der Schülerleistung ein hoher Stellenwert beigemessen wird (vgl. Beitrag 7 von Kämper-van den Boogaart & Reh in diesem Band). Dies spiegelt sich auch in den Bildungsstandards für die Allgemeine Hochschulreife, in denen die Selbstständigkeit der Prüfungsleistung betont wird. Insofern steht die Orientierung an einem EWH mit verbindlichen Kriterien, die eine möglichst objektive Bewertung ermöglichen sollen, in einer gewissen Spannung zu der Erwartung von kreativen und eigenständigen Lösungen.

2.4 Anlass der Studien

Die Leistungsbeurteilung im Abitur sowie die Gestaltung der EWH im Fach Deutsch stehen seit Jahren im Fokus fachdidaktischer Diskussionen (z. B. Köster, 2006; Zabka & Stark, 2010; Disdorn-Liesen, 2016). Bedingt durch die Herausforderungen bei der Datenerhebung, die sich für empirische Studien zur Leistungsbeurteilung von Abschlussprüfungen ergeben, war die Bewertung von Abiturarbeiten im Fach Deutsch bislang noch nicht Gegenstand quantitativer Untersuchungen. Auch zur Bewertung von Schülerarbeiten in anderen Jahrgangsstufen liegen für Deutschland nur wenige Ergebnisse vor. In einer explorativen Studie zur Beurteilung von Aufsätzen aus den Jahrgangsstufen 8, 9 und 10 untersuchten Grzesik und Fischer (1984) die Anwendung von Kriterienrastern im Vergleich zu einem globalen Ersteindruck. Im Rahmen der

Untersuchung kamen geübte Bewerter:innen zu einer höheren Übereinstimmung, wenn sie die Aufsätze aufgrund des ersten Gesamteindrucks bewerteten, was darauf zurückgeführt wurde, dass sie bei der Bewertung unbewusst ihre aus langjähriger eigener Erfahrung gewonnenen Bewertungskriterien zugrunde gelegt hatten. In der Studie von Birkel und Birkel (2002) wurde für die Grundschule untersucht, inwieweit Deutschlehrkräfte bei einer holistischen Bewertung von Aufsätzen zu übereinstimmenden Urteilen gelangen. Im Ergebnis der Untersuchung zeigte sich, dass die an der Studie teilnehmenden Lehrkräfte oftmals sehr diskrepant bewerteten. Dabei streuten die Notenurteile für ein- und denselben Aufsatz in Einzelfällen über bis zu vier Notenstufen. Überaus zufriedenstellende Übereinstimmungen der Beurteilungen wurden hingegen – allerdings erst nach intensiver Schulung der Bewerter:innen – für die Hamburger Aufsatzstudie berichtet, die als Teil einer internationalen Studie zu Schreibfähigkeiten von Schüler:innen der elften Jahrgangsstufe durchgeführt wurde (Hartmann, 1989; Lehmann, 1988). Darüber hinaus haben Studien zur Notengebung gezeigt, dass die Aufsatzbewertung im Fach Deutsch sehr anfällig für Urteilsverzerrungen ist. Neben Halo-Effekten (Böhme et al., 2009) sind dabei auch Positions- und Referenzgruppeneffekte dokumentiert (Ingenkamp, 1995).

Bisherige Forschungsergebnisse zum Vergleich von analytischen und holistischen Urteilen lassen sich demnach nur bedingt auf den deutschen Abiturprüfungskontext übertragen. Auch liegen bislang keine spezifischen Befunde zur Objektivität der Bewertung von Abiturarbeiten im Fach Deutsch vor. Folglich gibt es keine evidenzbasierten Erkenntnisse darüber, wie Vorgaben zu Kriterien und Gewichtung die Bewertung von Abiturarbeiten beeinflussen. Im Rahmen des Annäherungsprozesses zwischen den Ländern wurde deshalb entschieden, die Entwicklung einer ländergemeinsamen Regelung zur Bewertung von Abiturarbeiten im Fach Deutsch durch eine empirische Grundlage zu unterstützen. Zu diesem Zweck hat das IQB zwei experimentelle Studien mit den folgenden Untersuchungsschwerpunkten durchgeführt:

1. **Bewertung der Abiturarbeiten**
   - Zu welchen Notenurteilen kommen Lehrkräfte bei der Bewertung von Abiturarbeiten?
   - Ergeben sich Unterschiede bei der Bewertung in Abhängigkeit vom verwendeten EWH?
2. **Übereinstimmung der Bewertungen**
   - Inwieweit stimmen Lehrkräfte in ihren Bewertungen überein?
   - Unterscheidet sich die Übereinstimmung der Bewertungen in Abhängigkeit vom verwendeten EWH?
3. **Anfälligkeit für Urteilsverzerrungen**
   - Welcher EWH ist anfälliger für Urteilsverzerrungen?
4. **Nutzung und Akzeptanz der EWH**
   - Wie verwenden Lehrkräfte die EWH?
   - Wie schätzen Lehrkräfte den Nutzen der EWH ein?
   - Wie sollten EWH aus der Sicht von Lehrkräften gestaltet sein?

Studie 1 zielte darauf ab, empirische Erkenntnisse zu den Vor- und Nachteilen des holistischen sowie des analytischen Bewertungsmodells im spezifischen Kontext von Abiturprüfungen im Fach Deutsch zu gewinnen. Der Fokus lag hierbei insbesondere auf der Frage, welches der beiden Bewertungsmodelle objektivere Bewertungen ermöglicht. Hierfür wurden im Jahr 2018 Deutschlehrkräfte aus zwei Ländern im Rahmen einer experimentellen Studie gebeten, Abiturarbeiten aus dem Prüfungsjahr 2017 auf der Grundlage von holistisch oder analytisch ausgerichteten EWH zu beurteilen.

Aufbauend auf den Ergebnissen dieser Studie ging Studie 2 der Frage nach, welche Auswirkungen die Vorgabe unterschiedlich differenzierter Kriterien und deren Gewichtung auf die Bewertung von Abiturarbeiten haben. Neben eindeutig holistischen und analytischen EWH wurden hier auch Mischformen untersucht. Dazu wurden im Jahr 2020 zu einer Abituraufgabe fünf verschiedene EWH-Varianten entwickelt, auf deren Grundlage erfahrene Deutschlehrkräfte aus mehreren Ländern Abiturarbeiten aus dem Prüfungsjahr 2017 bewerteten.

3 Studie 1

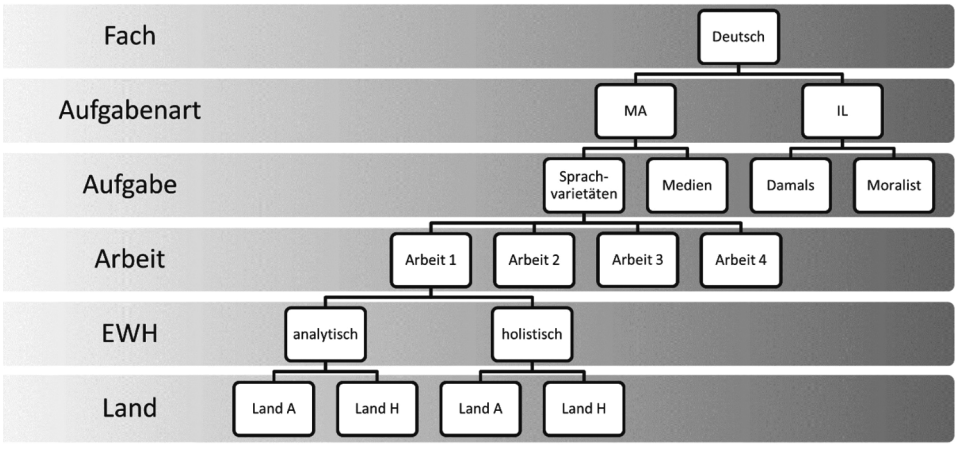
3.1 Zielstellung

Auch wenn sich die in den Abiturprüfungen der Länder eingesetzten EWH in vielen Aspekten unterscheiden, so lassen sie sich vor dem Hintergrund der Bewertungstraditionen der Länder meist in eher analytische und eher holistische Bewertungsmodelle differenzieren. Lehrkräfte aus dem Land A[1], das im Vergleich zu anderen Ländern ein ausgeprägt analytisches Vorgehen bei der Bewertung von Abiturarbeiten vorsieht, sind dazu angehalten, für den Grad der Erfüllung jedes einzelnen Kriteriums Punkte zu vergeben. Im Rahmen der Gesamtbewertung werden die bei den einzelnen Kriterien erreichten Punkte addiert und anhand einer Tabelle der entsprechenden Notenstufe zugeordnet. Demgegenüber erhalten Lehrkräfte aus dem Land H[1], in dem – wie in einigen anderen Ländern – eine ausgeprägt holistische Vorgehensweise bei der Bewertung von Abiturarbeiten praktiziert wird, keine Vorgaben zur Gewichtung von Kriterien. Die für die Gesamtleistung erteilte Notenpunktzahl wird hier in Form eines Wortgutachtens begründet. Studie 1 hatte zum Ziel, diese beiden Herangehensweisen bei der Bewertung miteinander zu vergleichen und Vor- und Nachteile zu identifizieren. Untersucht wurden dabei Unterschiede zwischen analytischen und holistischen EWH in Bezug auf die erteilten Notenpunkte, die Übereinstimmung der Bewertungen, die Anfälligkeit für Urteilsverzerrungen sowie die Nutzung und Akzeptanz durch die Lehrkräfte.

1 Zur Wahrung ihrer Anonymität werden die teilnehmenden Länder mit „Land A“ und „Land H“ bezeichnet.

3.2 Methode

Zur Untersuchung der Fragestellungen wurde ein experimentelles Between-Subjects-Design gewählt, das in Abbildung 1 schematisch dargestellt ist. Über die Bezirksregierungen der Länder A und H wurden 33 erfahrene Deutsch-Lehrkräfte rekrutiert (17 Lehrkräfte aus Land A und 16 Lehrkräfte aus Land H). Alle Lehrkräfte verfügten über Erfahrung mit der Bewertung von Abiturarbeiten im Fach Deutsch, nahmen freiwillig an der Studie teil und erhielten vom IQB eine Aufwandsentschädigung. Jede Lehrkraft wurde gebeten, vier Abiturarbeiten zu zwei verschiedenen Aufgaben zu bewerten (insgesamt also acht Arbeiten), darunter jeweils eine Aufgabe zur Aufgabenart „Interpretation literarischer Texte“ (IL) und eine Aufgabe zur Aufgabenart „materialgestütztes Verfassen argumentierender Texte“ (MA). Aufgrund der Vielzahl an durch die Bildungsstandards vorgegebenen Aufgabenarten im Fach Deutsch war es erforderlich, zur Erhöhung der Generalisierbarkeit der Ergebnisse mehr als ein Aufgabenformat in die Studie einzubeziehen. Die Auswahl dieser beiden Aufgabenarten ist darauf zurückzuführen, dass die Interpretation literarischer Texte eine in allen Ländern bereits seit Langem etablierte Aufgabenart darstellt, während das materialgestützte Verfassen von informierenden und argumentierenden Texten in den meisten Ländern erst mit der Verabschiedung der Bildungsstandards für die Allgemeine Hochschulreife im Jahr 2012 und deren sukzessiver Implementation in den Lehrplänen der Länder Bestandteil des Unterrichts in der Qualifikationsphase und der Abiturprüfung im Fach Deutsch wurde. Hierdurch konnten Unterschiede in der Bewertung, die gegebenenfalls auf den unterschiedlichen Grad an Erfahrung der Lehrkräfte mit diesen Aufgabenarten zurückgehen, in die Betrachtung der Effekte einbezogen werden.



Anmerkung: Die Ebenen „Arbeit“, „EWH“ und „Land“ sind hier aus Platzgründen nur für die Aufgabe „Sprachvarietäten“ dargestellt, beziehen sich aber gleichermaßen auf alle Aufgaben.

Abbildung 1: Schematische Darstellung des Forschungsdesigns von Studie 1

Zu jeder der beiden Aufgabenarten wurden zwei Aufgaben aus dem ländergemeinsamen Pool für das Prüfungsjahr 2017 ausgewählt: die IL-Aufgaben mit den Kurzbezeichnungen „Damals“ und „Moralist“ sowie die MA-Aufgaben „Sprachvarietäten“ und „Medien“. Diese Aufgaben waren weder in Land A noch in Land H in der Abiturprüfung zum Einsatz gekommen, sodass keine der an der Studie beteiligten Lehrkräfte über Korrekturerfahrungen mit diesen Aufgaben verfügte. Jeweils die Hälfte der Lehrkräfte aus Land A und Land H erhielt als Grundlage für die Bewertung der Arbeiten einen analytischen EWH, die andere Hälfte verwendete einen holistischen EWH. Insgesamt wurden alle Arbeiten von 17 Lehrkräften nach analytischem und von 16 Lehrkräften nach holistischem Modell bewertet. Die Zuweisung der Lehrkräfte zu den Aufgaben und Bewertungsbedingungen erfolgte zufällig.

Stichprobe

Die Gesamtstichprobe setzte sich aus 14 weiblichen und 19 männlichen Lehrkräften zusammen, wobei sich die Geschlechterverteilung nicht statistisch signifikant zwischen den Ländern ($\chi 2(1) < .01$, $p = 1$) oder Bewertungsbedingungen ($\chi 2(1) = .25$, $p = .62$) unterschied. Die Teilnehmer:innen waren im Mittel 46 Jahre alt ($SD = 9$ Jahre) und verfügten über durchschnittlich 17 Jahre Unterrichtserfahrung ($SD = 9$ Jahre), die zwischen fünf und 38 Jahren variierte. Zwischen den Lehrkräften aus Land A und Land H bestanden weder hinsichtlich ihres Alters ($t(31) = .95$, $p = .35$) noch ihrer Unterrichtserfahrung ($t(30) = .78$, $p = .44$) statistisch signifikante Unterschiede. Auch zwischen den Gruppen der Bewertungsbedingungen unterschieden sich die Teilnehmer:innen weder in ihrem Alter ($t(31) = 1.31$, $p = .20$) noch in der Unterrichtserfahrung ($t(28) = 1.23$, $p = .21$) statistisch signifikant.

Material

Für die Studie wurden 16 authentische Abiturarbeiten aus dem Prüfungsjahr 2017 transkribiert[2], die dem IQB für die Evaluation der Poolaufgaben als anonymisierte Kopien vorliegen. Zu jeder der vier Aufgaben wurden vier Abiturarbeiten mit vergleichbarer Wörterzahl ausgewählt, die im Rahmen der Abiturprüfung mit 7 bis 9 Notenpunkten bewertet worden waren. Die Auswahl von ausschließlich aus dem mittleren Leistungsspektrum stammenden Arbeiten geht auf die Annahme zurück, dass in diesem Bereich im Vergleich zu sehr guten und weniger guten Arbeiten das größte Potenzial für Diskrepanzen in der Bewertung besteht und die Wahrscheinlichkeit für Decken- bzw. Bodeneffekte gering ist. Angaben zur Wörterzahl und zur ursprünglichen Bewertung der für die Studie ausgewählten Arbeiten befinden sich in Tabelle 1.

Für jede der vier Aufgaben wurde in Zusammenarbeit mit Expert:innen aus der Fachdidaktik eine holistische und eine analytische Version des EWH entwickelt, wobei die EWH landeseigener Abiturprüfungsaufgaben aus den Ländern A und H als Vorlagen herangezogen wurden. Beide Versionen unterschieden sich in Bezug auf die erwartete Verstehensleistung lediglich in ihrer Struktur, nicht jedoch in Bezug auf

2 Die Transkription der Arbeiten erfolgte unverändert, d. h. inklusive aller Fehler in Rechtschreibung, Grammatik und Zeichensetzung.

deren Inhalt. Die inhaltlichen Anforderungen wurden aus dem EWH der entsprechenden Poolaufgaben übernommen. Die holistische Version umfasste neben den inhaltlichen Erwartungen an die Prüfungsleistung die Bewertungskriterien für eine „gute" (11 Punkte) und eine „ausreichende" (5 Punkte) Aufgabenbearbeitung. Die analytische Version enthielt die Erwartungen an die Prüfungsleistung getrennt nach Verstehens- und Darstellungsleistung in tabellarischer Form und mit Angabe der maximal zu erreichenden Punkte für jedes Kriterium. Die Punkteverteilung entsprach dabei einer Gewichtung von Verstehens- und Darstellungsleistung im Verhältnis von ca. 70 % zu ca. 30 % bei IL-Aufgaben und ca. 60 % zu ca. 40 % bei MA-Aufgaben.

Tabelle 1: Wörterzahl und ursprüngliche Bewertung der ausgewählten Arbeiten

Aufgabe	Wörterzahl			Bewertung		
	M	*Min*	*Max*	*M*	*Min*	*Max*
Damals	1456	1261	1688	7.75	7	9
Moralist	1062	888	1323	7.00	7	7
Medien	1089	960	1168	7.75	7	9
Sprachvarietäten	812	782	860	7.25	7	8

Anmerkungen: Wörterzahl = mittlere Zahl der Wörter (ohne Überschriften) aller vier Arbeiten zu den einzelnen Aufgaben, Bewertung = ursprüngliche Bewertung der Abiturprüfungsarbeiten (in Notenpunkten), *M* = Mittelwert, *Min* = kleinster Wert, *Max* = größter Wert.

Instrumente

Die Bewertungen der Abiturarbeiten durch die Lehrkräfte erfolgten anhand eines Bewertungsbogens. In der holistischen Bewertungsbedingung wurden die Lehrkräfte gebeten, zur Bewertung in Form von Notenpunkten eine schriftliche Begründung in Form eines Kurzgutachtens zu verfassen. Darüber hinaus sollten sie angeben, wie viele Notenpunkte sie jeweils für die Verstehens- und Darstellungsleistung erteilt hätten, wenn diese Beurteilungsbereiche getrennt voneinander zu bewerten gewesen wären. Der Bewertungsbogen in der analytischen Bewertungsbedingung enthielt eine tabellarische Auflistung der Kriterien zur Vergabe der Punkte sowie eine Tabelle, um die insgesamt erreichten Punkte der entsprechenden Notenpunktzahl zuzuordnen. Die Lehrkräfte hatten in beiden Bewertungsbedingungen keine Kenntnis davon, wie die Arbeiten im Rahmen der Abiturprüfung bewertet worden waren.

Jede Lehrkraft wurde gebeten, im Anschluss an die Bewertung der Arbeiten einen Fragebogen auszufüllen. Dieser enthielt vornehmlich Multiple-Choice-Items, mit denen Angaben und Einschätzungen zu den in der Studie verwendeten EWH sowie Vorstellungen von EWH im Allgemeinen erfasst wurden. Zu jeder Fragestellung wurde eine Reihe von Aussagen präsentiert, die auf einer vierstufigen Likert-Skala (von „trifft gar nicht zu" bis „trifft voll zu") eingeschätzt werden sollten.

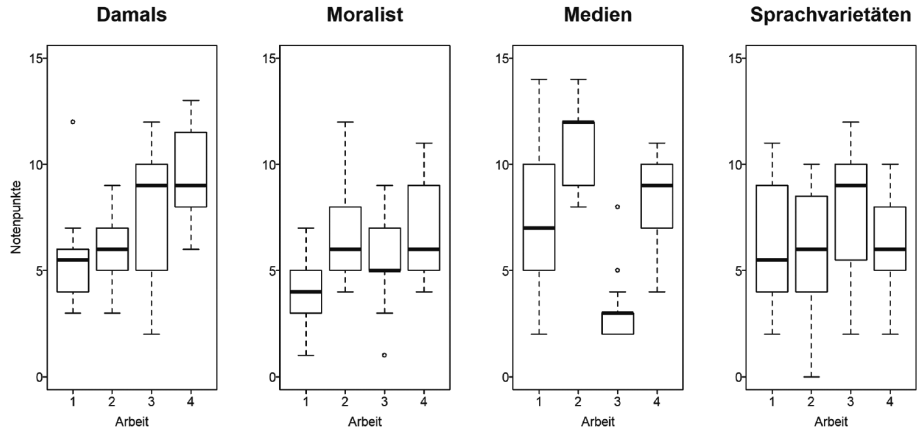
Durchführung

Die Datenerhebung erfolgte von Mitte August bis Mitte Oktober 2018. Alle Materialien wurden den Lehrkräften postalisch zugesendet und konnten bei freier Zeiteinteilung bearbeitet werden. Die Lehrkräfte erhielten jeweils vier transkribierte Abiturarbeiten zu zwei Aufgaben, acht Bewertungsbögen, zwei „Korrekturhilfen" (bestehend aus EWH und Bewertungshinweisen für jede Aufgabe) und einen Fragebogen. Sie wurden gebeten, sich bei der Korrektur und Bewertung der Arbeiten ausschließlich an den beiliegenden „Korrekturhilfen" zu orientieren und den Fragebogen erst zum Schluss auszufüllen. Innerhalb von acht Wochen Bearbeitungszeit wurden die Materialien in einem beigelegten Rücksendeumschlag postalisch an das IQB übermittelt. Alle eingesendeten Materialien wurden mit einem Code versehen und getrennt von den personenbezogenen Informationen gespeichert. Die Eingabe und Auswertung der Daten erfolgten anonymisiert.

3.3 Ergebnisse

Bewertung der Abiturarbeiten

In Abbildung 2 wird die Verteilung der Notenpunkte für jede Arbeit zu den vier verschiedenen Aufgaben in Form von Boxplots grafisch dargestellt. Bei allen Aufgaben ist eine sehr starke Streuung der erteilten Notenpunkte festzustellen. Dieser Befund zeigt sich sowohl bei Verwendung von analytischem EWH als auch bei Verwendung von holistischem EWH. Im Mittel wurden alle Arbeiten zu einer Aufgabe mit sechs bis sieben Notenpunkten bewertet ($M = 6.79$ NP, $SD = 3.01$ NP), wobei sich Differenzen für die Bewertung derselben Arbeit von bis zu 12 Notenpunkten ergaben. Diese Ergebnisse stimmen mit den Befunden von Birkel und Birkel (2002) überein, die für die Bewertung von Aufsätzen durch Deutschlehrkräfte in der Grundschule teilweise sehr starke Streuungen in der Benotung einzelner Arbeiten feststellten.



Anmerkungen: In dem Rechteck (Box) befinden sich 50 Prozent der Werte, der Querstrich markiert die Mitte der Verteilung (Median), die gestrichelten Ausläufer kennzeichnen die Minima und Maxima der Verteilung.

Abbildung 2: Grafische Darstellung der Notenpunkteverteilung in Form von Boxplots für jede Arbeit, getrennt nach Aufgaben

Während extreme Differenzen in der Bewertung bei den meisten Aufgaben nur für einzelne Arbeiten zu beobachten sind, divergieren bei der MA-Aufgabe „Sprachvarietäten“ die Bewertungen aller Arbeiten stark. Dies wird bei Betrachtung der zu erwartenden Anteile notwendiger Drittkorrekturen[3] bei den einzelnen Arbeiten besonders deutlich. Zur Berechnung der Anteile wurden in einem ersten Schritt nach den Regeln der Kombinatorik für jede einzelne Arbeit alle möglichen Zweierpaarungen der Lehrkräfte ermittelt, die die betreffende Arbeit in der Studie bewertet hatten. In einem zweiten Schritt wurde für jedes Paar berechnet, in welchem Maße beide Notenurteile voneinander abwichen. Für jedes Paar, bei dem diese Differenz größer als drei Notenpunkte war, wurde vermerkt, dass im betreffenden Fall unter realen Bedingungen in der Mehrzahl der Länder eine Drittkorrektur erfolgt wäre. Die Anteile der zu erwartenden Drittkorrekturen wurden schließlich als Anteil der Paare pro Arbeit berechnet, deren Bewertungen um mehr als drei Notenpunkte voneinander abwichen. Diese Anteile sind in Abbildung 3 dargestellt. Es wird deutlich, dass bei einigen Arbeiten der Anteil an notwendigen Drittkorrekturen relativ gering ausfällt, während andere Arbeiten stark polarisieren. Sechs Arbeiten, darunter drei Arbeiten zur MA-Aufgabe „Sprachvarietäten“, wären unter Prüfungsbedingungen in 25 bis 35 Prozent der Fälle einer Drittkorrektur unterzogen worden.

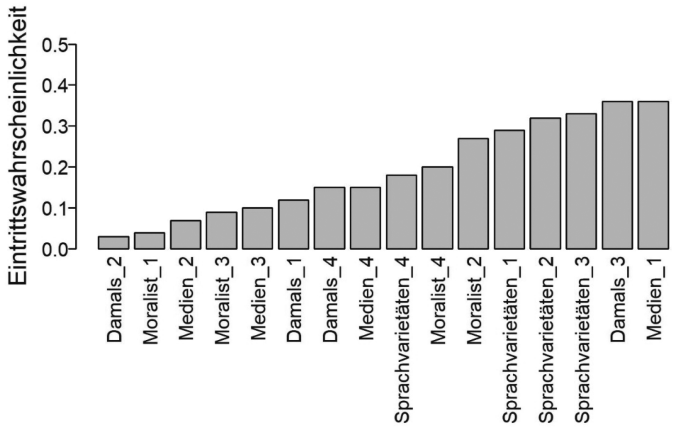


Abbildung 3: Anteile der (unter Prüfungsbedingungen zu erwartenden) Drittkorrekturen für jede Arbeit

Tabelle 2 zeigt die mittlere Bewertung der Arbeiten pro Land und Aufgabe getrennt nach den Bewertungsbedingungen. Unterschiede zwischen den Bewertungsbedingungen wurden durch lineare Regressionsanalysen ermittelt, bei denen das Land, aus dem die Lehrkräfte kamen, statistisch kontrolliert wurde. Tendenziell wurden die Arbeiten in der analytischen Bewertungsbedingung von den Lehrkräften im Mittel etwas strenger bewertet. Über alle Aufgaben hinweg ergab sich jedoch kein statistisch signi-

3 Bei größeren Abweichungen der Bewertungsvorschläge von Erst- und Zweitkorrektor:in ist in der Mehrzahl der Länder eine Drittkorrektur vorgesehen.

fikanter Unterschied in der Bewertung zwischen den Bewertungsbedingungen ($t(261) = 1.50$, $p = .13$). Dieser Befund lässt darauf schließen, dass analytische Bewertungen nicht generell strenger ausfallen als holistische. Lediglich bei der Aufgabe „Moralist" wurden die Arbeiten im Mittel signifikant strenger bewertet, wenn die Lehrkräfte einen analytischen EWH für die Bewertung nutzten ($t(65) = 3.40$, $p = .001$). In einer nachträglichen Begutachtung der in der Studie eingesetzten Abiturarbeiten durch fachdidaktische Expert:innen wurde deutlich, dass vor allem bei der Aufgabe „Moralist" viele Lösungen der Prüflinge von den im EWH ausgeführten inhaltlichen Erwartungen abwichen. Vermutlich hat dies vor allem bei der analytischen Beurteilung die Punktevergabe bei einzelnen Kriterien erschwert.

Tabelle 2: Bewertung der Arbeiten getrennt nach Bewertungsbedingungen

	Analytischer EWH					Holistischer EWH				
	M	*SD*	*SE*	*Min*	*Max*	*M*	*SD*	*SE*	*Min*	*Max*
insgesamt	6.52	3.10	0.27	0	14	7.08	2.91	0.26	2	14
Land										
Land A	6.36	3.26	0.38	0	14	7.30	3.14	0.39	2	14
Land H	6.70	2.92	0.36	2	14	6.86	2.67	0.33	2	12
Aufgabe										
Damals	7.12	2.67	0.47	2	12	7.19	2.96	0.52	3	13
Moralist	5.03	2.42	0.40	1	11	6.91	2.2	0.39	3	12
Medien	7.44	3.66	0.61	2	14	7.53	3.64	0.64	2	14
Sprachvarietäten	6.56	2.98	0.53	0	11	6.69	2.72	0.48	2	12

Anmerkung: Mittlere Bewertung aller Arbeiten in Notenpunkten: *M* = Mittelwert, *SD* = Standardabweichung, *SE* = Standardfehler, *Min* = kleinster Wert, *Max* = größter Wert.

Übereinstimmung der Bewertungen

Als etabliertes Maß für den Grad der Übereinstimmung der Bewertungen wurde zunächst der Intra-Klassen-Korrelationskoeffizient (ICC) ermittelt. ICCs können Werte zwischen 0 und 1 annehmen, wobei ICC-Werte ab .7 als akzeptable Übereinstimmung und ICC-Werte ab .9 als exzellente Übereinstimmung unabhängiger Beurteilungen gelten (Döring & Bortz, 2016; Koo & Li, 2016). Für den analytischen EWH wurde hier ein Wert von 0.22, für den holistischen EWH ein Wert von 0.43 ermittelt. Da die Interpretation der ICCs bei kleineren Stichproben jedoch nur eingeschränkt möglich ist, wurde anhand des Abstands der einzelnen Bewertungen zum Mittelwert aller Bewertungen pro Arbeit (in Notenpunkten) noch ein zweites Maß für den Grad der Übereinstimmung zwischen den Lehrkräften bestimmt. Für die durchgeführten Regressionsanalysen wurde dieser Abstand quadriert, um eine Gewichtung vorzunehmen; kleinere Abweichungen fallen damit weniger, größere Abweichungen hingegen stärker ins Ge-

wicht. Tabelle 3 enthält die absoluten Abweichungen zwischen den einzelnen Bewertungen und dem Mittelwert aller Bewertungen pro Land und Aufgabe getrennt nach Bewertungsbedingungen. Je kleiner die über alle Arbeiten einer Aufgabe gemittelte Abweichung ausfällt, desto größer ist die Übereinstimmung der Bewertungen zwischen den Lehrkräften.

Tabelle 3: Abweichung vom Mittelwert getrennt nach Bewertungsbedingungen

	Analytischer EWH					Holistischer EWH				
	M	*SD*	*SE*	*Min*	*Max*	*M*	*SD*	*SE*	*Min*	*Max*
Insgesamt	2.86	2.05	0.18	0	10	2.37	1.54	0.14	0	6
Land										
Land A	3.12	2.05	0.24	0.12	10.0	2.44	1.52	0.19	0	6.94
Land H	2.57	2.02	0.25	0	10.0	2.3	1.58	0.20	0.12	6.00
Aufgabe										
Damals	1.71	1.55	0.27	0.19	6.50	1.82	1.32	0.23	0.12	4.12
Moralist	3.91	1.94	0.32	0.18	6.76	2.42	1.33	0.24	0.24	4.82
Medien	3.21	1.32	0.23	0.12	4.12	3.11	1.88	0.33	0	6.94
Sprachvar.	2.45	1.60	0.28	0.12	5.88	2.14	1.33	0.24	0.31	5.81

Anmerkung: Absolute Abweichung zwischen den einzelnen Bewertungen und dem Mittelwert aller Bewertungen gemittelt über alle Arbeiten in Notenpunkten: *M* = Mittelwert, *SD* = Standardabweichung, *SE* = Standardfehler, *Min* = kleinster angegebener Wert, *Max* = größter angegebener Wert.

Für beide Bewertungsbedingungen zeigen die Ergebnisse zum Grad der Übereinstimmung zwischen den Lehrkräften bei allen Aufgaben eine mittlere Abweichung zum Mittelwert von zwei bis drei Notenpunkten. Allerdings fallen die Diskrepanzen größer aus, wenn (bei statistischer Kontrolle für das Land, aus dem die Untersuchungsteilnehmer:innen jeweils stammten) für die Bewertung der Arbeiten analytische EWH verwendet wurden ($t(261) = -2.65$, $p = .009$). Auch dieser Effekt kann anhand der zu erwartenden Anteile an Drittkorrekturen veranschaulicht werden: In Abbildung 4 wird die Höhe dieser Anteile für jede einzelne Arbeit getrennt nach Bewertungsbedingung dargestellt. Die Ergebnisse zeigen, dass unter analytischen Bewertungsbedingungen eine größere Anzahl an Arbeiten mit höherer Wahrscheinlichkeit eine Drittkorrektur erfordert hätte als unter holistischen Bewertungsbedingungen.

Die Unterschiede in der Übereinstimmung zwischen den Lehrkräften aus den beiden Ländern sind statistisch nicht signifikant ($t(261) = 1.19$, $p = .23$). Dieser Befund weist darauf hin, dass der Grad der Übereinstimmung in der Bewertung von Abiturarbeiten in dieser Gruppe von Lehrkräften nicht von der Vertrautheit mit der Form der verwendeten EWH abhängt. Bei einer aufgabenspezifischen Betrachtung zeigt sich, dass der Unterschied im Grad der Übereinstimmung zwischen den Bewertungsbe-

dingungen erneut größtenteils auf die IL-Aufgabe „Moralist“ zurückzuführen ist ($t(257) = 4.36$, $p < .001$). Da die Arbeiten zu dieser Aufgabe im Vergleich zu den Arbeiten zu anderen Aufgaben viele vom EWH abweichende Lösungen enthielten, könnte auch dieser Befund darauf zurückgehen, dass bei alternativen Lösungen die Erfüllung einzelner Kriterien mittels eines eher starren analytischen Bewertungsrasters schwerer zu beurteilen ist als mithilfe eines holistischen EWH, der mehr Entscheidungsspielräume lässt. Im Rahmen des holistischen Bewertungsmodells können andere als im EWH ausgeführte Lösungen bei der Bewertung der Prüfungsleistung als gleichwertig gewürdigt werden, wenn sie der Aufgabenstellung entsprechen, sachlich richtig und nachvollziehbar sind. Im Vergleich dazu sind im analytischen Bewertungsraster für die Bewertung alternativer Lösungen jeweils maximal 3 Punkte für die Verstehens- und die Darstellungsleistung vorgesehen.

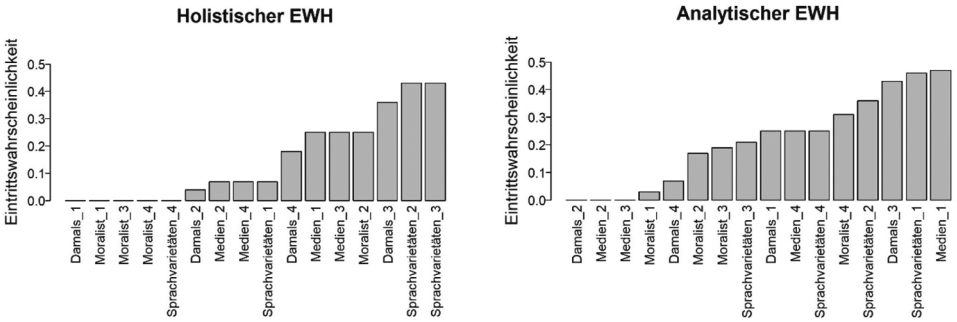


Abbildung 4: Anteile der (unter Prüfungsbedingungen zu erwartenden) Drittkorrekturen für jede Arbeit getrennt nach Bewertungsbedingung

Anfälligkeit für Urteilsverzerrungen

Um die Anfälligkeit der verschiedenen Bewertungsformen für Urteilsverzerrungen zu untersuchen, wurde der Zusammenhang zwischen der Einhaltung standardsprachlicher Normen und der Bewertung der Verstehensleistung der vorliegenden Abiturarbeiten modelliert. Dazu wurde die Korrelation zwischen dem Fehlerquotienten[4] einer Abiturarbeit und der Bewertung der Verstehensleistung berechnet. Abbildung 5 zeigt den über alle Arbeiten gemittelten linearen Zusammenhang zwischen den für die inhaltliche Leistung vergebenen Notenpunkten und den Fehlerquotienten unter Nutzung eines holistischen sowie analytischen EWH. Für den holistischen EWH fällt dieser Zusammenhang statistisch signifikant höher aus als für den analytischen EWH ($t(25) = -2.41$, $p = .02$). Dies weist darauf hin, dass sich Lehrkräfte unter Nutzung holistischer EWH bei der Bewertung der Verstehensleistung einer Abiturarbeit stärker davon beeinflussen lassen, ob die Arbeit viele Fehler im Hinblick auf Rechtschreibung, Grammatik und Zeichensetzung aufweist. Dieser Befund deckt sich mit den Erkenntnissen aus anderen empirischen Untersuchungen, wonach ho-

4 Als Fehlerquotient wird an dieser Stelle der Quotient bezeichnet, der aus der Anzahl der Fehler (in Rechtschreibung, Grammatik und Zeichensetzung) und der Gesamtwortzahl einer Arbeit gebildet wird.

listische Bewertungen im Vergleich zu analytischen Urteilen etwas anfälliger für Urteilsverzerrungen wie den Halo-Effekt sind (z. B. Böhme et al., 2009).

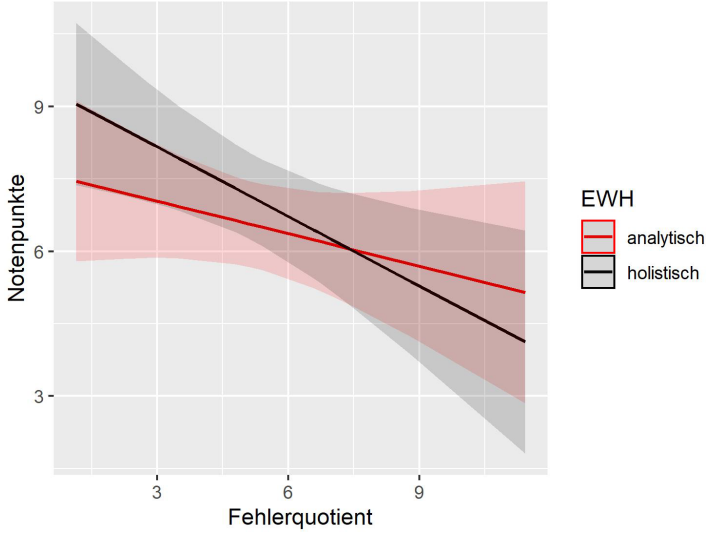


Abbildung 5: Korrelation zwischen den für die Verstehensleistung vergebenen Notenpunkten und den Fehlerquotienten aller Arbeiten unter Nutzung eines holistischen (schwarz) sowie analytischen (rot) EWH

Nutzung und Akzeptanz der EWH

In Bezug auf die Verwendung der EWH könnte vermutet werden, dass Lehrkräfte den Grad der Erfüllung der einzelnen Kriterien des analytischen EWH nicht wie intendiert sukzessive für jedes Kriterium beurteilen, sondern gemäß ihrem ersten Eindruck zunächst ein vorläufiges holistisches Gesamturteil bilden und sich bei der Bewertung der einzelnen Kriterien dann an diesem Urteil orientieren. Die hierzu mithilfe von Multiple-Choice-Items erhobenen Angaben der Studienteilnehmer:innen bestätigen diese Vermutung jedoch nicht (Abbildung 6).

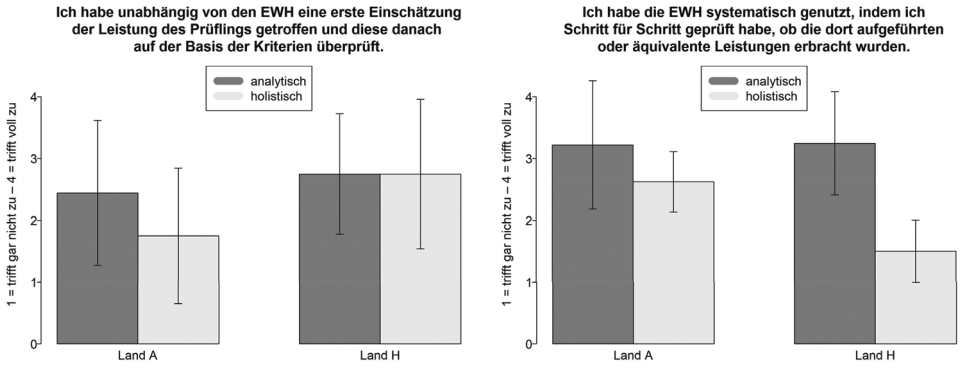


Abbildung 6: Mittelwerte (inkl. Standardabweichung) für die Angaben der Lehrkräfte bei ausgewählten Fragebogenitems zur Verwendung der EWH bei der Bewertung

Die in Abbildung 7 dargestellten Ergebnisse zum Vorgehen bei der Bewertung bestätigen, dass holistische EWH mehr Entscheidungsspielräume lassen, während analytische EWH eher dazu anhalten, die Bewertung vor allem auf der Grundlage der im EWH angeführten Kriterien vorzunehmen. Folglich liegt der Schluss nahe, dass holistische EWH den Vorteil bieten, Lösungen, die sich nicht im EWH finden, aber dennoch der Aufgabenstellung entsprechen und sachlich richtig sind, adäquater würdigen zu können.

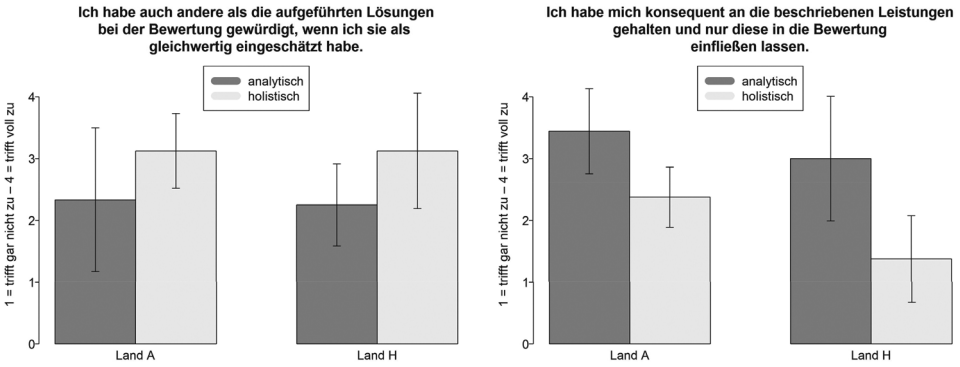


Abbildung 7: Mittelwerte für die Angaben der Lehrkräfte bei ausgewählten Fragebogenitems zum Vorgehen bei der Bewertung

Der Nutzen der EWH für die Bewertung der Abiturarbeiten wurde gemäß der jeweils vertrauten Bewertungspraxis von Lehrkräften aus Land A für analytische und von Lehrkräften aus Land H für holistische EWH jeweils geringfügig höher eingeschätzt (Abbildung 8, links). Eine positivere Beurteilung des Nutzens des EWH führte jedoch nicht zu einer höheren Übereinstimmung bei der Bewertung. Darüber hinaus sprachen sich die Lehrkräfte mehrheitlich für EWH aus, in denen die formulierten Erwartungen an die inhaltliche Leistung keine Ideallösung darstellen, sondern von sehr guten Prüflingen auch tatsächlich erfüllbar sein sollten (Abbildung 8, rechts).

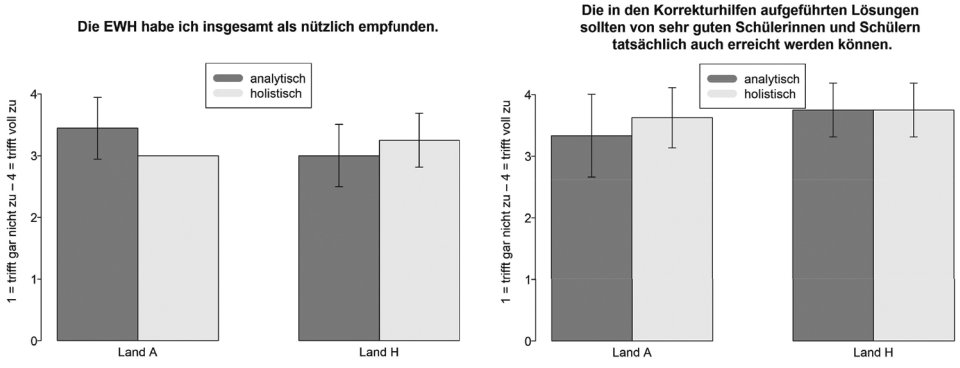


Abbildung 8: Mittelwerte für die Angaben der Lehrkräfte bei ausgewählten Fragebogenitems zur Einschätzung der EWH

3.4 Schlussfolgerungen

Zusammenfassend lässt sich anhand der vorliegenden Ergebnisse kein eindeutiger Vor- bzw. Nachteil für die Verwendung von holistischen oder analytischen EWH bei der Bewertung von Abiturarbeiten im Fach Deutsch ableiten. In beiden Bewertungsbedingungen war die Spanne der Bewertung mit durchschnittlich sechs bis sieben Notenpunkten sehr groß. Dieser Befund deckt sich sowohl mit den Befunden der Studie von Birkel und Birkel (2002) als auch mit den Ergebnissen der Evaluation der Poolaufgaben im Prüfungsjahr 2017 durch das IQB, in der zum Teil große Abweichungen zwischen der Bewertung durch Erst- und Zweitkorrektor:in identifiziert wurden. Obwohl der Grad der Übereinstimmung in beiden Bewertungsbedingungen insgesamt niedrig war, führte die Verwendung von holistischen EWH zu einer statistisch signifikant höheren Übereinstimmung in der Bewertung. Dieser Effekt ließ sich jedoch auf die Bewertung der Arbeiten zu einer einzelnen Aufgabe zurückführen, bei der die Lösungen der Prüflinge stark von den im EWH formulierten inhaltlichen Erwartungen abwichen. Hier fielen die Urteile unter Nutzung analytischer EWH deutlich diskrepanter aus, was darauf zurückzuführen sein könnte, dass bei alternativen Lösungen der Grad der Erfüllung einzelner Kriterien mit einem Bewertungsraster unter Umständen schwerer zu beurteilen ist. Anders als häufig angenommen, lieferten die durchgeführten Analysen somit keine Evidenz dafür, dass ein analytischer EWH zu reliableren Bewertungen führt.

Ein weiterer Unterschied zwischen den Bewertungsbedingungen fand sich in Bezug auf die Anfälligkeit für Urteilsverzerrungen bei der Bewertung. Die Untersuchung des Zusammenhangs zwischen den für die inhaltliche Leistung vergebenen Notenpunkten und den sprachlichen Oberflächenmerkmalen der Arbeiten wies darauf hin, dass die inhaltliche Bewertung stärker vom Fehlerquotienten beeinflusst wurde, wenn die Lehrkräfte einen holistischen EWH verwendeten. Dieser Befund repliziert die Ergebnisse aus anderen empirischen Studien, nach denen holistische im Vergleich zu analytischen Urteilen anfälliger für Urteilsverzerrungen sind (Böhme et al., 2009). Im Hinblick auf die Gestaltung von EWH für Prüfungsaufgaben im Fach Deutsch könnte dies darauf hinweisen, dass eine separate Bewertung der inhaltlichen Leistung und der sprachlichen Leistung sinnvoll ist. Dies ließe sich beispielsweise durch die Vorgabe einer Gewichtung von Verstehens- und Darstellungsleistung umsetzen, die in einigen Ländern bereits Bestandteil der EWH ist.

4 Studie 2

4.1 Zielstellung

Ausgehend von den Ergebnissen aus Studie 1 ging Studie 2 der Frage nach, welche Auswirkungen die Vorgabe einer separaten Bewertung und Gewichtung der Verstehens- und Darstellungsleistung hat. In der überwiegenden Zahl der Länder werden die Kriterien für die Bewertung der Darstellungsleistung differenziert ausgewiesen, wobei der Grad der Differenzierung variiert. In einzelnen Ländern ist vorgesehen,

dass auch die Verstehensleistung auf der Grundlage differenzierter Kriterien bewertet wird. In der Mehrzahl der Länder ist ferner eine Gewichtung von Verstehens- und Darstellungsleistung vorgesehen. Diese erfolgt entweder durch eine prozentuale Gewichtung (ungefähre Vorgaben oder feste Vorgaben, die in Teilnoten münden) oder durch ein Punktesystem (maximal erreichbare Anzahl an Punkten oder Bewertungseinheiten für den jeweiligen Leistungsbereich). Aus diesen unterschiedlichen Bewertungsvorgaben ergeben sich verschiedene Varianten in Bezug auf die Differenzierung und Gewichtung der Verstehens- und Darstellungsleistung in den EWH der Länder. Vor diesem Hintergrund bestand das Ziel von Studie 2 darin, die Vor- und Nachteile der unterschiedlichen Varianten im Hinblick auf die Bewertung von Abiturprüfungsarbeiten zu identifizieren. Untersucht wurden dabei Unterschiede zwischen den EWH-Varianten in Bezug auf die vier Schwerpunkte, die bereits in Studie 1 betrachtet wurden (Bewertung der Abiturarbeiten, Übereinstimmung der Bewertungen, Anfälligkeit für Urteilsverzerrungen, Nutzung und Akzeptanz der EWH).

4.2 Methode

Analog zu Studie 1 wurde ein experimentelles Between-Subjects-Design gewählt, um die Bewertungen von Abiturarbeiten durch Lehrkräfte mittels verschiedener EWH miteinander zu vergleichen. Um Aufgabeneffekte von vornherein auszuschließen, wurde anders als in Studie 1 nur eine Aufgabe in die Studie einbezogen. Hierbei handelte es sich um die IL-Aufgabe mit der Kurzbezeichnung „Damals“, die bereits in Studie 1 zum Einsatz kam und in ihrer Ausprägung als Gedichtvergleich eine in allen Ländern etablierte Aufgabenart darstellt. Die Aufgabe bestand aus zwei Teilaufgaben, deren Schwerpunkte auf der Interpretation eines Gedichts (Teilaufgabe 1) und dem Vergleich der Gestaltung der Gedichte (Teilaufgabe 2) lagen.

Um bestehende Länderunterschiede in Bezug auf die Gestaltung der EWH bestmöglich abzubilden, wurden von den Autor:innen der Studie fünf Varianten des EWH erstellt, die anschließend von Lehrkräften auf ihre Anwendbarkeit in der Praxis geprüft wurden. Als Vorlagen für die Entwicklung dieser Varianten wurden die EWH landeseigener Abiturprüfungsaufgaben aus allen 16 Ländern herangezogen. Die in Tabelle 4 dargestellten fünf EWH-Varianten bilden die möglichen Bewertungsmodelle im Hinblick auf die Verstehens- und Darstellungsleistung über alle Länder hinweg ab. Die Varianten unterscheiden sich dabei nicht in Bezug auf die Art und den Umfang der geforderten Verstehensleistung. Die Kriterien und zugehörigen Wissensbestände für die Bewertung der Verstehensleistung wurden aus dem EWH der Poolaufgabe übernommen und sind in allen EWH-Varianten identisch.

Tabelle 4: Überblick über Unterschiede und Gemeinsamkeiten der in der Studie eingesetzten EWH-Varianten

EWH-Variante	Grad der Differenzierung der Kriterien für die DL	Ausweisung der Bewertungshinweise für eine „gute" und eine „ausreichende" Leistung	Gewichtung von VL und DL	Vorgaben zur Vorgehensweise bei der Ermittlung der Gesamtnote
1	wenig differenziert	keine getrennte Ausweisung von VL und DL	keine Gewichtung	keine Vorgaben
2	stark differenziert	getrennte Ausweisung von VL und DL	keine Gewichtung	keine Vorgaben
3	stark differenziert	getrennte Ausweisung von VL und DL	Gewichtung (mit „ca." versehene Prozentangaben)	keine Vorgaben
4	stark differenziert	getrennte Ausweisung von VL und DL	Gewichtung (genaue Prozentangaben)	Vorgabe für die Berechnung des Gesamtergebnisses aus VL und DL
5	stark differenziert	getrennte Ausweisung von VL und DL	Gewichtung (genaue Prozentangaben)	Vorgabe für die Berechnung des Gesamtergebnisses aus kriterial gewichteter DL und VL

Anmerkung: Beispiele für die Gestaltung von EWH-Variante 3 können im Rahmen der veröffentlichten Abiturprüfungsaufgaben für das Fach Deutsch auf der Webseite des IQB eingesehen werden; VL = Verstehensleistung, DL = Darstellungsleistung.

Alle EWH-Varianten sahen eine Gewichtung der Teilaufgaben vor (Teilaufgabe 1: ca. 60 %, Teilaufgabe 2: ca. 40 %) und umfassten Kriterien für die Bewertung der Verstehensleistung beider Teilaufgaben. Ebenso beinhalteten alle EWH-Varianten kriteriale Bewertungshinweise für eine „gute" (11 Punkte) und eine „ausreichende" (5 Punkte) Leistung. EWH-Variante 1 enthielt diese Bewertungshinweise nur in Bezug auf die Gesamtleistung, während alle übrigen Varianten die Bewertungshinweise getrennt nach Verstehens- und Darstellungsleistung auswiesen. In den Varianten 2 bis 5 waren die Hinweise für die Bewertung der Darstellungsleistung dabei separat für vier verschiedene Kriterien angegeben: Textaufbau, sprachliche Gestaltung (Fachsprache, Ausdruck, Syntax), Umgang mit Textbelegen und Einhaltung standardsprachlicher Normen. In Variante 5 waren die Bewertungshinweise auch für die Verstehensleistung für jede Teilaufgabe differenziert ausgewiesen. Die Unterschiede zwischen den EWH-Varianten werden im Folgenden skizziert:

- Bei der Bewertung der Gesamtleistung sind in **EWH-Variante 1** sowohl die Verstehensleistung als auch die Darstellungsleistung zu berücksichtigen. Die Notenbildung erfolgt jedoch nicht durch eine Gewichtung der Verstehensleistung (weder der Teilaufgaben 1 und 2 noch im Verhältnis zur Darstellungsleistung).
- In **EWH-Variante 2** wird im Gegensatz zu Variante 1 die Darstellungsleistung für beide Teilaufgaben entlang von vier Kriterien (s. o.) differenziert und jeweils für eine gute und eine ausreichende Aufgabenbearbeitung ausgewiesen. Auch hier gibt es keine Vorgaben zum Umgang mit der Verstehensleistung und der Darstellungsleistung bei der Ermittlung der Gesamtnote.

- **EWH-Variante 3** ist analog zur Variante 2 aufgebaut, verlangt jedoch zusätzlich eine Gewichtung von Verstehens- und Darstellungsleistung im Verhältnis von ca. 70 zu ca. 30 Prozent. Wie diese Gewichtung bei der Ermittlung der Gesamtnote umgesetzt wird, ist nicht vorgegeben.
- **EWH-Variante 4** unterscheidet sich in Bezug auf den Umgang mit der vorgegebenen Gewichtung von den Varianten 2 und 3 in einer zusätzlichen Berechnungsgrundlage zur Ermittlung der Teilnoten und der Gesamtnote. Dabei werden jeweils die Noten für die Verstehensleistung für Teilaufgabe 1 und 2 festgelegt und zueinander in das vorgegebene Verhältnis gesetzt (Gesamtwert „Verstehensleistung"). Festgesetzt wird ferner die Note für die Darstellungsleistung bei der Gesamtbearbeitung der Aufgabe. Zur Ermittlung der Gesamtnote wird die Notenpunktzahl für die Verstehensleistung mit der Notenpunktzahl für die Darstellungsleistung im vorgegebenen Verhältnis verrechnet (siehe Abbildung 9a).
- **EWH-Variante 5** ist analog zu Variante 4 aufgebaut, weist aber einen noch höheren Differenzierungsgrad der Kriterien für die Bewertung der Verstehens- und der Darstellungsleistung auf, der bei der Berechnung der Gesamtnote umgesetzt wird. Für die Teilaufgaben 1 und 2 wird jeweils die Note für die Verstehensleistung auf der Grundlage prozentual gewichteter Teilleistungen und unter Berücksichtigung der vorgegebenen Gewichtung der Teilaufgaben berechnet. Die Note für die Darstellungsleistung wird über die bereits genannten vier Kriterien ermittelt, wobei die Kriterien jeweils prozentual gewichtet sind. Die Gesamtnote wird aus der Notenpunkzahl für die Verstehensleistung und der Notenpunktzahl für die Darstellungsleistung im vorgegebenen Verhältnis berechnet (siehe Abbildung 9b).

Über eine Ausschreibung auf der Webseite des IQB wurden bundesweit 91 Deutsch-Lehrkräfte rekrutiert, die über mindestens drei Jahre Erfahrung mit der Korrektur und Bewertung von schriftlichen Abiturarbeiten im Fach Deutsch verfügten. Die Teilnahme an der Studie war freiwillig und wurde vom IQB mit einer Aufwandsentschädigung vergütet. Jede Lehrkraft wurde gebeten, acht Abiturarbeiten anhand vorgegebener Materialien zu bewerten, wobei die Zuweisung zu einer EWH-Variante zufällig erfolgte. Bei der Zuweisung wurde jedoch darauf geachtet, dass Lehrkräfte aus Ländern mit holistischer sowie analytischer Bewertungstradition gleichmäßig auf die Varianten verteilt wurden.

	Verstehensleistung		Darstellungsleistung
Beurteilungsbereiche	Teilaufgabe 1 • Texterschließung (Inhalt und Aufbau) • Erschließung textkonstituierender sprachlicher und formaler Mittel in ihrem Wirkungszusammenhang • Deutung des Gedichts Teilaufgabe 2 • Vergleich beider Gedichte • Erfassen zentraler inhaltlicher Gemeinsamkeiten und Unterschiede • Vergleich der sprachlichen Gestaltung		• Textaufbau • sprachliche Gestaltung (Fachsprache, Ausdruck, Syntax) • Umgang mit Textbelegen • Standardsprachliche Normen
Ermittlung der Gesamtnote für die Verstehensleistung der Teilaufgaben	Gewichtung Teilaufgabe 1: 60 %	Gewichtung Teilaufgabe 2: 40 %	
Ermittlung der Gesamtnote	Gewichtung: 70 %		Gewichtung: 30 %

Abbildung 9a: Schematische Darstellung der Vorgehensweise zur Ermittlung der Gesamtnote in EWH-Variante 4

Ermittlung der Note für die Verstehensleistung	**Teilaufgabe 1**		**Teilaufgabe 2**	
	Teilleistungen	**Prozent**	**Teilleistungen**	**Prozent**
	Texterschließung	30 %	Erfassen zentraler inhaltlicher Gemeinsamkeiten und Unterschiede	80 %
	Erschließung text-konstituierender sprachlicher und formaler Mittel in ihrem Wirkungszusammenhang	50 %	Vergleich der sprachlichen Gestaltung	20 %
	Deutung des Gedichts	20 %		
	Teilaufgabe 1:	**60 %**	**Teilaufgabe 2:**	**40 %**
Ermittlung der Note für die Darstellungsleistung	**Teilleistungen**			**Prozent**
	Textaufbau			30 %
	Sprachliche Gestaltung (Fachsprache, Ausdruck, Syntax)			40 %
	Umgang mit Textbelegen			10 %
	Standardsprachliche Normen			20 %
Ermittlung der Gesamtnote	Note für die Verstehensleistung:	**70 %**	Note für die Darstellungsleistung:	**30 %**

Abbildung 9b: Schematische Darstellung der Vorgehensweise zur Ermittlung der Gesamtnote in EWH-Variante 5

Stichprobe

Die Gesamtstichprobe setzte sich aus 69 weiblichen und 22 männlichen Lehrkräften aus insgesamt neun Ländern zusammen. 35 kamen aus Ländern mit einer eher analytischen Bewertungstradition, während 56 aus Ländern mit einer eher holistischen Bewertungstradition stammten. Die Teilnehmer:innen waren im Mittel 49 Jahre alt ($SD = 9$ Jahre) und verfügten über durchschnittlich 20 Jahre Unterrichtserfahrung ($SD = 10$ Jahre). Etwa die Hälfte der Lehrkräfte (48 %) gab an, mehr als zwölfmal in einem Abiturjahrgang im Fach Deutsch eingesetzt worden zu sein (neun- bis zwölfmal: 16 %, fünf- bis achtmal: 25 %, ein- bis viermal: 11 %).

Material

Aus den für die Evaluation der Poolaufgaben 2017 vorliegenden Kopien von Abiturarbeiten für das Fach Deutsch wurden 40 Arbeiten zur Aufgabe „Damals" ausgewählt und transkribiert[5]. Alle Arbeiten waren von vergleichbarer Länge und deckten in Bezug auf die Bewertung der Verstehens- sowie Darstellungsleistung ein breites Leistungsspektrum ab. In einem nächsten Schritt wurde diese Vorauswahl einem Rating unterzogen. Dafür wurden zehn Expert:innen (Lehrkräfte und Fachdidaktiker:innen) gebeten, jeweils 12 Arbeiten auf einer vierstufigen Likert-Skala hinsichtlich der Verstehensleistung und der Darstellungsleistung einzuschätzen und mit Notenpunkten zu bewerten. Die Bewertung der Darstellungsleistung erfolgte anhand der Kriterien „Textaufbau", „sprachliche Gestaltung", „Umgang mit Textbelegen" sowie „Einhaltung standardsprachlicher Normen". Insgesamt lagen für jede der 40 Arbeiten die Einschätzungen und Bewertungen von drei Expert:innen vor.

Die Ergebnisse des Ratings stellten die Grundlage für die finale Auswahl der Arbeiten für die Studie dar. Als Auswahlkriterien wurden eine hohe Übereinstimmung in den Expertenurteilen (ICC > .8) sowie, anders als in Studie 1, eine breite Abdeckung des Leistungsspektrums sowohl in Bezug auf die Verstehens- als auch auf die Darstellungsleistung zugrunde gelegt. Ausgewählt wurden insgesamt 24 Arbeiten, die in der mittleren Bewertung der Gesamtleistung einen Bereich von 3 bis 14 Notenpunkten abdeckten ($M = 8.5$ NP, $SD = 3.1$ NP).

Instrumente

Die Bewertung der einzelnen Abiturarbeiten durch die Lehrkräfte erfolgte anhand von Bewertungsbögen, die in ihrer Struktur an die jeweilige EWH-Variante angepasst waren. Lehrkräfte mit den Varianten 1 und 2 erhielten zu jeder Abiturarbeit einen Bewertungsbogen, auf dem die Bewertung der Gesamtleistung eingetragen werden sollte. Lehrkräfte mit der EWH-Variante 3 erhielten Bewertungsbögen, auf denen die Bewertung der Verstehens- und Darstellungsleistung sowie die daraus ermittelte Bewertung der Gesamtleistung festgehalten werden sollte. Lehrkräfte mit der EWH-Variante 4 sollten auf den Bewertungsbögen die Bewertung der Verstehensleistung für jede Teilaufgabe, die daraus ermittelte Bewertung der Verstehensleistung insgesamt, die Be-

5 Die Transkription der Arbeiten erfolgte unverändert, d. h. inklusive aller Fehler in Rechtschreibung, Grammatik und Zeichensetzung.

wertung der Darstellungsleistung und die daraus ermittelte Bewertung der Gesamtleistung vermerken. Lehrkräfte mit der EWH-Variante 5 wurden zusätzlich gebeten, die Bewertung der einzelnen Kriterien für die Verstehens- und Darstellungsleistung zu vermerken. Um ihnen die Berechnungen zur Ermittlung der Notenpunkte zu erleichtern, erhielten die Lehrkräfte mit den Varianten 4 und 5 zusätzlich zu den Bewertungsbögen ein auf ihre EWH-Variante zugeschnittenes Instrument (in Form einer Excel-Datei), das anhand der eingetragenen Notenpunkte die Bewertung der Gesamtleistung automatisch berechnete. Die Nutzung der Bewertungsbögen ermöglichte es den Lehrkräften, die Bewertung der Abiturarbeiten örtlich sowie zeitlich flexibel vorzunehmen. Erfasst wurden die Ergebnisse zur Bewertung anschließend mittels eines webbasierten Eingabeinstruments. Mithilfe dieses Eingabeinstruments wurden die auf den Bewertungsbögen festgehaltenen Ergebnisse von den Lehrkräften in passende Eingabemasken übertragen und anonymisiert gespeichert.

Alle Lehrkräfte wurden gebeten, im Anschluss an die Eingabe ihrer Ergebnisse zur Bewertung einen Fragebogen auszufüllen. Dieser enthielt Multiple-Choice-Items, mit denen weitere Angaben und Einschätzungen zu der ihnen vorgelegten EWH-Variante sowie zu EWH im Allgemeinen erhoben wurden. Zu jeder Fragestellung wurde eine Reihe von Aussagen präsentiert, die auf einer vierstufigen Likert-Skala (von „trifft gar nicht zu“ bis „trifft voll zu“) bewertet werden sollten. Am Ende des Fragebogens boten offene Kommentarfelder den Lehrkräften die Möglichkeit, sich zu Vor- und Nachteilen der ihnen vorgelegten EWH-Variante zu äußern sowie Verbesserungsvorschläge, Herausforderungen, Fortbildungswünsche und allgemeine Anmerkungen zu formulieren.

Durchführung

Die Datenerhebung erfolgte Anfang Januar bis Ende Februar 2021. Alle Materialien wurden den Lehrkräften postalisch zugesendet und konnten bei freier Zeiteinteilung bearbeitet werden. Die Postsendung enthielt jeweils acht transkribierte Abiturarbeiten, acht Bewertungsbögen und einen Erwartungshorizont mit Bewertungshinweisen. Die Lehrkräfte wurden gebeten, sich bei der Korrektur und Bewertung der Arbeiten ausschließlich an den beiliegenden Materialien zur Bewertung zu orientieren und den Fragebogen erst zum Schluss auszufüllen. Alle Daten wurden mittels des webbasierten Eingabeinstruments erfasst und getrennt von den personenbezogenen Informationen gespeichert. Die Erfassung und Auswertung der Daten erfolgten vollständig anonymisiert.

4.3 Ergebnisse

Bewertungen der Abiturarbeiten

Nach Abschluss der Datenerhebung lagen für jede der 24 Arbeiten durchschnittlich 30 Bewertungen vor. Abbildung 10 zeigt die Verteilung der erteilten Notenpunkte für die Gesamtleistung in Form von Boxplots. Die mittlere Bewertung der Expert:innen ist für jede Arbeit als Punkt abgebildet. Ähnlich zu den Ergebnissen aus Studie 1 weisen die erteilten Notenpunkte für die Gesamtleistung auch hier eine sehr große Streu-

ung auf. Differenzen in den Bewertungen einer einzelnen Arbeit reichen von minimal 6 bis maximal 12 Notenpunkte. Zwar sind extreme Spannen in der Bewertung nur bei einzelnen Arbeiten zu beobachten, jedoch ist festzuhalten, dass nahezu alle Arbeiten unabhängig von ihrer Position im Leistungsspektrum sehr unterschiedlich bewertet werden. Dieser Befund bleibt bestehen, wenn die Verteilung der erteilten Notenpunkte getrennt nach EWH-Varianten betrachtet wird. Tabelle 5 zeigt die mittlere Bewertung der Gesamtleistung aller Arbeiten insgesamt sowie getrennt nach den EWH-Varianten. In Bezug auf die Unterschiede in der Streuung zwischen den Arbeiten ist zu vermuten, dass die Realisierung bestimmter Kriterien in manchen Schülertexten einfacher mit den in den EWH formulierten Kriterien ins Verhältnis gesetzt werden kann. Welche Faktoren jedoch konkret dazu führten, dass die Diskrepanzen in den Bewertungen bei einigen Arbeiten größer ausfielen als bei anderen, wird Gegenstand zukünftiger, qualitativer Analysen der Prüfungsarbeiten sein.

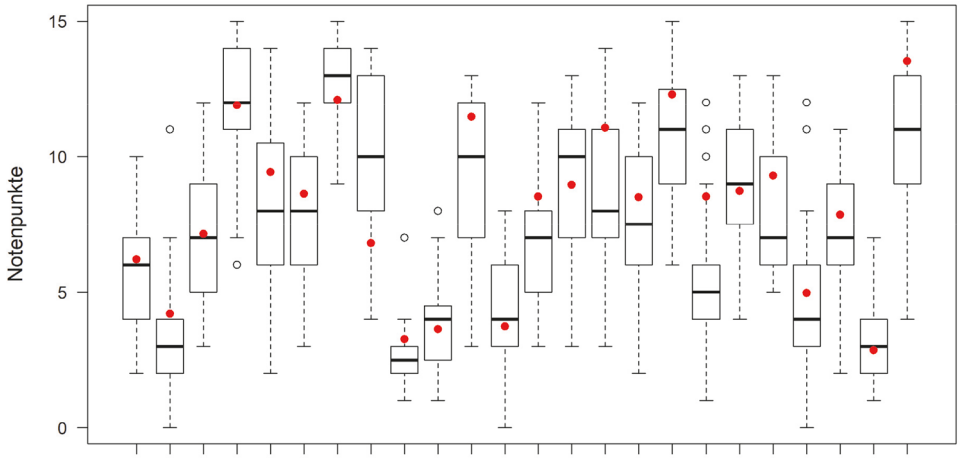


Abbildung 10: Grafische Darstellung des mittleren Expertenurteils (roter Punkt) und der Notenpunkteverteilung in Form von Boxplots für jede Arbeit

Tabelle 5: Mittlere Bewertung der Gesamtleistung aller Arbeiten insgesamt sowie getrennt nach den EWH-Varianten

Variante	*N*	*M*	*SD*	*SE*	*ICC*
insgesamt	91	7.45	3.67	0.14	0.58
EWH 1	17	7.35	3.90	0.33	0.60
EWH 2	18	7.49	3.74	0.32	0.51
EWH 3	18	7.52	3.61	0.30	0.56
EWH 4	19	7.56	3.65	0.30	0.50
EWH 5	19	7.34	3.50	0.28	0.59

Unterschiede zwischen den EWH-Varianten wurden mittels linearer Regressionsanalysen untersucht. In Bezug auf die Höhe der Bewertung der Gesamtleistung zeigte sich hier kein statistisch signifikanter Unterschied zwischen den fünf EWH-Varianten ($F(4,723) = 0.11$, $p = .98$). Diese Befunde lassen darauf schließen, dass keine der in der Studie eingesetzten EWH-Varianten einen Einfluss auf die Anzahl oder die Streuung der erteilten Notenpunkte für die Gesamtleistung hat.

Übereinstimmung der Bewertungen

Der ICC-Wert als Maß für die Übereinstimmung unabhängiger Bewertungen innerhalb einer Gruppe von Lehrkräften, die mittels derselben EWH-Variante bewertet hatte, lag bei allen Varianten zwischen .5 und .6, was als mittlere Übereinstimmung interpretiert werden kann. Als zweites Maß für den Grad der Übereinstimmung zwischen den EWH-Varianten wurde (anders als in Studie 1) der mittlere Abstand der Bewertungen zum Mittelwert der Expertenurteile für jede Arbeit in Notenpunkten gemessen. In Tabelle 6 sind die mittleren Abweichungen vom Mittelwert der Expertenurteile insgesamt und getrennt nach EWH-Variante für alle Arbeiten sowie separat für gute bis sehr gute (10–15 NP), befriedigende (7–9 NP) und ausreichende bis ungenügende Arbeiten (0–6 NP) aufgeführt (Grundlage für die Gruppierung war die Bewertung der Gesamtleistung durch die Expert:innen). Zunächst lässt sich festhalten, dass die Arbeiten von den Lehrkräften durchschnittlich um 0,6 Notenpunkte schlechter eingeschätzt wurden als von den Expert:innen. Dabei spielte es keine Rolle, welche EWH-Variante der Bewertung zugrunde lag ($F(4,713) = 0.21$, $p = .93$). Die Abweichung zu den Expertenurteilen lag bei allen EWH-Varianten im Mittel zwischen 0,5 und 0,8 Notenpunkten. Auffällig ist hierbei, dass die Arbeiten umso strenger bewertet wurden, je besser deren Gesamtleistung von den Expert:innen eingeschätzt worden war ($F(2,713) = 9.93$, $p < .001$). Während die Bewertungen der Arbeiten im unteren Leistungsspektrum fast gar nicht von den Expertenurteilen abwichen, schätzten die Lehrkräfte die Arbeiten im mittleren Leistungsbereich durchschnittlich einen halben Notenpunkt schlechter ein als die Expert:innen. Bei den Arbeiten aus dem oberen Leistungsbereich fielen die Bewertungen der Lehrkräfte durchschnittlich sogar mehr als einen ganzen Notenpunkt strenger aus. Dabei ergab sich jedoch kein statistisch signifikanter Unterschied zwischen den EWH-Varianten ($F(8,713) = 0.62$, $p = .76$). Aus diesen Befunden lässt sich schlussfolgern, dass weder der Differenzierungsgrad der Bewertungskriterien noch die Vorgabe einer Gewichtung von Verstehens- und Darstellungsleistung eine Auswirkung auf die Übereinstimmung der Bewertungen mit den Expertenurteilen haben. Vielmehr scheinen die großen Diskrepanzen in der Bewertung auf Faktoren zurückzugehen, die mit dem subjektiven Verständnis der Bewertungskriterien zusammenhängen.

Tabelle 6: Mittlere Abweichung (inkl. Standardabweichung) der Bewertung der Gesamtleistung von den Expertenurteilen insgesamt sowie getrennt für gute bis sehr gute (10–15 NP), befriedigende (7–9 NP) und weniger befriedigende Arbeiten (0–6 NP)

	alle Arbeiten	0–6 NP	7–9 NP	10–15 NP
insgesamt	-0.63 (2.71)	-0.12 (1.93)	-0.52 (2.91)	-1.34 (2.83)
EWH 1	-0.69 (2.74)	-0.52 (1.79)	-0.54 (3.03)	-1.18 (2.96)
EWH 2	-0.56 (2.94)	-0.05 (2.40)	-0.20 (2.98)	-1.78 (3.09)
EWH 3	-0.57 (2.75)	0.03 (1.71)	-0.50 (3.11)	-1.31 (2.70)
EWH 4	-0.53 (2.49)	0.05 (1.90)	-0.69 (2.79)	-0.78 (2.34)
EWH 5	-0.77 (2.66)	-0.11 (1.80)	-0.65 (2.72)	-1.66 (3.05)

Anfälligkeit für Urteilsverzerrungen

Um die Anfälligkeit der EWH-Varianten für Verzerrungen durch Urteilsheuristiken zu prüfen, wurde in Studie 2 untersucht, welchen Einfluss die jeweilige Bewertungstradition der Lehrkräfte auf die Übereinstimmung bei der Bewertung hat. Aufbauend auf dem Befund aus Studie 1, dass der Nutzen der EWH für die Bewertung der Abiturarbeiten von Lehrkräften höher eingeschätzt wird, wenn das vorliegende Bewertungsmodell der Bewertungstradition des eigenen Landes entspricht, wurden die an Studie 2 teilnehmenden Lehrkräfte gemäß ihrer Landeszugehörigkeit in zwei Gruppen eingeordnet: Lehrkräfte aus Ländern mit einer eher holistischen Bewertungstradition und Lehrkräfte aus Ländern mit einer eher analytischen Vorgehensweise bei der Bewertung. Da bei der Zuweisung der Studienteilnehmer:innen darauf geachtet wurde, Lehrkräfte aus Ländern mit verschiedenen Bewertungstraditionen gleichmäßig auf die EWH-Varianten zu verteilen, ergibt sich eine relativ ausgeglichene Verteilung der vertretenen Traditionen pro EWH-Variante. Tabelle 7 (linke Seite) enthält die mittlere Bewertung (in Notenpunkten) insgesamt sowie separat für jede EWH-Variante getrennt nach Bewertungstradition. Über alle EWH-Varianten hinweg zeigte sich ein statistisch signifikanter Unterschied zwischen den beiden Gruppen ($t(718) = -3.17$, $p = .002$). Im Mittel vergaben Lehrkräfte aus Ländern mit holistischer Tradition 0,8 Notenpunkte weniger als Lehrkräfte aus Ländern mit analytischer Tradition. Dieser Unterschied ist jedoch vollständig auf die Nutzung von EWH-Variante 1 zurückzuführen, die bei Lehrkräften aus Ländern mit holistischer Tradition zu Bewertungen führte, die durchschnittlich fast zwei Notenpunkte strenger waren als die Einschätzungen von Lehrkräften aus Ländern mit analytischer Tradition ($t(718) = -2.97$, $p = .003$). Auch unter Nutzung von EWH-Variante 2 vergaben Lehrkräfte aus holistisch geprägten Ländern durchschnittlich über einen Notenpunkt weniger für dieselben Arbeiten als ihre Kolleg:innen aus analytisch geprägten Ländern. Dieser Unterschied war jedoch statistisch nicht signifikant ($t(718) = 0.81$, $p = .42$).

Im Hinblick auf die mittlere Abweichung der Bewertung der Gesamtleistung von den Expertenurteilen (Tabelle 7, rechte Seite) zeigt sich, dass die Übereinstimmung mit den Bewertungen der Expert:innen bei Lehrkräften aus Ländern mit holistischer

Tradition im Mittel geringer ausfällt als bei Lehrkräften aus Ländern mit analytischer Tradition ($F(1,718) = 9.78$, $p = .002$). Besonders unter Nutzung von EWH-Variante 1 ergab sich hier eine große Diskrepanz in den Bewertungen zwischen den Gruppen ($t(718) = -3.66$, $p = <.001$). Diese Ergebnisse zeigen, dass eine Bewertung der Arbeiten unter Nutzung von EWH-Variante 1 nicht unabhängig davon ist, welches Bewertungsmodell die Lehrkräfte in der Praxis anwenden. Die Differenzierung von Bewertungskriterien sowie die Vorgabe einer Gewichtung der Teilleistungen scheint die Wirkung solcher Urteilsheuristiken hingegen abzuschwächen. Dieses Ergebnis deckt sich mit dem Befund aus anderen Studien, dass holistische Bewertungen häufig etwas anfälliger für Urteilsverzerrungen sind als analytische Bewertungen (z. B. Böhme et al., 2009; Jonsson & Svingby, 2007; Kreuzer, 2018).

Tabelle 7: Mittlere Bewertung der Gesamtleistung (links) und mittlere Abweichung von den Expertenurteilen (rechts) insgesamt sowie separat für jede EWH-Variante getrennt nach Bewertungstradition (in NP)

	mittlere Bewertung		mittlere Abweichung	
	analytisch	holistisch	analytisch	holistisch
insgesamt	7.99 (3.75)	7.12 (3.58)	-0.23 (2.46)	-0.88 (2.83)
EWH 1	8.46 (3.99)	6.58 (3.67)	0.32 (2.17)	-1.40 (2.89)
EWH 2	8.14 (3.65)	6.96 (3.75)	-0.20 (2.67)	-0.85 (3.13)
EWH 3	7.88 (3.72)	7.30 (3.53)	-0.34 (2.82)	-0.72 (2.71)
EWH 4	7.50 (3.86)	7.59 (3.54)	-0.62 (2.20)	-0.48 (2.65)
EWH 5	7.94 (3.59)	7.07 (3.44)	-0.30 (2,34)	-0.98 (2.79)

Anmerkung: Die Verteilung der vertretenen Bewertungstraditionen pro EWH-Variante variiert zwischen 41 % analytisch und 59 % holistisch (EWH 1), 44 % und 56 % (EWH 2), 39 % und 61 % (EWH 3), 37 % und 63 % (EWH 4), 32 % und 68 % (EWH 5).

Nutzung und Akzeptanz der EWH

Die Einschätzung der Praktikabilität der EWH-Varianten wurde separat in Bezug auf die Bewertung der Verstehens- und Darstellungsleistung erhoben. Die in Abbildung 11 dargestellten Angaben der Lehrkräfte zu ausgewählten Multiple-Choice-Items zeigen, dass es in Bezug auf die empfundene Praktikabilität des EWH für die Bewertung der Verstehensleistung keine Unterschiede zwischen den EWH-Varianten gab. Für die Bewertung der Darstellungsleistung hingegen bewerteten Lehrkräfte die EWH-Version 1 als etwas weniger hilfreich ($t(4,86) = -2.44$, $p = .02$) und schätzten sowohl den Grad der Detailliertheit ($t(4,84) = -3.42$, $p = <.001$) als auch die Klarheit der Formulierungen ($t(4,86) = -3.04$, $p = .003$) weniger positiv ein als Lehrkräfte, denen eine andere EWH-Version vorgelegen hatte. Diese Ergebnisse legen den Schluss nahe, dass eine Differenzierung der Bewertungskriterien und die Vorgabe einer Gewichtung von Teilleistungen unabhängig von der jeweiligen Ausgestaltung (EWH-Varian-

ten 2 bis 5) dazu führen, dass sich Lehrkräfte bei der Bewertung von Prüfungsarbeiten besser durch den EWH unterstützt fühlen.

Bestätigt wird diese Schlussfolgerung durch Angaben der Lehrkräfte zu ihren Vorstellungen von einer idealen Gestaltung der EWH. Unabhängig von der verwendeten EWH-Version und der eigenen Bewertungstradition sprachen sich alle Lehrkräfte dafür aus, dass sowohl für die Bewertung der Verstehensleistung als auch für die Bewertung der Darstellungsleistung klare Kriterien im EWH vorgegeben sein sollten (Abbildung 12, links). Die Zustimmung zu der Aussage, dass die Prüfungsarbeit eher auf der Grundlage einer Gesamtwürdigung der erbrachten Leistung und ohne die Ausweisung des Grads der Erfüllung von Kriterien bewertet werden sollte, fiel über alle EWH-Varianten hinweg gering aus. Auch hier gab es keinen Unterschied zwischen den Einschätzungen von Lehrkräften aus Ländern mit analytischer und holistischer Bewertungstradition (Abbildung 12, rechts). Im Hinblick auf die Gewichtung von Teilleistungen sprachen sich alle Lehrkräfte in der Tendenz für eindeutige prozentuale Vorgaben zur Gewichtung von Verstehens- und Darstellungsleistung aus (Abbildung 13, links). Lehrkräfte aus analytisch geprägten Ländern befürworteten dies jedoch stärker als Lehrkräfte aus holistisch geprägten Ländern ($t(9{,}90) = -2.62$, $p = .01$). Nur wenige Studienteilnehmer:innen stimmten der Aussage zu, dass die Gewichtung der Kriterien zur Bewertung den Bewerter:innen überlassen bleiben sollte (Abbildung 13, rechts). Auch hier zeigte sich allerdings, dass Lehrkräfte aus Ländern mit holistischer Bewertungstradition diese Überzeugung eher vertreten als Lehrkräfte, denen das analytische Bewertungsmodell vertrauter ist ($t(9{,}79) = -2.24$, $p = .03$).

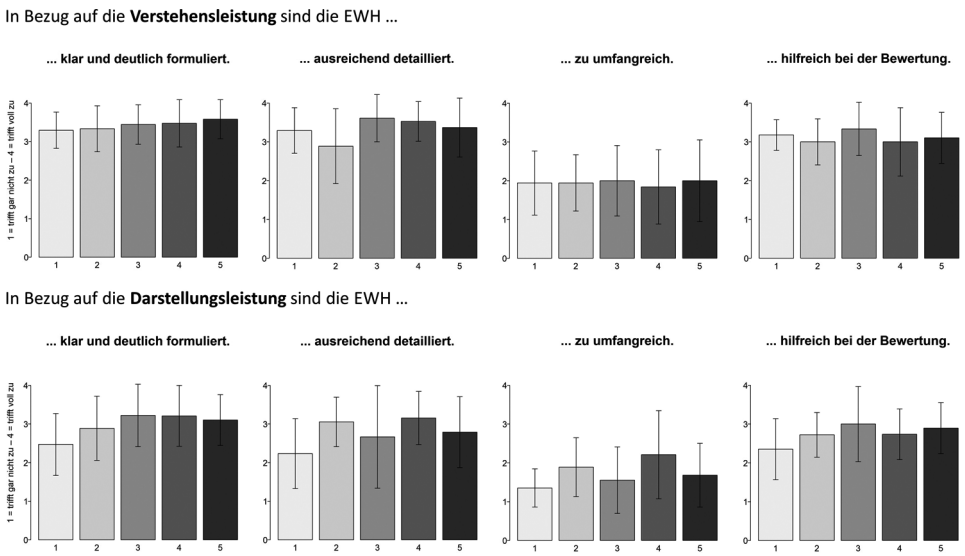


Abbildung 11: Mittelwerte (inkl. Standardabweichung) für die Angaben der Lehrkräfte bei ausgewählten Fragebogenitems zur Praktikabilität der EWH-Varianten bei der Bewertung

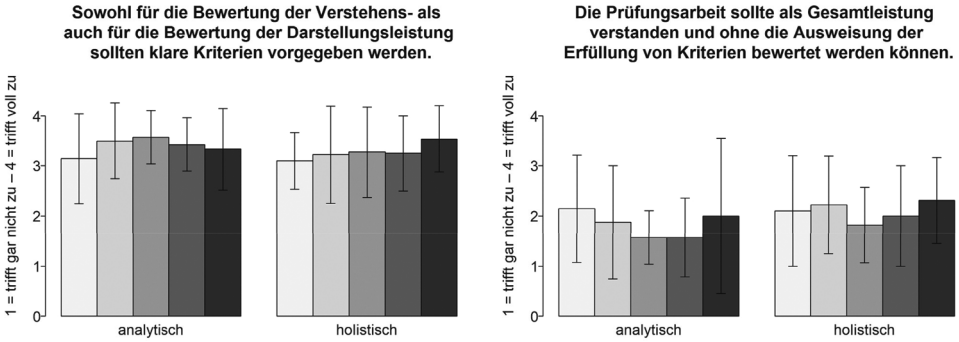


Abbildung 12: Mittelwerte (inkl. Standardabweichung) für die Angaben der Lehrkräfte zur Ausweisung von Bewertungskriterien, getrennt nach Bewertungstradition

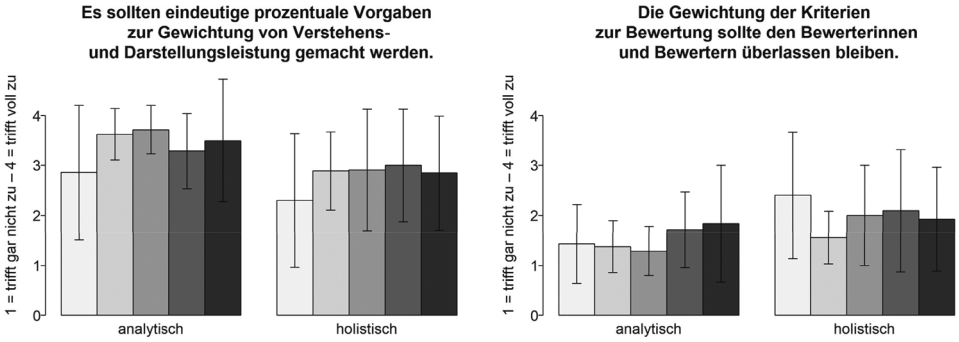


Abbildung 13: Mittelwerte (inkl. Standardabweichung) für die Angaben der Lehrkräfte zur Gewichtung von Verstehens- und Darstellungsleistung, getrennt nach Bewertungstradition

4.4 Schlussfolgerungen

Basierend auf den vorliegenden Befunden lässt sich kein eindeutiger Vorteil für den Einsatz einer der untersuchten EWH-Varianten bei der Bewertung von Abiturarbeiten im Fach Deutsch feststellen. Keine der in der Studie untersuchten fünf EWH-Varianten führte zu einer Reduktion der Streuung der erteilten Notenpunkte oder zu einer höheren Übereinstimmung der Bewertungen mit den Expertenurteilen. Dies lässt vermuten, dass sich Unterschiede in der Bewertung der Arbeiten weniger aus der Gestaltung der EWH ergeben und vielmehr auf Unterschiede in der subjektiven Einschätzung der Bewertungskriterien bzw. auf den individuellen Umgang mit diesen Kriterien durch die Lehrkräfte zurückzuführen sind.

Die durchschnittliche Bewertung der Abiturarbeiten lag bei allen EWH-Varianten unter dem Urteil der Expert:innen, wobei Lehrkräfte aus Ländern mit holistischer Bewertungstradition die Abiturarbeiten im Durchschnitt strenger bewerteten als ihre Kolleg:innen aus Ländern mit analytischer Tradition. Dieser Unterschied ist jedoch auf die Nutzung von EWH-Variante 1 zurückzuführen, die bei Lehrkräften aus holistisch geprägten Ländern zu einer um fast zwei Notenpunkte schlechteren Bewertung führte als bei Lehrkräften aus Ländern mit analytischer Tradition. Auch die mittlere

Abweichung von den Expertenurteilen war mit Abstand am größten, wenn Lehrkräfte aus Ländern mit holistischer Bewertungstradition zur Bewertung der Abiturarbeiten die EWH-Variante 1 nutzten. Aus diesen Ergebnissen lässt sich schlussfolgern, dass ein EWH ohne differenzierte Bewertungskriterien und ohne Vorgabe einer Gewichtung von Teilleistungen anfälliger für Verzerrungen durch Urteilsheuristiken ist. Dieser Befund repliziert die durch Studie 1 und andere Studien gewonnene Erkenntnis, dass holistische Bewertungen eine höhere Anfälligkeit dafür aufweisen, durch für die Beurteilung der Leistung irrelevante Faktoren beeinflusst zu werden (Böhme et al. 2009; Ingenkamp, 1995).

Auch anhand der Einschätzung der EWH durch die Lehrkräfte lässt sich ein Nachteil für EWH-Variante 1 im Vergleich zu den anderen Varianten feststellen: Für die Bewertung der Darstellungsleistung wurde sowohl der Grad der Detailliertheit als auch die Klarheit der Formulierungen in EWH-Variante 1 von den Lehrkräften als weniger ausreichend und der EWH insgesamt als weniger hilfreich bewertet. Angaben der Lehrkräfte zeigten weiterhin, dass unabhängig von der eigenen Bewertungstradition die Vorgabe klarer Bewertungskriterien für die Verstehens- und Darstellungsleistung im EWH befürwortet wird und Prüfungsarbeiten durch die Ausweisung der Erfüllung vorgegebener Kriterien bewertet werden sollten. Alle Lehrkräfte sprachen sich tendenziell für eindeutige Vorgaben zur Gewichtung von Verstehens- und Darstellungsleistung aus, wobei die Zustimmung von Lehrkräften aus analytisch geprägten Ländern hier höher war als die von Lehrkräften aus Ländern mit holistischer Bewertungstradition.

Zusammenfassend lässt sich festhalten, dass sich zwischen den verschiedenen Modellen der Differenzierung und Gewichtung der Verstehens- und Darstellungsleistung im EWH zwar keine Vorteile finden, aber deutliche Nachteile für das Fehlen solcher Vorgaben festgestellt werden konnten. Dies legt den Schluss nahe, dass eine ländergemeinsame, einheitliche Regelung zur Bewertung von schriftlichen Prüfungsleistungen im Fach Deutsch sowohl differenzierte Bewertungskriterien als auch eine Gewichtung der beiden Leistungsbereiche vorsehen sollte. Zusätzlich sollte ein einheitliches Verständnis vorgegebener Bewertungskriterien sichergestellt werden.

Fazit

Die Ergebnisse beider Studien weisen nicht darauf hin, dass eines der untersuchten Bewertungsmodelle (holistisch/analytisch) bzw. eine der untersuchten EWH-Varianten (1 bis 5) entscheidende Vorteile in Bezug auf die Übereinstimmung der Bewertungen sowie auf die subjektive Akzeptanz der Lehrkräfte hat. Empirische Belege fanden sich jedoch für die Bedeutung hinreichend differenzierter Bewertungskriterien sowie für die wichtige Rolle der Gewichtung von Verstehens- und Darstellungsleistung. Beides wird in den Bewertungsmodellen zahlreicher Länder bereits verwirklicht. Des Weiteren wurde in beiden Studien deutlich, dass es unabhängig vom implementierten Bewertungsmodell geboten ist, die Objektivität der Bewertung von Abiturprü-

fungsarbeiten durch geeignete Maßnahmen zu erhöhen. Konkrete Aussagen der bildungswissenschaftlichen und fachdidaktischen Forschung dazu, wie eine objektive Bewertung der Bearbeitung von offenen, komplexen Aufgaben, wie sie durch die Bildungsstandards im Fach Deutsch für die Allgemeine Hochschulreife vorgegeben sind, mithilfe des EWH angeleitet werden kann, liegen jedoch (noch) nicht vor (Kötter-Mathes, 2020).

Die Abituraufgaben einschließlich der Vorgaben zur Bewertung werden in nahezu allen Ländern zentral gestellt, wohingegen die Anwendung der Erwartungshorizonte und die Umsetzung der Bewertungshinweise letztlich in der Hand der korrigierenden Lehrkraft liegen. Die Unterschiede bei der Bewertung derselben Arbeit durch mehrere Lehrkräfte – unabhängig von der verwendeten EWH-Variante – scheinen sich aus individuellen Unterschieden bei der Interpretation und Einschätzung der zugrunde liegenden Kriterien zu ergeben. Die Herausforderungen bei der Bewertung von Verstehens- und Darstellungsleistungen werden möglicherweise unterschätzt, sodass entsprechende Kompetenzen eher selten explizit gefördert werden. Ansatzpunkte, um die Vergleichbarkeit der Bewertungen von Prüfungsarbeiten zu erhöhen, könnten demnach in der Lehrkräftebildung sowie in der berufsbegleitenden Förderung der Bewertungskompetenz von Lehrkräften liegen. Ergebnisse aus Studien, die die Wirkung von Beurteilungstrainings auf die Beurteilungsgenauigkeit untersucht haben, stützen diese Überlegungen. So konnten Fahim und Bijani (2011) zeigen, dass ein Training zur Anwendung der EWH die im Vorfeld identifizierte Urteilsverzerrung sowie die Strenge der Bewerter:innen reduzieren konnte. Beurteilungstrainings scheinen sich jedoch hauptsächlich bei unerfahrenen Personen auszuwirken (Greatorex & Bell, 2008).

Eine besondere Herausforderung stellt die Bewertung im Bereich der Darstellungsleistung dar, was z. B. am Kriterium der Verwendung einer adäquaten und abwechslungsreichen Wortwahl deutlich wird. Es ist nicht möglich, diese Teilleistung eindeutig mit den Kategorien „richtig“ und „falsch“ zu bewerten (Baurmann & Dehn, 2004). Stattdessen müssen differenziertere Kategorien für die Einschätzung herangezogen werden, z. B. inwieweit die Wortwahl hinsichtlich des Adressatenkreises oder der Textfunktion angemessen ist. Die Bewertung und Benotung sprachlicher Leistungen anzuleiten, ist entsprechend anspruchsvoll. Deshalb wäre es wichtig, in einem kontinuierlichen Reflexionsprozess, z. B. in der Fachkonferenz oder in Fortbildungen, zu diskutieren, wie sprachliche Äußerungen bewertet und welche Kriterien hierzu verwendet werden. Damit die Kriterien der jeweiligen Bewertungsvorgaben für den praktischen Einsatz konkretisiert werden können, sollte ermittelt werden, in Bezug auf welche Aspekte Dissens besteht und wie ein gemeinsames Verständnis der Lehrkräfte erreicht werden kann.

Ähnliche Herausforderungen zeigen sich auch bei der Bewertung der Verstehensleistung. So müssen die Lehrkräfte nicht nur das Textverständnis der Schüler:innen erfassen, sondern zudem auch einschätzen, in welchem Verhältnis dieses Textverständnis zu den im EWH genannten Kriterien steht. Zudem ist, wie bei der Bewertung der Darstellungsleistung, nicht zwingend davon auszugehen, dass die im EWH

genannten Kriterien von den Lehrkräften in ähnlicher Weise aufgefasst werden. Dies betrifft sowohl Kriterien, die sich auf thematische Aspekte bzw. Deutungsmöglichkeiten der Texte beziehen, als auch Kriterien, welche die sprachliche Gestaltung eines Textes, insbesondere aber die Gewichtung und Relationierung dieser Aspekte beschreiben.

Insgesamt lässt sich festhalten, dass dem Bereich der professionellen Bewertungskompetenz von Deutschlehrkräften – nicht nur im Rahmen der Abiturprüfungen – zukünftig deutlich mehr Aufmerksamkeit gewidmet werden sollte. Dies gilt zum einen in Hinblick auf die Schulung von Kompetenzen, die für die differenzierte Einschätzung von Verstehens- und Darstellungsleistung grundlegend sind, zum anderen aber auch hinsichtlich des kontinuierlichen Austausches zu Fragen der Anwendbarkeit von Bewertungskriterien.

Literatur

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice, 18*(3), 279–293.

Baurmann, J. & Dehn, M. (2004). Beurteilen im Deutschunterricht. *Praxis Deutsch, 184*, 6–13.

Birkel, P. & Birkel, C. (2002). Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? *Psychologie in Erziehung und Unterricht, 49*(3), 219–224.

Böhme, K., Bremerich-Vos, A. & Robitzsch, A. (2009). Aspekte der Kodierung von Schreibaufgaben. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 290–329). Weinheim: Beltz.

Bouwer, R., Koster, M. & Van den Bergh, H. (2016). *Benchmark Rating Procedure: Best of Both Worlds. Comparing Procedures to Rate Text Quality in a Reliable and Valid Manner*. https://dspace.libraryuu.nl/bitstream/1874/338041/1/bouwerkoster.pdf [20.06.2022]

Bramley, T. (2007). Mark scheme features associated with different levels of marker agreement. Verfügbar unter: http://www.cambridgeassessment.org.uk/Images/109770-mark-scheme-features-associated-with-different-levels-of-marker-agreement.pdf [26.07.2021]

Cumming, A., Kantor, R. & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67–96.

Disdorn-Liesen, V. (2016). *Vergleichbarkeit in der Vielfalt. Leistungsanforderungen und Leistungsfeststellung im Zentralabitur Deutsch*. Wiesbaden: Springer VS.

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation*. Wiesbaden: Springer Verlag.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing, 14*(2), 88–115.

Elliot, N. (2005). *On a Scale: A social history of writing assessment in America*. New York: Peter Lang.

Elliot, N. & Haswell, R. (2019). *Early Holistic Scoring of Writing: A Theory, a History, a Reflection*. Louisville, CO: University Press of Colorado.

Fahim, M. & Bijani, H. (2011). The Effects of Rater Training on Raters' Severity and Bias in Second Language Writing Assessment. *Iranian Journal of Language Testing, 1*(1), 1–16.

Greatorex, J. & Bell, J. F. (2008). What makes AS marking reliable? An experiment with some stages from the standardisation process. *Research Papers in Education, 23*(3), 333–355.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E. & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment, 12*(1), 19.

Grzesik, J. & Fischer, M. (1984). *Was leisten Kriterien für die Aufsatzbeurteilung? Theoretische, empirische und praktische Aspekte des Gebrauchs von Kriterien und der Mehrfachbeurteilung nach globalem Ersteindruck*. Opladen: Westdeutscher Verlag.

Hamp-Lyons, L. (2016). Farewell to Holistic Scoring? *Assessing Writing* 27, A1–A2.

Harsch, C. & Martin, G. (2013). Comparing holistic and analytic scoring methods: issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice, 20*(3), 281–307.

Hartmann, W. (1989). Die Hamburger Aufsatzstudie. *Der Deutschunterricht, 41*(3), 92–98.

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

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130–144.

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., Comfort, K. & Othman, A. R. (1998). Analytic Versus Holistic Scoring of Science Performance Tasks. *Applied Measurement in Education, 11*(2), 121–137.

KMK (2012). *Bildungsstandards im Fach Deutsch für die Allgemeine Hochschulreife*. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK). Köln: Wolters Kluwer.

Koo, T. K. & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine, 15*(2), 155–163.

Köster, J. (2006). Das Deutschabitur in Zeiten von Bildungsstandards – Vergleichbarkeit der Prüfungsleistungen und ihre Bedeutung. *Didaktik Deutsch* 21, 78–90.

Kötter-Mathes, S. (2020). *Leistungsbeurteilung in zentralen Prüfungen. Lehrkräftewahrnehmungen der landesweit vorgegebenen Erwartungshorizonte im Prüfungsfach Deutsch*. Wiesbaden: Springer VS.

Kreuzer, P. (2018). *Kriterienraster* (Handreichung der Prüfungswerkstatt). Verfügbar unter: https://www.zq.uni-mainz.de/files/2018/08/6_Kriterienraster-erstellen.pdf [26.06.2021]

Lehmann, R. H. (1988). Reliabilität und Generalisierbarkeit der Aufsatzbeurteilung im Rahmen des Hamburger Beitrags zur internationalen Aufsatzstudie der IEA. Empirische Pädagogik, *Zeitschrift zu Theorie und Praxis erziehungswissenschaftlicher Forschung, 2*(4), 349–365.

Lindauer, N, & Sommer, T. (2018). Verfahren der Textbeurteilung. Merkmale und Vorzüge eines holistischen Benchmarkratings. *Leseräume: Zeitschrift für Literalität in Schule und Forschung, 5*, 1–14.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246–276.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Neumann, A. (2007a). *Briefe schreiben in Klasse 9 und 11. Beurteilungskriterien, Messungen, Textstrukturen und Schülerleistungen*. Münster: Waxmann.

Neumann, A. (2007b). Schreiben. Ausgangspunkt für eine kriteriengeleitete Ausbildung in der Schule. In H. Willenberg (Hrsg.), *Kompetenzhandbuch für den Deutschunterricht. Auf der empirischen Basis des DESI-Projekts* (S. 74–83). Baltmannsweiler: Schneider.

Rezaei, A. R. & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18–39.

Reznitskaya, A., Kuo, L.-j., Glina, M., & Anderson, R. C. (2009). Measuring argumentative reasoning: What's behind the numbers? *Learning and Individual Differences, 19*(2), 219–224.

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159–179

Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modeling. *Language Testing, 22*(1), 1–5

Shabani, E. A. & Panahi, J. (2020). Examining consistency among different rubrics for assessing writing. *Language Testing in Asia, 10*(1), 1–25.

Silvestri, L. & Oescher, J. (2006). Using rubrics to increase the reliability of assessment in health classes. *International Electronic Journal of Health Education, 9*, 25–30.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*, 25–29.

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.

Van den Bergh, H., De Maeyer, S., Van Weijen, D. & Tillema, M. (2012). Generalizability of Text Quality Scores. In: E. Van Steendam, M. Tillema, G. Rijlaarsdam, H. Van den Bergh (Hrsg.), *Measuring Writing: Recent Insights Into Theory, Methodology and Practices* (S. 23–32). Leiden: Brill.

Zabka, T. & Stark, T. (2010). Aufgabenstellungen und Erwartungshorizonte als Steuerungsinstrumente. Zum Umgang mit Problemen der Literaturinterpretation im Zentralabitur. *Der Deutschunterricht, 62*(1), 19–29.

Zlatkin-Troitschanskaia, O., Shavelson, R. J., Schmidt, S. & Beck, K. (2019). On the complementarity of holistic and analytic approaches to performance assessment scoring. *British Journal of Educational Psychology, 89*(3), 468–484.

Abbildungsverzeichnis

Tabellenverzeichnis

9 Und wenn man die Abiturprüfungen einfach ausfallen ließe? Empirische Befunde zu Unterschieden zwischen Abiturprüfungsnoten und Kursnoten

Lars Hoffmann, Nicolas Hübner, Marko Neumann & Pauline Schröter

Zusammenfassung

Im vorliegenden Beitrag werden die Ergebnisse von drei Studien dargestellt, die jeweils die Passung zwischen Abiturprüfungsnoten auf der einen Seite und den in der Qualifikationsphase der gymnasialen Oberstufe erreichten Kursnoten auf der anderen Seite anhand unterschiedlicher Datensätze (d. h. Daten aus der Evaluation zur Bewährung der Poolaufgaben, Daten aus einer Zusatzstudie des Nationalen Bildungspanels, Daten aus der BERLIN-Studie) und mit jeweils anderen Schwerpunktsetzungen untersucht haben. Insgesamt verdeutlichen die Ergebnisse der drei Studien, dass die in den Abiturprüfungen erzielten Noten im Mittel etwas schlechter ausfallen als die Kursnoten der Qualifikationsphase. Darüber hinaus zeigen die Befunde, dass die Passung zwischen Prüfungsnoten und Kursnoten erheblich zwischen Ländern, Schulen und Fächern variiert und möglichweise auch von schulsystemischen Änderungen (z. B. Oberstufenreformen) beeinflusst wird. Die Ergebnisse werden im Beitrag auch vor dem Hintergrund der in der Corona-Pandemie verschiedentlich diskutierten Möglichkeit eines Anerkennungsabiturs unter Verzicht auf die Abiturprüfung beleuchtet.

Einleitung

Während der ersten und dritten Welle der Corona-Pandemie im Frühling der Jahre 2020 und 2021 wurde von verschiedenen Seiten – wie etwa vom Ministerium für Bildung, Wissenschaft und Kultur des Landes Schleswig-Holstein oder von der Gewerkschaft für Erziehung und Wissenschaft (GEW) – vorgeschlagen, auf die Durchführung schriftlicher Abiturprüfungen zu verzichten.[1] Begründet wurde dieser Vorschlag vor allem mit den außergewöhnlichen Umständen, die aus den Maßnahmen zur Bekämpfung der Pandemie für das Lehren und Lernen an Schulen resultierten. Im Vergleich zu den Vorjahren war die Vorbereitung auf die schriftlichen Abiturprüfungen für viele Schülerinnen und Schüler durch Schulschließungen, Einschränkungen des

1 https://www.schleswig-holstein.de/DE/Landesregierung/III/Presse/PI/2020/Maerz_2020/III_abschlusspruefungen_2019_20.html [02.06.2021]; https://www.tagesschau.de/inland/abitur-gew-101.html [18.06.2021].

Präsenzunterrichts und Distanzlernen deutlich erschwert. Als Alternative wurde für ein „Anerkennungsabitur" geworben, bei dem sich die Abiturgesamtnote ausschließlich aus den bewerteten Leistungen zusammengesetzt hätte, die in der Qualifikationsphase, also den letzten beiden Jahren der gymnasialen Oberstufe, erreicht wurden. Historisch erinnert das vorgeschlagene Vorgehen an die Praxis des „Not- oder Kriegsabiturs" zuzeiten der beiden Weltkriege, als wehrtüchtige junge Männer das Abitur nach einer in Umfang und Anforderungsniveau deutlich reduzierten Prüfung oder sogar ohne Prüfung ablegten (Bölling, 2010).

Letzten Endes beschloss die Kultusministerkonferenz (KMK) jedoch, die Abiturprüfungen in den Prüfungsjahren 2020 und 2021 stattfinden zu lassen.[2] Um den Folgen der Corona-Pandemie für die Prüfungsvorbereitung Rechnung tragen zu können, wurden den Ländern aber verschiedene Möglichkeiten eingeräumt, die Rahmenbedingungen und Regelungen der Abiturprüfungen in dem aus ihrer Sicht erforderlichen Maße anzupassen.[3] Diese Möglichkeiten, von denen die Länder in unterschiedlichem Maße Gebrauch machten, umfassten die Option, auf den Einsatz von Aufgaben aus den Gemeinsamen Abituraufgabenpools der Länder zu verzichten. Ferner konnten die Länder ihre Prüfungen auf einen späteren Zeitpunkt verschieben, nicht wenige Länder boten den Prüflingen sogar verschiedene Prüfungstermine zur Auswahl an. Einige Länder veränderten auch die Vorgaben für die Abiturprüfungen selbst, indem sie Schulen, Lehrkräften und Prüflingen mehr Wahlmöglichkeiten bei den Prüfungsaufgaben einräumten oder die für die Prüfungsaufgaben zur Verfügung stehende Bearbeitungszeit erhöhten. Darüber hinaus wurde in einigen Ländern die Abiturprüfungsnote für bestimmte Fächer angehoben.[4]

Doch warum wurde der Vorschlag zum „Anerkennungsabitur" im Pandemiefrühling der Jahre 2020 und 2021 nicht weiterverfolgt? Weshalb sprachen sich sowohl die Bildungsministerien der meisten Länder als auch Expertinnen und Experten aus der Bildungsforschung[5] und sogar der Bundeselternrat[6] sowie die meisten Landesschülervertretungen[7] gegen diesen Vorschlag aus? Die Gegner des Vorschlags argumentierten insbesondere, dass ein Anerkennungsabitur durch deutlich bessere Abiturgesamtnoten als in den vorherigen Jahrgängen gekennzeichnet wäre. Das Abitur der Jahre 2020 und 2021 wäre folglich nur bedingt mit dem der Vorjahre vergleichbar. Dementsprechend wurde die Sorge geäußert, dass ein Verzicht auf die Durchführung von Abiturprüfungen eine Entwertung des Abiturs zur Folge hätte, den Absolventen der betreffenden Jahrgänge also stets der Makel eines „geschenkten" Abiturs anheften

2 https://www.kmk.org/presse/pressearchiv/mitteilung/detail/News/abschlusspruefungen-finden-auch-2021-statt.html [02.06.2021]; https://www.kmk.org/presse/pressearchiv/mitteilung/detail/News/kmk-pruefungen-finden-wie-geplant-statt.html [02.06.2021].

3 Vgl. hierzu die Beschlüsse der KMK vom 25.03.2020 (https://www.kmk.org/presse/pressearchiv/mitteilung/detail/News/kmk-pruefungen-finden-wie-geplant-statt.html) sowie vom 21.01.2021 (https://www.kmk.org/presse/pressearchiv/mitteilung/detail/News/abschlusspruefungen-finden-auch-2021-statt.html) [28.03.2022].

4 https://www.ndr.de/nachrichten/mecklenburg-vorpommern/Schwaches-Mathe-Abi-Zwei-Extra-Punkte-vom-Land,coronavirus5350.html [16.08.2021].

5 https://www.br.de/nachrichten/wissen/kultusminister-nach-abi-streit-einig-pruefungen-finden-statt,RuEoVo5 [02.06.2021].

6 https://www.welt.de/politik/deutschland/article223792940/Schulen-Corona-Die-grosse-Angst-vor-dem-Notabitur.html [22.06.2016].

7 https://www1.wdr.de/nachrichten/themen/coronavirus/corona-abitur-pruefungen-100.html [02.06.2021].

könnte. In diesem Sinne wurde vermutet, dass die Absolventinnen und Absolventen eines Not-Abiturs für ihren weiteren Bildungsweg und den zukünftigen beruflichen Werdegang systematische Nachteile gegenüber den Abiturientinnen und Abiturienten anderer Jahrgänge zu befürchten hätten.

Aus einer historischen Perspektive erscheinen diese Befürchtungen nicht vollkommen aus der Luft gegriffen. So standen etwa die jungen Männer, die während der Zeit des zweiten Weltkrieges ein Abitur mit Reifevermerk erhielten, nach dem Krieg häufig vor dem Problem, dass dieses „Notabitur" nicht anerkannt wurde und sie stattdessen ein zweites Mal Abitur machen mussten (Bölling, 2010). Doch wie sind die oben skizzierten Argumente gegen ein Anerkennungsabitur aus der heutigen Perspektive, also vor dem Hintergrund der geltenden Abiturbestimmungen und der empirischen Befundlage zu bewerten?

Die „Feststellung der Gesamtqualifikation", und mithin auch die Berechnung der Abiturgesamtnote, ist länderübergreifend in der „Vereinbarung zur Gestaltung der gymnasialen Oberstufe und der Abiturprüfung" (KMK, 2021) geregelt. Demnach werden die in den vier Halbjahren der Qualifikationsphase erzielten Kursnoten (Block I) in einem Verhältnis von 2 zu 1 mit den Ergebnissen der Abiturprüfungen (Block II) verrechnet. In Block I können dabei maximal 600 Punkte, in Block II höchstens 300 Punkte erreicht werden. Die Summe der von einem Prüfling in den Blöcken I und II erzielten Punkte werden mittels einer Formel in die Abiturgesamtnote umgerechnet, wobei 180 Punkte einer Notenstufe entsprechen. Führt man sich vor Augen, dass zur Bestimmung der Punkte für Block II die in den Abiturprüfungen erzielten Leistungen je nach Anzahl der Prüfungsfächer jeweils mit dem Faktor vier oder fünf multipliziert werden, so wird deutlich, welches Gewicht die Abiturprüfungen an der Abiturgesamtnote haben. Vor dem Hintergrund dieses Gewichts erscheint die Befürchtung, dass ein Anerkennungsabitur ohne Prüfungen subjektiv als weniger wertvoll wahrgenommen wird, zweifellos nachvollziehbar. Weniger klar ist hingegen die Frage, ob ein Verzicht auf die Durchführung von Abiturprüfungen tatsächlich erheblich bessere Abiturgesamtnoten zur Folge haben würde.

Die Befürchtung einer „Noteninflation" beim Anerkennungsabitur impliziert die Annahme, dass Schülerinnen und Schüler in der Qualifikationsphase, also etwa im Rahmen von Klausuren, Kurzarbeiten oder mündlichen Leistungskontrollen, in der Regel besser abschneiden als in den Abiturprüfungen. Aus einer theoretischen Perspektive erscheint diese Annahme zunächst durchaus plausibel: Insbesondere die schriftlichen Abiturprüfungen sind in Umfang und Niveau der zu erfüllenden Anforderungen deutlich komplexer als die schriftlichen und mündlichen Leistungskontrollen, die Schülerinnen und Schüler im Rahmen der Qualifikationsphase bewältigen müssen, und dabei am ehesten noch mit Klausuren vergleichbar, die gegen Ende der gymnasialen Oberstufe geschrieben werden und zum Teil sogar als „Probeabitur" angelegt sind. Darüber hinaus ist zu bedenken, dass die Abiturprüfungen im Leben vieler Schülerinnen und Schüler die ersten Prüfungen dieser Art darstellen. Die Prüflinge stehen also nicht nur vor der Herausforderung, die Abiturprüfungsaufgaben kompetent zu lösen, sondern müssen sich außerdem den ihnen in diesem Maße bis-

lang noch kaum bekannten Schwierigkeiten und Anforderungen von Prüfungssituationen stellen. Nicht zuletzt dürften sie beim Abitur in besonderem Maße mit potenziell leistungshemmenden Faktoren wie Leistungsdruck, Stress, Aufregung oder sogar Prüfungsangst konfrontiert sein (Pekrun & Götz, 2006).

Indessen lassen sich auch Argumente finden, die der Annahme widersprechen, dass Abiturprüfungsnoten in der Regel schlechter ausfallen als die Kursnoten. So ist etwa zu bedenken, dass die meisten Schülerinnen und Schüler zum einen durch umfangreiche Lernaktivitäten und zum anderen durch die spezifische Prüfungsvorbereitung ihrer Lehrerinnen und Lehrer in besonderem Maße, und vermutlich deutlich besser und intensiver als bei „einfachen" Klausuren und Leistungskontrollen, auf die Anforderungen von Abiturprüfungen vorbereitet sein dürften. Möglicherweise werden in den Prüfungen daher teilweise sogar bessere Noten erzielt als in der Qualifikationsphase.

Ein Blick auf die derzeitige empirische Befundlage lässt nur bedingt Rückschlüsse auf die Frage zu, inwieweit Schülerinnen und Schüler in den Abiturprüfungen besser oder schlechter abschneiden als in den Klausuren und Leistungskontrollen der Qualifikationsphase. So sind die bildungswissenschaftlichen Forschungsarbeiten zum Abitur, die sich vor allem auf der Grundlage von Daten aus der in Baden-Württemberg durchgeführten TOSCA-Studie (Köller et al., 2004), aus der in Hamburg durchgeführten Studie LAU-13 (Lehmann et al., 2006; Trautwein et al., 2007) oder aus den Evaluationen zur Einführung des Zentralabiturs in den Ländern Hessen und Bremen (Maag Merki, 2012) in den letzten Jahren mit Abiturprüfungsnoten beschäftigt haben, in ihrer Zahl insgesamt überschaubar und setzen inhaltlich andere Schwerpunkte. Zum Beispiel wurde ermittelt, wie Abiturprüfungsnoten mit den Ergebnissen standardisierter Leistungstests zusammenhängen (Nagy et al., 2007; Jonkmann et al., 2007; Maag Merki & Holmeier, 2015; Hübner et al., 2020), oder es wurde untersucht, welche Effekte zentrale Abiturprüfungen auf die Art der Bezugsnormorientierung bei der Bewertung von Prüfungsarbeiten haben (Neumann et al., 2011). Statistiken, aus denen hervorgeht, wie viele Notenpunkte die Prüflinge im Mittel in den Abiturprüfungen und in den Halbjahren der Qualifikationsphase erreichen, werden dabei (wenn überhaupt) nur am Rande berichtet. Insgesamt lassen die wenigen hierzu publizierten Zahlen vermuten, dass die Differenzen zwischen Prüfungsleistungen und Vorleistungen überwiegend eher gering bis moderat sind und je nach Fach sowohl zugunsten der in der Qualifikationsphase erzielten Leistungen als auch zugunsten der Abiturprüfungsergebnisse ausfallen können (Neumann et al., 2009, 2011; Hübner et al., 2020). Darüber hinaus belegen die Evaluationsberichte des IQB zur Bewährung der Gemeinsamen Abituraufgabenpools der Länder, dass die Klausurnoten und Halbjahresnoten der Qualifikationsphase sowohl mit den Gesamtergebnissen in den Abiturprüfungen als auch mit den Ergebnissen bei einzelnen Abiturprüfungsaufgaben hoch korreliert sind (vgl. Beitrag 5 in diesem Band; Hoffmann et al., 2018, 2020).

Weitere Informationen zu Abiturprüfungsnoten lassen sich aus amtlichen Statistiken entnehmen, die regelmäßig, etwa von der KMK, den Bildungsministerien der

Länder oder den statistischen Landesämtern, veröffentlicht werden. In der Regel geben diese Daten zwar Auskunft über die Abiturdurchschnittsnoten, die von Abiturientinnen und Abiturienten verschiedener Jahrgänge im Mittel erzielt wurden, für viele Länder finden sich aber auch Informationen zu den im Mittel erreichten Abiturprüfungsnoten. Öffentlich zugängliche Statistiken, in denen die Differenzen zwischen den in der Qualifikationsphase erzielten Leistungen und den bewerteten Leistungen in den Abiturprüfungen systematisch ausgewiesen werden, sucht man hingegen weitgehend vergebens. Eine Ausnahme hiervon bilden die Bildungsberichte des Landes Hamburg für die Jahre 2017 und 2020 (Hildenbrand, 2017, 2020). Die dort berichteten Ergebnisse stützen insgesamt die Annahme, dass in den Abiturprüfungen im Mittel weniger gute Ergebnisse erzielt werden als in den Halbjahren der gymnasialen Oberstufe. Gleichzeitig verdeutlichen die Befunde aber auch, dass die Höhe des Unterschieds zwischen den Abiturprüfungsnoten und den Kursnoten je nach Schulart, Prüfungsart und Fach erheblich variiert. So fällt die Diskrepanz zwischen Prüfungsergebnissen und Vorleistungen an Stadtteilschulen sehr viel höher aus als an Gymnasien. Bei den schriftlichen Abiturprüfungen ist die Diskrepanz insgesamt deutlicher ausgeprägt als bei den mündlichen Abiturprüfungen, und im Fach Mathematik ist der Unterschied zwischen Abiturprüfungsnoten und Kursnoten sehr viel größer als in den Fächern Deutsch und Englisch.

Insgesamt lässt sich also festhalten, dass bislang nur wenige Befunde aus empirischen Forschungsarbeiten und amtlichen Statistiken vorliegen, die eine Beurteilung der Annahme erlauben, dass in den Abiturprüfungen weniger gute Ergebnisse erzielt werden als in den Klausuren und Leistungskontrollen der Qualifikationsphase. Gestützt wird diese Annahme von den für das Land Hamburg veröffentlichten Befundmustern, die sich auf zwei Schuljahre (2015/16 und 2018/19) beziehen und dementsprechend robust erscheinen (Hildenbrand, 2017, 2020). Fraglich ist allerdings, inwieweit das Hamburger Befundmuster auf andere Länder generalisiert werden kann. So haben insbesondere die Ergebnisse nationaler Schulleistungsstudien wie die IQB-Bildungstrends gezeigt, dass am Ende der Sekundarstufe I zum Teil erhebliche Länderunterschiede bei den im Mittel von Schülerinnen und Schülern erreichten Kompetenzen bestehen. Dies gilt auch für Schülerinnen und Schüler, die die Allgemeine Hochschulreife anstreben (Stanat et al., 2016, 2019). Dass sich diese Unterschiede in der gymnasialen Oberstufe nivellieren, erscheint wenig plausibel, zumal auch in Analysen von Daten der TOSCA- und LAU-13-Studie für das Fach Mathematik deutliche Länderunterschiede bei den in einem standardisierten Leistungstest im Mittel erreichten Ergebnissen zugunsten von Baden-Württemberg und zuungunsten von Hamburg festgestellt wurden (Nagy et al., 2007; Neumann et al., 2009, 2011). Der in den Mathematiktests ermittelte Leistungsunterschied spiegelte sich dabei nicht in den Kursnoten wider – diese fielen in beiden Ländern im Mittel sehr ähnlich aus –, wohl aber in den Noten für die Abiturprüfungen, bei denen die Prüflinge aus Baden-Württemberg deutlich besser abschnitten als die Prüflinge aus Hamburg. Dementsprechend unterschieden sich beide Länder auch in der Differenz zwischen Abi-

turprüfungsnoten und Kursnoten. Diese fiel in Baden-Württemberg deutlich geringer aus als in Hamburg.

Das Ziel des vorliegenden Beitrags besteht darin, die derzeitige Befundlage zur Frage, ob und in welchem Maße sich Abiturprüfungsnoten und Kursnoten unterscheiden, zu erweitern. Der Beitrag gliedert sich im Folgenden in drei Abschnitte, in denen die Passung bzw. der Zusammenhang zwischen Abiturprüfungsnoten und Kursnoten anhand von jeweils unterschiedlichen Datensätzen und mit jeweils anderen Schwerpunktsetzungen untersucht wird. In Abschnitt 1 wird zunächst aus einer länderübergreifenden Perspektive eine größere Zahl von Ländern in den Blick genommen. Dabei wird anhand der bereits in Beitrag 5 vorgestellten Daten aus den vom IQB durchgeführten Evaluationen zur Bewährung der ländergemeinsamen Poolaufgaben und mithilfe metanalytischer Methoden untersucht, ob in den Fächern Deutsch, Englisch und Mathematik systematische Unterschiede zwischen Abiturprüfungsnoten und Kursnoten bestehen. Der Abschnitt 2 geht insbesondere der Frage nach, ob sich die in den Jahren zwischen 2001 und 2012 in vielen Ländern durchgeführten Reformen der gymnasialen Oberstufe, mit denen verpflichtende (Kern-)Fächer auf einem einheitlichen Niveau für alle Schülerinnen und Schüler eingeführt wurden, auf die Passung zwischen Abiturprüfungsnoten und Kursnoten ausgewirkt haben. Der Beitrag nutzt Daten einer Zusatzstudie des Nationalen Bildungspanels (NEPS) für das Land Thüringen. Der Abschnitt 3 analysiert schließlich die Ebene der beiden Abiturblöcke. Auf der Grundlage von Daten der BERLIN-Studie (Maaz et al., 2013; Neumann et al., 2017a) wird untersucht, ob und in welchem Ausmaß die in Block I (Kursblock) zur Bestimmung der Abiturgesamtnote eingebrachten Notenpunkte von den in Block II (Abiturprüfungsblock) erreichten Notenpunkten differieren. Der vorliegende Beitrag schließt mit einer knappen Diskussion der Ergebnisse und wesentlicher Desiderata.

1 Befunde aus Metanalysen zu Unterschieden zwischen Abiturprüfungsnoten und Kursnoten

Parallel zum ersten Einsatz von Aufgaben aus den Gemeinsamen Abituraufgabenpools der Länder im Jahr 2017 hat das IQB damit begonnen, die Bewährung dieser Aufgaben systematisch zu evaluieren. Im Rahmen der Evaluation des IQB wurden in den Prüfungsjahren 2017 und 2019 umfangreiche Daten zum Abschneiden von Prüflingen in den schriftlichen Abiturprüfungen und in der Qualifikationsphase erfasst. Diese Daten ermöglichen es, unabhängig von der Frage, ob und in welchem Umfang die Länder Aufgaben aus den Pools eingesetzt haben, länderübergreifend zu untersuchen, ob sich Unterschiede zwischen Abiturprüfungsnoten und Kursnoten feststellen lassen. Im Folgenden soll zunächst dargestellt werden, welche Forschungsfragen im Fokus der hierzu durchgeführten Analysen stehen, um darauffolgend das methodische Vorgehen zu skizzieren. Im Anschluss daran werden die Ergebnisse der Analysen dargestellt und diskutiert.

1.1 Forschungsfragen

1. Die in diesem Abschnitt vorgestellten Analysen sind von der Frage geleitet, ob sich bei einer länderübergreifenden, bundesweiten Betrachtung in den Fächern Deutsch, Englisch und Mathematik systematische Unterschiede zwischen den Abiturprüfungsnoten und den in der Qualifikationsphase erreichten Kursnoten ergeben. Ferner wird explorativ geprüft, inwieweit der Unterschied zwischen Abiturprüfungsnoten und Kursnoten mit dem Anforderungsniveau, auf dem die Prüfungen geschrieben werden, und dem Geschlecht der Prüflinge variiert.
2. Zusätzlich wird der Frage nachgegangen, ob sich Ländereffekte finden, also ob sich die für ein bestimmtes Fach länderübergreifend festgestellte Differenz aus Abiturprüfungsnoten und Kursnoten zwischen den Ländern systematisch unterscheidet.
3. Darüber hinaus wird untersucht, inwieweit Schuleffekte vorliegen. Hierzu wird für jedes Fach und jedes einzelne Land ermittelt, ob und in welchem Maße der Unterschied zwischen Abiturprüfungsnoten und Kursnoten zwischen den Schulen des jeweiligen Landes variiert.

1.2 Methodisches Vorgehen

1.2.1 Daten

Die in diesem Abschnitt vorgestellten Analysen und Befunde basieren auf den Daten der vom IQB für die Prüfungsjahre 2017 und 2019 durchgeführten Evaluationen zur Bewährung der aus den Gemeinsamen Abituraufgabenpools entnommenen Aufgaben in den schriftlichen Abiturprüfungen der Länder. Die konzeptionellen Grundlagen, das methodische Vorgehen und die zentralen Befunde dieser Evaluationen sind ausführlich in Beitrag 5 dieses Bandes dokumentiert. Informationen zu den Stichproben, die den in diesem Abschnitt vorgestellten Analysen zu Unterschieden zwischen den Noten für die schriftlichen Abiturprüfungen und den Kursnoten in den Fächern Deutsch, Englisch und Mathematik zugrunde liegen, können der Tabelle 1 entnommen werden. In dieser Tabelle ist zum einen jeweils die Anzahl der Prüflinge angegeben, von denen Daten zu Abiturprüfungsergebnissen und Kursnoten in die Analysen für die Fächer Deutsch, Englisch und Mathematik eingeflossen sind. Zum anderen wird über die Anzahl der Länder informiert, aus denen diese Prüflinge stammen.

Aus den bisherigen Evaluationen des IQB zur Bewährung der aus den Gemeinsamen Abituraufgabenpools entnommenen und in den schriftlichen Abiturprüfungen der Länder eingesetzten Aufgaben liegen für das Prüfungsjahr 2017 Daten zu den Fächern Deutsch, Englisch und Mathematik und für das Prüfungsjahr 2019 Daten zu den Fächern Deutsch und Englisch vor.[8] Im Rahmen der Evaluationen für die Prüfungsjahre 2017 und 2019 erfolgte die Datenerhebung mit unterschiedlichen Verfahrensweisen. Die allermeisten Länder nutzten ein vom IQB programmiertes digitales Eingabeinstrument, das in den betrachteten Fächern für jeden Prüfling der für die Evaluation ausgewählten Kurse Angaben zu den Abiturprüfungsnoten sowie zu den

8 In beiden Prüfungsjahren wurden zudem Daten für das Fach Französisch erhoben, die aber aufgrund der geringen Stichprobenumfänge im Rahmen des vorliegenden Beitrags nicht ausgewertet wurden.

in den vier Halbjahren der Qualifikationsphase erreichten Halbjahresnoten erfasste. Dabei wurden zusätzlich zu den Halbjahresnoten die in der Qualifikationsphase erzielten Klausurnoten erhoben. Einige Länder verzichteten dagegen ganz oder in Teilen auf den Einsatz des Eingabeinstruments und übermittelten dem IQB stattdessen Daten aus eigenen Erhebungen, in denen zum Teil weniger Variablen (d. h. nur Halbjahresnoten, jedoch keine Klausurnoten) erfasst werden als vom Eingabeinstrument des IQB. Aus diesem Grund sind in Tabelle 1 je nachdem, ob die Differenz zwischen Abiturprüfungsnoten und Kursnoten in Bezug auf die Halbjahresnoten oder in Bezug auf die Klausurnoten untersucht wurde, unterschiedliche Stichprobenumfänge ausgewiesen.

Tabelle 1: Überblick der Analysestichproben für die Fächer Deutsch, Englisch und Mathematik

Fach	Prüfungsjahr	Art der Vorleistung	Anzahl der Länder	Anzahl der Prüflinge
Deutsch	2017	Halbjahresnoten	14	5085
		Klausurnoten	11	3051
	2019	Halbjahresnoten	16	6861
		Klausurnoten	14	4962
Englisch	2017	Halbjahresnoten	13	4636
		Klausurnoten	10	3038
	2019	Halbjahresnoten	16	8569
		Klausurnoten	14	6759
Mathematik	2017	Halbjahresnoten	13	6702
		Klausurnoten	10	3555

In allen Fällen lagen die von den Prüflingen in den Abiturprüfungen und in der Qualifikationsphase erzielten Ergebnisse in Form von Notenpunkten vor, also auf einer von 0 bis 15 Punkte reichenden Notenskala.

1.2.2 Statistische Auswertung

Die Auswertung der Daten zu den oben skizzierten Fragestellungen erfolgte in zwei Schritten. Das Ziel des ersten Analyseschritts bestand darin, zunächst für jedes Prüfungsjahr, jedes Fach und jedes Land die Differenz (sowie den Standardfehler der Differenz) zwischen den in der schriftlichen Abiturprüfung erzielten Notenpunkten einerseits und dem Mittelwert der in den Halbjahren der Qualifikationsphase bzw. dem Mittelwert der in den Klausuren erzielten Notenpunkten andererseits zu schätzen.[9]

9 In den Analysen wurden die erzielten Vorleistungen somit auf zweifache Weise operationalisiert: Zum einen wurde der Mittelwert der Halbjahresnoten herangezogen, die auf den in einem Semester erzielten Noten basieren. Hierunter fallen zum Beispiel Noten aus mehrstündigen schriftlichen Klausuren, aber auch Noten, die in schriftlichen und mündlichen Kurzkontrollen erzielt wurden. Zum anderen wurde als zweiter Indikator ausschließlich der Mittelwert der Klausurnoten verwendet.

Hierzu wurden mittels des Pakets *lme4* (Bates et al., 2015) in der Statistiksoftware R gemischte Modelle spezifiziert, bei denen die jeweils erzielten Notenpunkte durch die jeweilige Art der Leistung (Abiturprüfung vs. Halbjahresnoten bzw. Abiturprüfung vs. Klausurnoten) statistisch vorhergesagt und gleichzeitig die Prüflinge sowie deren Schulen als zufällige Effekte modelliert wurden. Diese Form der Analyse stellt ein Äquivalent zu konventionellen *t*-Tests für abhängige Stichproben dar, bietet jedoch den Vorteil, dass im Rahmen der Schätzung von Standardfehlern auch die hierarchische Struktur der Daten (im vorliegendem Fall sind Prüflinge in Schulen geschachtelt) Berücksichtigung findet (Hox et al., 2017). Für die in den Fragestellungen adressierten Subgruppen, also jeweils für weibliche und männliche Prüflinge sowie für das erhöhte und grundlegende Anforderungsniveau, wurden separat analoge statistische Modelle spezifiziert. Zur Untersuchung von Schuleffekten wurde zudem für jedes Fach und jedes Land die Intraklassenkorrelation der Variable „Schule" berechnet. Die dabei ermittelten Kennwerte liegen im Wertebereich zwischen 0 und 1 und geben an, welcher Anteil der Varianz der in einem Fach und Land ermittelten Differenz zwischen Abiturprüfungsergebnissen und Kursnoten auf Unterschiede zwischen Prüflingen innerhalb oder zwischen Schulen zurückgeführt werden kann.

Im zweiten Analyseschritt wurden für jedes Prüfungsjahr die zuvor landesspezifisch geschätzten Parameter mittels metaanalytischer Methoden und unter Verwendung des R-Pakets *meta* (Balduzzi et al., 2019) zu länderübergreifenden Gesamtergebnissen zusammengefasst. Hierbei wurden in jedem Fach die zuvor für jedes Land geschätzten Differenzen zwischen Abiturprüfungsergebnissen und Kursnoten jeweils als separate „Studien" in die Auswertung einbezogen. Zudem wurden Differenzierungen nach dem Geschlecht der Prüflinge und dem Anforderungsniveau der Prüfung vorgenommen. Als Ergebnis der durchgeführten *Random-Effects*-Metaanalysen wurde jeweils länderübergreifend die mittlere Differenz zwischen Abiturprüfungsergebnissen und Kursnoten geschätzt. Dieser Schätzung liegt ein Algorithmus zugrunde, der jeder einzelnen „Studie" ein Gewicht zuweist, das bestimmt, mit welchem Anteil die Daten des betreffenden Landes bei der Ermittlung des Gesamteffekts Berücksichtigung finden. Die Höhe dieser Gewichte bemisst sich nach der statistischen Aussagekraft der einzelnen Studien. Dabei sind Länderergebnisse, die anhand großer Stichproben gewonnen werden und bei denen die Differenz zwischen Abiturprüfungsergebnissen und Kursnoten über alle Prüflinge hinweg relativ homogen ausgeprägt ist, aus statistischer Sicht aussagekräftiger als Länderergebnisse, die auf kleinen Stichproben basieren und durch eine hohe Streuung gekennzeichnet sind (Borenstein et al., 2009).

Wenn die in den Metaanalysen geschätzte mittlere Differenz ein positives Vorzeichen hat, so bedeutet dies, dass in der Abiturprüfung im Mittel bessere Ergebnisse erzielt wurden als in der Qualifikationsphase. Bei einer negativen Differenz sind die Prüfungsergebnisse schlechter ausgefallen als die in der Qualifikationsphase erzielten Ergebnisse. Zu den für die einzelnen Prüfungsjahre und Fächer ländergreifend geschätzten mittleren Differenzen wurde außerdem der Vertrauensbereich berechnet, der jeweils angibt, in welchen Grenzen die „wahre" Differenz zwischen Abitur-

prüfungsergebnissen und Kursnoten mit einer Wahrscheinlichkeit von 95 Prozent liegt. Zudem wurde ermittelt, ob sich die jeweils geschätzten Differenzen statistisch signifikant von Null unterscheiden. Darüber hinaus wurde geprüft, ob sich zwischen den betrachteten Subgruppen (männliche vs. weibliche Prüflinge, Abiturprüfung auf erhöhtem vs. grundlegendem Niveau) statistisch signifikante Unterschiede in der Höhe der geschätzten mittleren Differenzen finden.

Zur Untersuchung von Ländereffekten wurde zusätzlich analysiert, wie heterogen die für die einzelnen Länder geschätzten Differenzen zwischen Abiturprüfungsergebnissen und Kursnoten ausfallen. Als Indikator für die Heterogenität wurde erstens der Kennwert *Cochran's Q* bestimmt, der die (gewichtete) Summe der quadratischen Abweichungen der für die einzelnen Länder geschätzten Differenzen von der länderübergreifend geschätzten Differenz abbildet (Borenstein et al., 2009). Basierend auf *Cochran's Q* wurden zudem Signifikanztests für die Heterogenität durchgeführt. Diese Tests prüfen, ob sich der „wahre" Wert der Differenz von Abiturprüfungsergebnissen und Kursnoten zwischen den Ländern unterscheidet. Als zweiter Indikator für die Heterogenität wurde jeweils der Kennwert I^2 („Jota-Quadrat") berechnet, für den sich ebenfalls ein Vertrauensbereich bestimmen lässt (Higgins & Thompson, 2002; Higgins et al., 2003). Der Kennwert ist auf einen Wertebereich von 0 bis 100 Prozent normiert und bildet den Anteil der Varianz zwischen den einzelnen Studien an der Gesamtvarianz einer Metanalyse ab. Einer Daumenregel des Cochrane-Instituts zufolge sind dabei Werte zwischen 0 und 40 Prozent zu vernachlässigen. Werte von 30 bis 60 Prozent sind als „moderat", Werte zwischen 50 und 90 Prozent als „substanziell" einzustufen. Ein I^2 zwischen 75 und 100 Prozent gilt als „beträchtlich" (Deeks et al., 2021). Für den Kennwert I^2 wurde ebenfalls ein Vertrauensbereich bestimmt.

1.3 Ergebnisse

Für jedes Fach erfolgt die Ergebnisdarstellung in Form einer separaten Tabelle. Den Tabellen kann jeweils in der vierten Spalte die länderübergreifend geschätzte mittlere Differenz in Notenpunkten (*NP*) zwischen Abiturprüfungsergebnissen und Kursnoten entnommen werden. Jeweils in Klammern dahinter ist der dazugehörige Vertrauensbereich angegeben. Die mithilfe der Metaanalysen geschätzten mittleren Differenzen sind dabei sowohl für die jeweilige Gesamtstichprobe als auch separat für die oben genannten Subgruppen (männliche vs. weibliche Prüflinge, Abiturprüfung auf erhöhtem vs. grundlegendem Niveau) ausgewiesen. Darüber hinaus ist als deskriptives Maß für die Heterogenität der für die einzelnen Länder geschätzten Differenzen in der fünften Spalte der jeweils im Rahmen der Metanalysen für die Gesamtstichprobe ermittelte Kennwert I^2 (inklusive des zugehörigen Vertrauensintervalls) angegeben. Die Ergebnisse der zusätzlich für die Heterogenität durchgeführten Signifikanztests können der sechsten Spalte entnommen werden. In der siebten Spalte sind schließlich die Ergebnisse von Signifikanztests zu Subgruppenunterschieden ausgewiesen.

1.3.1 Ergebnisse für das Fach Deutsch

Wie anhand von Tabelle 2 ersichtlich, zeigt sich im Fach Deutsch auf der Grundlage der Daten für das Prüfungsjahr 2017, dass die schriftlichen Abiturprüfungen länderübergreifend im Mittel signifikant schlechter ausfallen als die Kursnoten. Für den Vergleich zwischen Abiturprüfungsnoten und Halbjahresnoten wurde dabei eine Differenz von -1.20 *NP* ($z = -18.86$, $p < 0.001$) ermittelt. Ein deutlich geringer ausgeprägter, aber dennoch statistisch signifikanter Unterschied von -0.34 *NP* ($z = -3.43$, $p < 0.001$) konnte beim Vergleich zwischen Abiturprüfungsnoten und Klausurnoten festgestellt werden. Die für die einzelnen Länder berechneten Differenzen variieren in Bezug auf die Halbjahresnoten zwischen -0.66 *NP* und -1.69 *NP* und in Bezug auf die Klausurnoten zwischen -0.62 *NP* und 0.12 *NP*. Trotz dieser Spannweite legen die Ausprägung des Kennwerts I^2 und die hierzu durchgeführten Signifikanztests den Schluss nahe, dass die Heterogenität der für die einzelnen Länder geschätzten Differenzen insgesamt zu vernachlässigen ist und somit keine systematischen Ländereffekte vorliegen. Bemerkenswert ist aber, dass die Differenz zwischen Abiturprüfungsnoten und Kursnoten je nach Land in unterschiedlichem Maße auf Schuleffekte zurückzuführen ist. So variiert die für die Variable „Schule" berechnete Intraklassenkorrelation je nach Land bei den Halbjahresnoten zwischen .03 und .22 ($\widetilde{x} = .08$) und bei den Klausurnoten zwischen .00 und .25 ($\widetilde{x} = .08$). Dies bedeutet, dass vorfindbare Differenzen zwischen Prüfungs- und Kursnoten in einigen Ländern stärker und in anderen Ländern weniger bzw. kaum auf Unterschiede zwischen den Schulen zurückführbar sind. Die Analysen zu Subgruppenunterschieden ergaben keine statistisch signifikanten Ergebnisse.

Die anhand der Daten zum Prüfungsjahr 2019 berechneten Ergebnisse sind ebenfalls in Tabelle 2 zu finden. Da diese nahezu identisch mit den für das Prüfungsjahr 2017 ermittelten Resultaten sind, wird auf eine ausführliche Darstellung der Ergebnisse an dieser Stelle verzichtet.

1.3.2 Ergebnisse für das Fach Englisch

Die im Fach Englisch länderübergreifend geschätzten mittleren Differenzen zwischen Abiturprüfungsnoten und Kursnoten fallen deutlich geringer aus als im Fach Deutsch (s. Tabelle 3). So wurde anhand der Daten zum Prüfungsjahr 2017 in Bezug auf die in der Qualifikationsphase erreichten Halbjahresnoten eine statistisch signifikante Differenz von -0.57 *NP* ($z = -5.93$, $p < 0.001$) berechnet. In Bezug auf die in der Qualifikationsphase erzielten Klausurnoten wurde sogar eine Differenz mit positivem Vorzeichen gefunden, die aber nicht statistisch signifikant ist (0.02 *NP*, $z = -0.16$, $p = 0.88$).

Die für die einzelnen Länder ermittelten Differenzen zwischen Abiturprüfungsnoten und Halbjahresnoten streuen mit einer Spannweite von rund 1 *NP* (*max* = -0.07 *NP*, *min* = -1.08 *NP*). Insgesamt ist die Heterogenität der für die einzelnen Länder festgestellten Differenzwerte als „substanziell" einzustufen ($I^2 = 65.1\%$, $Q(12) = 34.36$, $p < .001$). Eine Spannweite von ähnlicher Höhe wurde für die landesspezifisch ermittelten Differenzen zwischen Abiturprüfungsnoten und Klausurnoten festgestellt

(*max* = 0.33 *NP*, *min* = -0.58 *NP*), wobei hier die Heterogenität der Einzelkennwerte insgesamt betrachtet vernachlässigbar zu sein scheint ($I^2 = 0\,\%$, $Q(9) = 7.12$, $p = .63$). Auch für das Fach Englisch finden sich Hinweise auf Schuleffekte: Je nach Land variiert die für die Variable „Schule" berechnete Intraklassenkorrelation bei den Halbjahresnoten zwischen .01 und .33 ($\widetilde{x} = .10$) und bei den Klausurnoten zwischen .06 und .37 ($\widetilde{x} = .21$). Die Analysen zu Subgruppenunterschieden ergaben keine statistisch signifikanten Ergebnisse.

Ähnlich wie im Fach Deutsch wird auch im Fach Englisch das für das Prüfungsjahr 2017 ermittelte Befundmuster im Wesentlichen von den für das Prüfungsjahr 2019 berechneten Ergebnissen bestätigt (s. Tabelle 3). Hervorzuheben ist lediglich, dass anhand der Daten zum Prüfungsjahr 2019 ein statistisch signifikanter Subgruppenunterschied zwischen den Halbjahresnoten einerseits und den Ergebnissen von Prüfungen auf erhöhtem oder grundlegendem Niveau andererseits ermittelt wurde ($Q(1) = 5.78$, $p = .016$). Mit 0.29 *NP* (erhöhtes Niveau: -0.57 *NP*; grundlegendes Niveau: 0.28 *NP*) ist dieser Unterschied jedoch relativ gering ausgeprägt.

1.3.3 Ergebnisse für das Fach Mathematik

Für das Fach Mathematik wurden im Quervergleich der Fächer die höchsten mittleren Differenzen zwischen Abiturprüfungsnoten und Kursnoten ermittelt (s. Tabelle 4). Länderübergreifend zeigt sich, dass die Ergebnisse der schriftlichen Abiturprüfungen im Fach Mathematik im Mittel um rund -1.70 *NP* ($z = -6.45$, $p < 0.001$) von den Halbjahresnoten bzw. um rund -0.80 *NP* ($z = -3.02$, $p = 0.003$) von den Klausurnoten differieren. Die Spannweite der für die einzelnen Länder geschätzten Differenzen ist im Fach Mathematik deutlich größer als in den anderen beiden Fächern. Die länderspezifischen Differenzen variieren in Bezug auf die Halbjahresnoten zwischen -3.10 *NP* und 0.16 *NP*, in Bezug auf die Klausurnoten reichen sie je nach Land von -2.12 *NP* bis 0.84 *NP*. Die Heterogenität der Differenzwerte zwischen den Ländern ist dementsprechend jeweils als beträchtlich einzustufen (Halbjahresnoten: $I^2 = 88.9\,\%$, $Q(12) = 108.53$, $p < .001$; Klausurnoten: $I^2 = 77.4\,\%$; $Q(9) = 39.79$, $p = .002$).

Im Fach Mathematik finden sich ebenfalls Hinweise auf Schuleffekte. Die für die Variable „Schule" berechneten Intraklassenkorrelationen fallen dabei höher als in den beiden anderen Fächern aus und variieren je nach Land beim Vergleich von Abiturprüfungsnoten und Halbjahresnoten zwischen .08 und .38 ($\widetilde{x} = .21$) sowie beim Vergleich von Abiturprüfungsnoten und Klausurnoten zwischen .11 und .34 ($\widetilde{x} = .22$). Die Analysen zu Subgruppenunterschieden ergaben keine statistisch signifikanten Ergebnisse.

Tabelle 2: Ergebnisse der Metaanalysen zu Unterschieden zwischen Abiturprüfungsnoten und Halbjahres- sowie Klausurnoten im Fach Deutsch

Prüfungs-jahr	Art der Vorleistung	Subgruppe	Differenz in Notenpunkten	Heterogenität I^2	Test auf Heterogenität	Test auf Subgruppen-unterschiede
2017	Halbjahresnoten	Alle Prüflinge	-1.20 [-1.33; -1.08]	0.0 % [0.0 %; 31.5 %]	$Q(13) = 8.53$, $p = .81$	
		Weibliche Prüflinge	-1.22 [-1.37; -1.07]			$Q(1) = 0.29$, $p = .59$
		Männliche Prüflinge	-1.15 [-1.34; -0.96]			
		Erhöhtes Niveau	-1.21 [-1.34; -1.08]			$Q(1) = 0.59$, $p = .44$
		Grundlegendes Niveau	-1.04 [-1.46; -0.63]			
	Klausurnoten	Alle Prüflinge	-0.34 [-0.54; -0.15]	0.0 % [0.0 %; 19.1 %]	$Q(10) = 4.91$, $p = .90$	
		Weibliche Prüflinge	-0.32 [-0.54; -0.09]			$Q(1) = 0.29$, $p = .59$
		Männliche Prüflinge	-0.41 [-0.68; -0.15]			
		Erhöhtes Niveau	-0.35 [-0.56; -0.14]			$Q(1) = 0.05$, $p = .82$
		Grundlegendes Niveau	-0.29 [-0.80; 0.22]			
2019	Halbjahresnoten	Alle Prüflinge	-1.19 [-1.30; -1.08]	0.0 % [0.0 %; 51.1 %]	$Q(15) = 14.63$, $p = .48$	
		Weibliche Prüflinge	-1.19 [-1.33; -1.05]			$Q(1) = 0.33$, $p = .57$
		Männliche Prüflinge	-1.13 [-1.29; -0.97]			
		Erhöhtes Niveau	-1.18 [-1.30; -1.06]			$Q(1) = 0.03$, $p = .87$
		Grundlegendes Niveau	-1.20 [-1.39; -1.01]			
	Klausurnoten	Alle Prüflinge	-0.37 [-0.52; -0.22]	4.2 % [0.0 %; 56.9 %]	$Q(13) = 13.56$, $p = 0.41$	
		Weibliche Prüflinge	-0.32 [-0.51; -0.13]			$Q(1) = 0.06$, $p = .81$
		Männliche Prüflinge	-0.36 [-0.57; -0.15]			
		Erhöhtes Niveau	-0.26 [-0.46; -0.07]			$Q(1) = 2.37$, $p = .12$
		Grundlegendes Niveau	-0.48 [-0.68; -0.28]			

Anmerkung: Negative Differenzen in den Notenpunkten zeigen an, dass in den Abiturprüfungen im Mittel schlechtere Ergebnisse erzielt wurden als in der Qualifikationsphase. Positive Differenzen in den Notenpunkten indizieren, dass in den Abiturprüfungen im Mittel bessere Ergebnisse erzielt wurden als in der Qualifikationsphase.

Tabelle 3: Ergebnisse der Metaanalysen zu Unterschieden zwischen Abiturprüfungsnoten und Halbjahres- sowie Klausurnoten im Fach Englisch

Prüfungs-jahr	Art der Vorleistung	Subgruppe	Differenz in Notenpunkten	Heterogenität I^2	Test auf Heterogenität	Test auf Subgruppen-unterschiede
2017	Halbjahresnoten	Alle Prüflinge	-0.57 [-0.76; -0.38]	65.1 % [37.0 %; 80.6 %]	$Q(12) = 34.36$, $p < .001$	
		Weibliche Prüflinge	-0.58 [-0.77; -0.39]			$Q(1) = 2.02$, $p = .15$
		Männliche Prüflinge	-0.38 [-0.58; -0.19]			
		Erhöhtes Niveau	-0.56 [-0.75; -0.36]			$Q(1) = 0.13$, $p = .72$
		Grundlegendes Niveau	-0.62 [-0.87; -0.36]			
	Klausurnoten	Alle Prüflinge	0.02 [-0.17; 0.20]	0.0 % [0.0 %; 52.4 %]	$Q(9) = 7.12$, $p = .63$	
		Weibliche Prüflinge	-0.09 [-0.33; 0.14]			$Q(1) = 0.77$, $p = .38$
		Männliche Prüflinge	0.06 [-0.19; 0.30]			
		Erhöhtes Niveau	0.03 [-0.18; 0.23]			$Q(1) = 0.81$, $p = .37$
		Grundlegendes Niveau	-0.16 [-0.52; 0.20]			
2019	Halbjahresnoten	Alle Prüflinge	-0.55 [-0.70; -0.40]	55.2 % [21.3 %; 74.5 %]	$Q(15) = 33.45$, $p = .004$	
		Weibliche Prüflinge	-0.65 [-0.83; -0.46]			$Q(1) = 2.52$, $p = .11$
		Männliche Prüflinge	-0.46 [-0.60; -0.32]			
		Erhöhtes Niveau	-0.57 [-0.73; -0.41]			$Q(1) = 5.78$, $p = .016$
		Grundlegendes Niveau	-0.28 [-0.45; -0.11]			
	Klausurnoten	Alle Prüflinge	-0.06 [-0.17; 0.04]	0.0 % [0.0 %; 40.4 %]	$Q(13) = 9.80$, $p = .71$	
		Weibliche Prüflinge	-0.15 [-0.29; 0.00]			$Q(1) = 1.62$, $p = .20$
		Männliche Prüflinge	-0.01 [-0.16; 0.13]			
		Erhöhtes Niveau	-0.11 [-0.24; 0.02]			$Q(1) = 1.56$, $p = .21$
		Grundlegendes Niveau	0.03 [-0.15; 0.20]			

Anmerkung: Negative Differenzen in den Notenpunkten zeigen an, dass in den Abiturprüfungen im Mittel schlechtere Ergebnisse erzielt wurden als in der Qualifikationsphase. Positive Differenzen in den Notenpunkten indizieren, dass in den Abiturprüfungen im Mittel bessere Ergebnisse erzielt wurden als in der Qualifikationsphase.

Tabelle 4: Ergebnisse der Metaanalysen zu Unterschieden zwischen Abiturprüfungsnoten und Halbjahres- sowie Klausurnoten im Fach Mathematik

Prüfungs-jahr	Art der Vorleistung	Subgruppe	Differenz in Notenpunkten	Heterogenität I^2	Test auf Heterogenität	Test auf Subgruppen-unterschiede
2017	Halbjahresnoten	Alle Prüflinge	-1.69 [-2.21; -1.18]	88.9 % [82.9 %; 92.8 %]	$Q(12) = 108.53$, $p < .001$	
		Weibliche Prüflinge	-1.93 [-2.43; -1.43]			$Q(1) = 1.42$, $p = .23$
		Männliche Prüflinge	-1.48 [-2.02; -0.95]			
		Erhöhtes Niveau	-1.63 [-2.14; -1.12]			$Q(1) = 1.30$, $p = .25$
		Grundlegendes Niveau	-2.16 [-2.91; -1.41]			
	Klausurnoten	Alle Prüflinge	-0.80 [-1.32; -0.28]	77.4 % [58.5 %; 87.7 %]	$Q(9) = 39.79$, $p < .001$	
		Weibliche Prüflinge	-1.06 [-1.59; -0.53]			$Q(1) = 1.74$, $p = .19$
		Männliche Prüflinge	-0.55 [-1.09; -0.02]			
		Erhöhtes Niveau	-0.75 [-1.29; -0.21]			$Q(1) = 2.29$, $p = .13$
		Grundlegendes Niveau	-1.54 [-2.40; -0.67]			

Anmerkung: Negative Differenzen in den Notenpunkten zeigen an, dass in den Abiturprüfungen im Mittel schlechtere Ergebnisse erzielt wurden als in der Qualifikationsphase. Positive Differenzen in den Notenpunkten indizieren, dass in den Abiturprüfungen im Mittel bessere Ergebnisse erzielt wurden als in der Qualifikationsphase.

1.4 Fazit

Insgesamt verdeutlichen die anhand der Evaluationsdaten des IQB für die Abiturprüfungen der Jahre 2017 und 2019 ermittelten Befunde, dass die schriftlichen Abiturprüfungen bei einer länderübergreifenden Betrachtung im Mittel tatsächlich schlechter ausfallen als die Kursnoten. Dieses Ergebnismuster findet sich insbesondere in Bezug auf die in der Qualifikationsphase erzielten Halbjahresnoten. In Bezug auf die Klausurnoten ist der Unterschied zwischen Abiturprüfungsnoten und Kursnoten hingegen deutlich geringer und für das Fach Englisch auch nicht statistisch signifikant. In die Halbjahresnoten fließen neben den Klausurergebnissen weitere Noten ein, die etwa im Rahmen kleinerer Leistungskontrollen oder für mündliche Leistungen vergeben werden. Offenbar erzielen die Schülerinnen und Schüler hier deutlich bessere Ergebnisse als in den schriftlichen Abiturprüfungen oder in den Klausuren der Qualifikationsphase.

Während sich für die in den Analysen betrachteten Subgruppen (männliche vs. weibliche Prüflinge, Abiturprüfung auf erhöhtem vs. grundlegendem Niveau) keine substanziellen Unterschiede zeigten, lassen die ermittelten Ergebnisse auf einen Fächereffekt schließen: Analog zu den in den Bildungsberichten des Landes Hamburg der Jahre 2017 und 2020 dokumentierten Befundmustern (Hildenbrand, 2017, 2020) ist die länderübergreifend geschätzte mittlere Differenz zwischen Abiturprüfungsnoten und Kursnoten im Quervergleich der Fächer für Mathematik deutlich größer als für Deutsch oder Englisch. Vor allem im Fach Mathematik weisen die Ergebnisse der Analysen darüber hinaus auf das Vorliegen markanter Ländereffekte hin. Während die Schülerinnen und Schüler der meisten Länder in den Abiturprüfungen deutlich schlechter abschnitten als in der Qualifikationsphase, fielen die Abiturprüfungsnoten in einigen Ländern sogar besser aus als die Halbjahresnoten bzw. Klausurnoten. Darüber hinaus wurden (insbesondere im Fach Mathematik) Hinweise auf Schuleffekte gefunden. Die Größe der Differenz zwischen Abiturprüfungsnoten und Kursnoten variiert also nicht nur zwischen den Ländern, sondern auch zwischen den einzelnen Schulen eines Landes.

Diese starke Heterogenität könnte verschiedene Ursachen und Hintergründe haben. Wie bereits oben erwähnt, erscheint plausibel, dass die im Rahmen von Schulleistungsstudien am Ende der Sekundarstufe I festgestellten Länderunterschiede (z. B. Stanat, 2016, 2019) auch in der gymnasialen Oberstufe bestehen. Diese Annahme wird durch die Befunde der in Baden-Württemberg und Hamburg durchgeführten TOSCA- und LAU-13-Studien unterstützt, die erhebliche Länderunterschiede bei den im Mittel erreichten mathematischen Kompetenzen identifizierten (Nagy et al., 2007; Neumann et al., 2009). Anhand dieser Daten konnten zudem Hinweise darauf gefunden werden, dass die Notengebung in der gymnasialen Oberstufe nicht frei von Referenzgruppeneffekten erfolgt (Neumann et al., 2009, 2011). So orientiert sich die Bewertung individueller Leistungen in der Qualifikationsphase offenbar in stärkerem Maße am mittleren Leistungsniveau des jeweiligen Kurses (soziale Bezugsnorm; Rheinberg, 2001) als bei den Abiturprüfungen. Als Konsequenz dürfte die Vergleichbarkeit der in der Qualifikationsphase vergebenen Noten deutlich eingeschränkt sein: Die Noten dürften zwar recht gut die Leistungsunterschiede innerhalb

eines Kurses und ggf. auch innerhalb einer Schule abbilden, nicht jedoch die Leistungsunterschiede, die zwischen Schulen oder Ländern bestehen. Vielmehr dürfte die Anwendung der sozialen Bezugsnorm bei der Notengebung in der Qualifikationsphase dazu führen, dass in jedem Kurs, in jeder Schule, in jedem Land bei den Kursnoten in etwa der gleiche Notendurchschnitt erreicht wird. Demgegenüber scheint die Notengebung bei der Abiturprüfung in stärkerem Maße als in der Qualifikationsphase an sachlichen Bewertungskriterien und inhaltlichen Normen orientiert zu sein (Neumann et al., 2009, 2011), die sich zum Beispiel in den Erwartungshorizonten und Bewertungshinweisen der Abituraufgaben finden (kriteriale Bezugsnorm; Rheinberg, 2001). Die in den schriftlichen Abiturprüfungen erzielten Noten sind also zu einem höheren Grad als die Kursnoten zwischen Schulen und Ländern vergleichbar und korrespondieren dabei besser mit den tatsächlich erreichten mathematischen Kompetenzen.

Darüber hinaus hat die Unterrichtsforschung gezeigt, dass Lehrkräfte ihre Instruktion an das Leistungsniveau der jeweiligen Lernklasse anpassen (Dreeben & Barr, 1988; Kunter et al., 2006). Es erscheint plausibel, dass entsprechend auch die Anforderungen von schriftlichen und mündlichen Leistungskontrollen sowie von Klausuren an das jeweilige Leistungsniveau adaptiert werden. Diese Anpassungen würden ebenfalls die Vergleichbarkeit der in der Qualifikationsphase erzielten Noten zwischen Schulen und Ländern reduzieren, da je nach Leistungsniveau des Kurses jeweils unterschiedlich hohe Anforderungen zum Erlangen bestimmter Noten zu bewältigen wären.

Insgesamt dürften die festgestellten Länder- und Schuleffekte also vor allem Länder- bzw. Schulunterschiede im mittleren Leistungsniveau der Schülerinnen und Schüler abbilden. Übereinstimmend mit diesem Erklärungsansatz zeigen die Befunde von zusätzlichen (nicht in diesem Beitrag dargestellten) länderscharfen Analysen einen engen Zusammenhang zwischen den Ergebnissen nationaler Schulleistungsstudien einerseits und den Unterschieden zwischen Abiturprüfungsnoten und den Kursnoten andererseits. Vor allem für das Fach Mathematik lässt sich feststellen, dass die größten Unterschiede zwischen Prüfungs- und Kursnoten für diejenigen Länder gefunden wurden, die im IQB-Bildungstrend 2018, in dem die mathematischen Kompetenzen von Schülerinnen und Schülern der 9. Jahrgangsstufe erfasst wurden, besonders schwach abgeschnitten haben. Die geringsten Unterschiede wurden hingegen für die Länder ermittelt, in denen im IQB-Bildungstrend 2018 die im Mittel höchsten Kompetenzen erreicht wurden.

Die durchgeführten Analysen sind stichprobenbasiert; dementsprechend stellt sich die Frage nach der Generalisierbarkeit der ermittelten Ergebnisse. Für die Robustheit der Befunde spricht, dass die in den Fächern Deutsch und Englisch für das Prüfungsjahr 2017 festgestellten Ergebnisse nahezu exakt durch die für das Prüfungsjahr 2019 durchgeführten Analysen repliziert werden. Für das Fach Mathematik liegen dagegen nur Evaluationsdaten zum Prüfungsjahr 2017 vor. Allerdings decken sich die für Mathematik festgestellten Ergebnisse zu den Unterschieden von Abiturprüfungsnoten und Kursnoten im Wesentlichen mit internen Statistiken der Länder für die Prüfungsjahre 2017, 2018 und 2019.

2 Zusammenhänge von Kurs- und Abiturprüfungsnoten vor und nach Oberstufenreformen

In den Jahren 2001 bis 2012 reformierte ein Großteil der Länder die gymnasiale Oberstufe, darunter auch Thüringen und Baden-Württemberg (vgl. Beitrag 2 in diesem Band, Trautwein & Neumann, 2008; Trautwein et al., 2010). Die implementierten Veränderungen sahen die Einführung von insgesamt fünf Fächern auf erhöhtem Anforderungsniveau vor, beispielsweise von drei verpflichtenden Kernfächern sowie einem Neigungs- und einem Profilfach[10], und gingen infolgedessen mit einer partiellen Auflösung von Kurswahloptionen in der Oberstufe einher (Neumann, 2010).

Die Einführung verpflichtender (Kern-)Fächer auf einem einheitlichen Niveau für alle Schülerinnen und Schüler ist ein bildungspolitischer Trend, der sich nicht nur in Deutschland, sondern weltweit beobachten lässt und wissenschaftlich oft unter dem Stichwort „curricular intensification" behandelt wird (Domina & Saldana, 2012; Hübner et al., 2019). Als zentrales bildungspolitisches Argument für die Implementation entsprechender Reformen wird oftmals angeführt, dass mit einer stärkeren Vereinheitlichung und Verpflichtung von Kursangeboten in Kernfächern eine geringere Leistungsstreuung (insb. durch einen Leistungsanstieg in niedrigen Leistungsbereichen) erreicht werden könne. Empirische Studien der vergangenen zwei Dekaden konnten noch nicht abschließend klären, inwieweit diese Annahme tatsächlich zutrifft. Insgesamt deutet die bisherige Befundlage jedoch darauf hin, dass die erzielten Effekte zumeist kleiner ausfallen als gewünscht und zudem je nach Reformkontext und Fach variieren (Domina & Saldana, 2012; Hübner et al., 2019; Hübner et al., 2020; Nagy et al., 2010).

Wie umfassend ausgeführt von Neumann (2010), wurden die strukturellen Veränderungen der Oberstufe in vielen Ländern durch eine Novelle rechtlicher Grundlagen ermöglicht. So verabschiedete die KMK im Rahmen der Husumer Beschlüsse im Jahr 1999 neue Vorgaben zur Gestaltung der gymnasialen Oberstufe. Diese Beschlüsse flexibilisierten die Regelungen zur Kursbelegung: Eine Reduktion des Umfangs von Leistungskursen von mindestens fünf auf vier Wochenstunden wurde möglich, wenn drei oder mehr Kurse auf erhöhtem Anforderungsniveau belegt wurden. Für den Freistaat Thüringen, der im Jahr 2009 die Oberstufe reformierte, sind die entsprechenden Veränderungen exemplarisch in Tabelle 5 dargestellt.

Die Oberstufenreformen der Länder lösten in der Regel dyadische Oberstufensysteme, bestehend aus zwei Leistungskursen und zwei Grundkursen in der Abiturprüfung, ab. Leistungskurse wurden zuvor auf einem erhöhten Anforderungsniveau unterrichtet und je nach Land fünf bis sechsstündig belegt. Grundkurse, in denen eine mündliche oder schriftliche Abiturprüfung abgelegt wurde, zielten auf die Vermittlung grundlegender fachbezogener Fähigkeiten und Fertigkeiten ab und wurden je nach Fach und Land drei- bis vierstündig unterrichtet (Hübner et al., 2019). Nach der Reform wurden Abiturfächer (Kern-, Neigungs-, und Profilfächer) jeweils vierstün-

10 Weitere Informationen hierzu finden sich bei Hübner et al. (2019).

dig auf erhöhtem Anforderungsniveau unterrichtet (Thüringer Kultusministerium, 2008). In der jüngeren Vergangenheit ist in mehreren Ländern wieder eine gewisse Rückbewegung hin zu weniger Leistungskursfächern und mehr Wahlfreiheit erkennbar (Neumann & Trautwein, 2019).

Tabelle 5: Exemplarische Übersicht zu Kurswahlmöglichkeiten vor und nach der Oberstufenreform

	Vor der Reform (2010)			Nach der Reform (2011)		
Abiturfachnummer	**Fach**	**Zeit (h/w)**	**Niveau**	**Fach**	**Zeit (h/w)**	**Niveau**
1	D/M	6	erhöht	D	4	erhöht
2	FS/NW/SW	6	erhöht	M	4	erhöht
3	M/D	4	grundlegend	FS	4	erhöht
4	FS/(INF)	3	grundlegend	NW	4	erhöht
5	–	–	–	SW	4	erhöht

Anmerkung: D = Deutsch, M = Mathematik, FS = Fremdsprache, NW = Naturwissenschaft, SW = Sozialwissenschaft, INF = Informatik, h/s = Unterrichtsstunden pro Woche (Übersetzt aus: Hübner et al., 2019).

In den für die Länder Baden-Württemberg und Thüringen durchgeführten empirischen Studien zu den Wirkungen der Oberstufenreformen wurde insbesondere untersucht, ob sich Effekte auf die im Mittel erreichten Kompetenzen feststellen lassen. Je nach Land fielen die Ergebnisse dieser Studien unterschiedlich aus. Während in Baden-Württemberg ein mittlerer Anstieg in der mit standardisierten Tests erfassten Mathematikleistung zu finden war und kein Unterschied in der Englischleistung, zeigten sich in Thüringen diesbezüglich keinerlei substanzielle Unterschiede zwischen den Kohorten vor und nach der Reform (Hübner et al., 2019; Trautwein et al., 2010). Des Weiteren wurde in einer Studie von Hübner et al. (2020) für beide Länder anhand des Zusammenhangs von Testleistungen und Kursnoten untersucht, inwieweit die Reformen Effekte auf die Notengebung in der Qualifikationsphase hatten. Die dabei ermittelten Ergebnisse legen den Schluss nahe, dass Oberstufenreformen auch innerhalb von Ländern die Bedeutung von Kursnoten verändern können. So zeigte sich, dass Schülerinnen und Schüler mit der gleichen Note im Kernfach (nach der Reform) im Mittel eine deutlich niedrigere standardisierte Leistung aufwiesen als Schülerinnen und Schüler im Leistungskurs (vor der Reform) und eine deutlich höhere Leistung als Schülerinnen und Schüler im Grundkurs (vor der Reform). Die Autoren der Studie führen aus, dass diese Unterschiede insbesondere auf die soziale Bezugsnormorientierung bei der Notenvergabe von Lehrkräften zurückzuführen seien. Diese Art der Benotung ist klassenintern konsistent, führt aber letztlich dazu, dass Noten kaum vergleichbar zwischen Kursen, Klassen und Schulen sind.

Bislang noch nicht im Fokus empirischer Studien stand die Frage, inwieweit sich die Reformen auf die erzielten Abiturprüfungsnoten ausgewirkt haben. Dieses Desiderat aufgreifend wird im Rahmen des vorliegenden Beitrags untersucht, wie die in den Abiturprüfungen erzielten Noten vor und nach der im Land Thüringen durchge-

führten Oberstufenreform ausfallen und wie gut dabei jeweils die Passung zwischen Prüfungsnoten und Kursnoten ist. Die Analysen basieren auf den Daten einer Zusatzstudie des nationalen Bildungspanels (NEPS) (Blossfeld & Roßbach, 2019).[11] Im Unterschied zu den in Abschnitt 1 dieses Beitrags dargestellten Analysen erfolgt eine differenzierte Betrachtung der einzelnen Semester bzw. Schulhalbjahre der Qualifikationsphase (hier: 11.1–12.2). Diese Differenzierung ist von der Hypothese geleitet, dass die Kursnoten der späteren Schulhalbjahre möglicherweise bereits von den Vorbereitungsaktivitäten auf die Abiturprüfung geprägt sind und daher insgesamt besser ausfallen könnten als die zu Beginn der Qualifikationsphase erzielten Noten (z. B. Arens et al., 2017).

2.1 Forschungsfragen

Basierend auf den obigen Ausführungen sollen Unterschiede zwischen Kurs- und Abiturprüfungsnoten in den drei Fächern Mathematik, Deutsch und Englisch näher untersucht werden. Konkret sollen die folgenden drei Fragestellungen beantwortet werden:

1. Finden sich in den drei Fächern Unterschiede im mittleren Niveau der Abiturprüfungsnoten in Abhängigkeit der besuchten Kursart (d. h. Kernfachkurse ohne Leistungsdifferenzierung im neuen System vs. Grundkurse und Leistungskurse im traditionellen System)?
2. Zeigen sich in den drei Fächern Unterschiede zwischen Abiturprüfungsnoten und Kursnoten innerhalb der drei betrachteten Kursarten? Fällt also die Passung von Prüfungs- und Kursnoten in Abhängigkeit der besuchten Kursart unterschiedlich aus?
3. Ist für die Passung von Kurs- und Prüfungsnoten im Längsschnitt der vier Schulhalbjahre ein Effekt der zunehmenden Vorbereitungsaktivitäten auf die Abiturprüfungen zu erkennen? Fallen also die Differenzen von Kurs- und Prüfungsnoten in den späteren Halbjahren der Qualifikationsphase geringer aus als in den früheren Halbjahren?

2.2 Methodisches Vorgehen

2.2.1 Daten

Zur Untersuchung der Fragestellungen wurden Daten von rund 2.260 Schülerinnen und Schülern aus 32 Schulen der Zusatzstudie Thüringen (Scientific Usefile 2.0.0) des nationalen Bildungspanels verwendet (Blossfeld & Rossbach, 2019; NEPS-Netzwerk, 2014). Die Daten wurden im Jahrgang der letzten Kohorte vor der Oberstufenreform im Jahr 2010 ($N = 1.374$) und im ersten Jahrgang nach der Oberstufenreform im Jahr 2011 jeweils am Ende der Sekundarstufe II erhoben.

11 Das NEPS wird vom Leibniz-Institut für Bildungsverläufe (LIfBi, Bamberg) in Kooperation mit einem deutschlandweiten Netzwerk durchgeführt.

2.2.2 Statistische Auswertung

Zur Beantwortung der Forschungsfragen erfolgte zunächst die Aufbereitung der Daten im Statistikprogramm R (R Development Core Team, 2021), bevor anschließend in Mplus 8.6 (Muthén & Muthén, 1998–2017) Mehrgruppenmodelle geschätzt wurden, die eine Untersuchung von Unterschieden zwischen zwei oder mehr Gruppen ermöglichen. Alle Analysen erfolgten mittels Full-Information-Maximum-Likelihood-(FIML-)Schätzung. Darüber hinaus wurden clusterrobuste Standardfehler berechnet.

Die vorliegenden Analysen berücksichtigen nur diejenigen Schülerinnen und Schüler, die eine schriftliche Prüfung im jeweiligen Fach ablegten. Konkret wurden in den Mehrgruppenmodellen separat für jede Kursart (Leistungskurs, Grundkurs, Kernfach) der Mittelwert und die Varianz der jeweils in den Abiturprüfungen und der Qualifikationsphase erreichten Notenpunkte geschätzt. Anschließend erfolgte die Prüfung möglicher Unterschiede zwischen den im Mittel bei den einzelnen Kursarten erreichten Prüfungsnoten (Forschungsfrage 1) und möglicher Unterschiede in den Differenzen von Prüfungsnoten und Kursnoten je nach besuchter Kursart (Forschungsfrage 2). Abschließend wurde deskriptiv exploriert, inwieweit sich bei den Differenzen zwischen Prüfungsnoten und Kursnoten über die vier Schulhalbjahre hinweg ein Trend im Sinne eines Effekts der zunehmenden Vorbereitungsaktivitäten auf die Abiturprüfungen abzeichnet (Forschungsfrage 3).

2.3 Ergebnisse

Zunächst wurde untersucht, ob sich zwischen den verschiedenen Kursarten (Leistungskurs, Grundkurs und Kernfach) statistisch signifikante Unterschiede bei den im Mittel erreichten Abiturprüfungsnoten feststellen lassen (Forschungsfrage 1; vgl. Tabelle 6). Beim Vergleich der im Fach Mathematik einerseits im Kernfach (= nach der Reform) und anderseits im Leistungs- und Grundkurs (= vor der Reform) im Mittel erzielten Abiturprüfungsnoten zeigten sich statistisch signifikante Differenzen, die mit jeweils besseren Prüfungsnoten im Kernfach einhergingen. Mit rund 3.5 Notenpunkten fiel dieser Unterschied besonders deutlich für den Vergleich zwischen Kernfach ($M = 9.78$) und Grundkurs ($M = 6.31$) aus ($p < .001$). Deskriptiv betrachtet wurden auch im Fach Deutsch nach der Reform bessere Prüfungsergebnisse erzielt als vor der Reform. Ein statistisch signifikanter Mittelwertunterschied von rund 0.6 *NP* wurde allerdings nur für den Vergleich zwischen den im Kernfach ($M = 9.12$) und im Leistungskurs ($M = 8.54$) erzielten Prüfungsnoten ermittelt ($p < .01$). In Englisch zeigten sich zwischen den drei Kursen nur sehr geringe, statistisch nicht signifikante Mittelwertunterschiede bei den erzielten Prüfungsnoten.

Tabelle 6: Durchschnittliche Note nach Fach und Kurs und Kursunterschiede in Punkten

Fach	Note	Leistungskurs (LK)			Grundkurs (GK)			Kernfach (KF)			Differenzen		
		N	*M*	*SD*	N	*M*	*SD*	N	*M*	*SD*	$Diff_{LKGK}$	$Diff_{LKKF}$	$Diff_{GKKF}$
Mathematik	Schriftliche Abiturnote	568	8.17	3.63	388	6.31	3.59	607	9.78	3.45	1.86***	-1.61***	-3.48***
	Kursnote_11.1	568	9.48	2.95	388	9.16	3.00	607	9.80	2.81	0.32	-0.32	-0.64**
	Kursnote_11.2	568	9.48	3.09	388	9.05	3.15	607	10.14	2.93	0.43	-0.66**	-1.08***
	Kursnote_12.1	568	9.81	2.97	388	9.68	2.98	607	10.29	2.88	0.13	-0.48*	-0.61**
	Kursnote_12.2	568	9.27	3.30	388	8.57	3.48	607	9.99	2.99	0.71*	-0.72*	-1.42***
Deutsch	Schriftliche Abiturnote	634	8.54	3.28	311	8.70	3.15	665	9.12	3.14	-0.15	-0.58**	-0.42
	Kursnote_11.1	634	9.28	2.29	311	9.81	2.49	665	10.22	2.26	-0.53**	0.94***	-0.41
	Kursnote_11.2	634	9.39	2.40	311	10.07	2.52	665	10.33	2.30	-0.68***	-0.94***	-0.26
	Kursnote_12.1	634	9.41	2.48	311	9.95	2.60	665	10.44	2.29	-0.54**	-1.03***	-0.48
	Kursnote_12.2	634	9.77	2.65	311	10.02	2.72	665	10.53	2.39	-0.25	-0.76***	-0.51*
Englisch	Schriftliche Abiturnote	388	9.19	2.65	307	9.30	2.56	585	9.37	2.69	-0.11	-0.18	-0.07
	Kursnote_11.1	388	9.81	2.21	307	9.55	2.50	585	10.00	2.30	0.26	-0.19	-0.45*
	Kursnote_11.2	388	9.93	2.26	307	9.86	2.53	585	10.08	2.43	0.07	-0.15	-0.22
	Kursnote_12.1	388	10.07	2.30	307	9.97	2.49	585	10.16	2.36	0.10	-0.10	-0.20
	Kursnote_12.2	388	10.65	2.23	307	10.22	2.48	585	10.29	2.40	0.43	0.36	-0.07

Anmerkung: Diff = Differenz in Punkten. 11.1–12.2 = Halbjahresnoten.
* $p < .05$, ** $p < .01$, *** $p < .001$

Im nächsten Schritt wurde untersucht, ob und wie stark die in den Abiturprüfungen im Mittel erzielten Noten von den in der Qualifikationsphase erreichten Kursnoten differieren (Forschungsfrage 2). Einen zusammenfassenden Überblick der Ergebnisse zu dieser Forschungsfrage bietet Tabelle 7. Für die Grund- und Leistungskurse zeigte sich in allen drei Fächern und über (nahezu) alle Schuljahre hinweg, dass die Abiturprüfungsnoten jeweils schlechter ausfielen als die Kursnoten. Für den Grundkurs wies das ermittelte Befundmuster dabei eine deutliche Fachspezifik auf: Die größten Unterschiede fanden sich im Fach Mathematik, wo Prüfungsnoten und Kursnoten um bis zu 3.4 *NP* ($p < .001$) differierten. Die geringsten Unterschiede wurden für das Fach Englisch ermittelt. Weniger deutlich ausgeprägt war die Fachspezifik für den Leistungskurs, wobei die Prüfungsnoten und Kursnoten auch hier, insgesamt betrachtet, im Fach Mathematik etwas stärker differierten als in den Fächern Deutsch und Englisch. Auch für das Kernfach zeigte sich, dass die Schülerinnen und Schüler in den Abiturprüfungen weniger gute Ergebnisse erzielten als in den Halbjahren der Qualifikationsphase. Bemerkenswert ist allerdings, dass die für das Fach Mathematik ermittelten Unterschiede zwischen Prüfungsnoten und Kursnoten hier deutlich geringer ausfielen als im Grund- oder Leistungskurs und dabei nur in einem Halbjahr mit 0.5 *NP* ($p < .05$) statistisch signifikant waren. Demgegenüber lagen die im Kernfach für die Fächer Deutsch und Englisch ermittelten Ergebnisse im Bereich der Werte, die für den Grund- bzw. Leistungskurs gefunden wurden.

Abschließend wurde exploriert, ob sich für die im Rahmen von Forschungsfrage 2 ermittelten Differenzen über die vier Schulhalbjahre hinweg Änderungen ergaben und ob sich dabei die Passung zwischen Prüfungsnoten und Kursnoten mit zunehmender zeitlicher Nähe zu den Abiturprüfungen erhöhte (Fragestellung 3). Hierzu zeigt ein Blick auf die in Tabelle 7 dargestellten Ergebnisse, dass die für die einzelnen Fächer zwischen Prüfungsnoten und Kursnoten festgestellten Differenzen bei allen drei Kursarten je nach Schulhalbjahr zum Teil erheblich variierten. Ein Trend im Sinne einer zunehmenden Erhöhung der Passung von Prüfungsnoten und Kursnoten zeigte sich jedoch nicht. In den Fächern Deutsch und Englisch ließen die Ergebnisse in der Tendenz sogar einen gegenteiligen Effekt vermuten.

Tabelle 7: Punktedifferenzen zwischen schriftlichen durchschnittlichen Abiturprüfungs- und Kursnoten nach Fach und Kurs

	Leistungskurs (LK)			
Fach	**11.1**	**11.2**	**12.1**	**12.2**
Mathematik	-1.31***	-1.31***	-1.64***	-1.10***
Deutsch[ü]	-0.73***	-0.85***	-0.87***	-1.23***
Englisch[ü]	-0.62***	-0.74***	-0.87***	-1.46***
	Grundkurs (GK)			
Mathematik[ü]	-2.85***	-2.74***	-3.37***	-2.26***
Deutsch	-1.11***	-1.37***	-1.26***	-1.33***
Englisch[ü]	-0.25	-0.56***	-0.67***	-0.92***

(Fortsetzung Tabelle 7)

	Kernfach (KF)			
Mathematik	-0.02	-0.35	-0.51*	-0.20
Deutsch[ü]	-1.10***	-1.21***	-1.32***	-1.42***
Englisch[ü]	-0.63***	-0.71***	-0.79***	-0.91***

Anmerkung: 11.1–12.2 = Halbjahresnoten. Differenz= Abiturprüfungsnote – Kursnote (d. h. positive Differenz = Abiturprüfungsnoten sind im Durchschnitt besser als die Kursnoten, negative Differenz = Abiturprüfungsnoten sind im Durchschnitt schlechter als die Kursnoten).
* $p < .05$, ** $p < .01$, *** $p < .001$

2.4 Fazit

Insgesamt lassen die hier dargestellten Ergebnisse darauf schließen, dass die im Land Thüringen durchgeführte Oberstufenreform mit einer Verbesserung der Abiturgesamtnoten einhergegangen ist. Dies zeigt sich *erstens* in den Noten der Abiturprüfungen, die in den Fächern Deutsch und Mathematik nach der Reform im Kernfachsystem signifikant besser ausfielen als im traditionellen Kurssystem vor der Reform. *Zweitens* verdeutlichen die Befunde zur Passung von Prüfungsnoten und Kursnoten, dass sich mit der Reform auch die Kursnoten verbessert haben. So fielen in den Fächern Deutsch und Englisch die Differenzen zwischen Prüfungsnoten und Kursnoten vor und nach der Reform jeweils ähnlich hoch aus, mit im Mittel schlechteren Abiturprüfungsnoten. Im Fach Mathematik war die Passung von Prüfungsnoten und Kursnoten nach der Reform deutlich verbessert, was bedeutet, dass sich hier keine statistisch signifikanten Unterschiede zwischen Prüfungs- und Kursnoten mehr zeigten.

Einschränkend ist zu erwähnen, dass die hier dargestellten Ergebnisse zu Auswirkungen von Oberstufenreformen keinen Anspruch auf bundesweite Repräsentativität und Generalisierbarkeit besitzen. Da ausschließlich Daten aus Thüringen und nur zwei Jahrgänge (2010 und 2011) ausgewertet wurden, ist offen, ob die in diesem Abschnitt festgestellten Effekte der Thüringer Oberstufenreform auch die weiteren Prüfungsjahre ab 2012 überdauert haben und inwieweit sie auf die Oberstufenreformen anderer Länder übertragbar sind. Darüber hinaus wurden in den vorliegenden Analysen lediglich schriftliche Abiturprüfungsnoten berücksichtigt und aufgrund der geringen Fallzahlen keine Ergebnisse von mündlichen Verbesserungsversuchen einberechnet. Unter Berücksichtigung dieser Limitationen zeigen die Ergebnisse dieses Abschnitts auch, dass sich Prüfungsnoten und Kursnoten zumeist substanziell voneinander unterscheiden: Eine lediglich auf Basis von Kursnoten berechnete Abiturgesamtnote würde den Ergebnissen dieser Untersuchung zur Folge etwas besser ausfallen als eine Abiturnote auf Basis von Kurs- und Abiturprüfungsnoten.

3 Globale Passung von Kurs- und Prüfungsnoten auf Ebene der Gesamtqualifikation – Was wäre, wenn …?

In den beiden vorangegangenen Abschnitten wurde die Korrespondenz von Halbjahres- und Klausurnoten aus der Kursphase der Oberstufe mit den Noten der schriftlichen Abiturprüfung aus einer fachbezogenen Perspektive für die drei Fächer Deutsch, Mathematik und Englisch heraus betrachtet. Die fachbezogene Perspektive ermöglicht einen differenzierten Blick auf potenziell differierende Ergebnismuster und ist weiterführend besonders dazu geeignet, Niveauunterschiede in den erreichten Fach- und Prüfungsnoten zu den konkreten Prozessen des Unterrichts, zu fachlichen Anforderungen und Leistungserwartungen sowie zur Leistungsbewertung in Unterrichts- und Prüfungssituationen in Beziehung zu setzen. Gleichwohl wird die Diskussion um die Wertigkeit des Abiturs insbesondere in der öffentlichen und politischen Debatte oftmals auf globaler Ebene geführt, etwa wenn angenommen wird, dass die Leistungsanforderungen in bestimmten Ländern allgemein höher sind als in anderen, oder wenn Unterschiede im Abiturdurchschnitt zwischen den Ländern (wie sie etwa in den entsprechenden KMK-Statistiken sichtbar sind) und die aufgrund steigender Abiturdurchschnitte angenommene Noteninflation diskutiert werden. Dies ist unmittelbar mit der weit verbreiteten Annahme verbunden, dass man in manchen Ländern (zunehmend) leichter, in anderen hingegen schwerer zu einer guten Abiturnote käme und dass dafür in nicht unerheblichem Maße auch Unterschiede in curricularen und organisatorischen Ausgestaltungsmerkmalen und Leistungserwartungen in der gymnasialen Oberstufe einschließlich der Abiturprüfungen verantwortlich seien (vgl. z. B. Brodkorb & Koch, 2020).

Vergegenwärtigt man sich vor diesem Hintergrund einerseits die in der langjährigen Historie der Oberstufe und des Abiturs zum Teil erbittert geführten Diskussionen um einzelne, nicht selten sehr detaillierte Ausgestaltungsmerkmale des Kurssystems und der Abiturprüfungen (vgl. im Überblick Neumann, 2010) und andererseits die Bedeutung der Abiturdurchschnittsnote für die Vergabe von knappen Studienplätzen (und hier die Bedeutung von Zehntelnotenunterschieden), so wird deutlich, dass die im Rahmen der Corona-Pandemie diskutierte Möglichkeit eines „Anerkennungsabiturs“ als tiefer Einschnitt wahrgenommen worden wäre. Über die möglichen Auswirkungen eines Wegfalls der Prüfungen auf die Lernergebnisse der Schülerinnen und Schüler – etwa aufgrund eines fehlenden motivationalen Lern- oder Bewährungsanreizes – lässt sich nur spekulieren. Eine empirische Annäherung lässt sich jedoch hinsichtlich der Frage vornehmen, ob und in welchem Maß die Abiturnoten auf Ebene der Gesamtqualifikation besser, schlechter oder unverändert ausfallen, wenn statt der Kombination aus Kurs- und Prüfungsergebnis im Verhältnis 2:1 eine reine Beschränkung auf die erzielten Bewertungen in den besuchten Kursen während der Oberstufe erfolgt. Dies soll im Folgenden beispielhaft für eine Stichprobe von Berliner Abiturientinnen und Abiturienten aus allgemeinbildenden Gymnasien erfolgen. Dieser dritte Teil des vorliegenden Beitrags ergänzt damit die vorangegangenen Ausführungen um einen Fokus auf die Gesamtqualifikation und die Bedeutung poten-

zieller Unterschiede zwischen erreichten Kurs- und Prüfungsblockergebnissen für die Gesamtpunktzahl im Abitur.

3.1 Forschungsfragen

Konkret sollen folgende Forschungsfragen untersucht werden:

1. Welche Zusammenhänge finden sich in der Gesamtqualifikation zwischen Abiturgesamtpunktzahl (maximal 900 Punkte), Kursblockpunktzahl (Block I, maximal 600 Punkte) und Prüfungsblockpunktzahl (Block II, maximal 300 Punkte)? Generell werden hier hohe Assoziationen erwartet, wobei die stärksten Zusammenhänge aufgrund seines hohen Gewichtes zwischen dem Kursblock und der Gesamtpunktzahl bestehen sollten, während für die Kurs- und Prüfungsblockpunktzahlen geringere (aber immer noch substanzielle) Zusammenhänge erwartet werden. In der Konsequenz sollten die Zusammenhänge zwischen Prüfungsblockpunktzahl und Gesamtpunktzahl dazwischen rangieren.
2. Wie stellt sich das mittlere Notenniveau von Gesamtpunktzahl, Kursblock- und Prüfungsblockpunktzahlen im Vergleich dar? Ausgehend von den in Abschnitt 1 und 2 berichteten Befunden auf Ebene der einzelnen Fächer, die überwiegend höhere Kursnoten indizieren, wird hier im Mittel ein niedrigeres Niveau des Prüfungsnotenblocks im Vergleich zum Kursblock erwartet. Die mittlere Gesamtpunktzahl müsste sich entsprechend dazwischen bewegen.
3. In welchem Maß bzw. für welche Schüleranteile wäre bei ausschließlicher Berücksichtigung der Kursnoten für die Gesamtqualifikation eine Verbesserung oder eine Verschlechterung der Gesamtabiturnote zu erwarten? Wie stark streuen also die Differenzen zwischen Kurs- und Prüfungsnoten zwischen den Schülerinnen und Schülern? Wir gehen hier von nicht unerheblichen Unterschieden zwischen den Schülerinnen und Schülern aus und erwarten für den überwiegenden Teil, korrespondierend mit den Annahmen aus Fragestellung 2, bessere Ergebnisse im Kursblock im Vergleich zum Prüfungsblock (und somit einen Notenvorteil zugunsten eines Anerkennungsabiturs).

3.2 Methodisches Vorgehen

3.2.1 Daten

Die Datengrundlage der vorliegenden Untersuchung bildet die BERLIN-Studie[12] zur Reform der Berliner Schulstruktur, deren Kernbestandteil die Umstellung auf ein zweigliedriges Schulsystem darstellt (Neumann et al., 2017a). Im Rahmen der BERLIN-Studie wurden in einem Mehrkohortenlängsschnittdesign Schülerinnen und Schüler sowohl aus nichtgymnasialen Schulformen als auch aus allgemeinbildenden Gymnasien zu mehreren Erhebungszeitpunkten getestet und befragt. Ergänzend

12 Die BERLIN-Studie ist ein Kooperationsprojekt des Max-Planck-Instituts für Bildungsforschung (MPIB, Berlin, Principal Investigator: Prof. Dr. Jürgen Baumert), des DIPF | Leibniz-Instituts für Bildungsforschung und Bildungsinformation (DIPF, Frankfurt am Main/Berlin, Principal Investigator: Prof. Dr. Kai Maaz) und des Leibniz-Instituts für die Pädagogik der Naturwissenschaften und Mathematik (IPN, Kiel, Principal Investigator: Prof. Dr. Olaf Köller). Die Untersuchung wird durch Mittel der Senatsverwaltung für Bildung, Wissenschaft und Forschung (SenBWF) des Landes Berlin und der Jacobs-Foundation (Projekt-Nr. 2013-1083) gefördert.

wurden Informationen aus den Schulakten erhoben (vgl. zu weiteren Einzelheiten im Erhebungsdesign, Stichprobe und Instrumentierung Neumann et al., 2017b; Becker et al., 2017). Die Analysen der vorliegenden Untersuchung basieren auf einer Teilstichprobe der BERLIN-Studie von 400 Abiturientinnen und Abiturienten mit bestandener Abiturprüfung aus 29 allgemeinbildenden Gymnasien (Abitur 2014), von denen vollständige Zeugnisangaben (erhoben zum Schuljahresende 2013/14 über die Schulakten) vorliegen. Für die Stichprobe der BERLIN-Studie wurden pro Schule 25 15-jährige Schülerinnen und Schüler (aus den Jahrgängen 7–12) sowie 10 nicht 15-jährige Neuntklässlerinnen und Neuntklässler gezogen (vgl. Becker et al., 2017). Schülerinnen und Schüler, die zum Zeitpunkt der Stichprobenziehung die neunte Jahrgangsstufe besuchten, wurden im Längsschnitt weiterverfolgt. Schülerinnen und Schüler, die die Schule vor Erwerb des Abiturs verlassen hatten (z. B. aufgrund von Abbruch oder Schulwechsel), sind nicht mehr enthalten. Insgesamt ist jedoch von einer weitgehend repräsentativen Stichprobe von Abiturientinnen und Abiturienten an allgemeinbildenden Berliner Gymnasien für den Abiturjahrgang 2014 auszugehen.

3.2.2 Statistische Analysen

Die Auswertungen sind überwiegend deskriptiv. Die Zusammenhangsanalysen (Fragestellung 1) erfolgen mittels Korrelationsberechnungen. Mittelwertsunterschiede im Notenniveau (Fragestellung 2) werden mittels *t*-Tests für abhängige Stichproben auf statistische Signifikanz geprüft. Die Streuung der Differenzen von Kurs- und Prüfungsnoten (Fragestellung 3) wird anhand der Standardabweichung sowie einer kategorialen Betrachtung der Differenzen untersucht. Die Analysen wurden mit der Statistiksoftware IBM SPSS (Version 23) durchgeführt.

3.3 Ergebnisse

In einem ersten Schritt wurden hinsichtlich Fragestellung 1 zunächst die korrelativen Zusammenhänge zwischen Kurs- und Prüfungsblockpunktzahl sowie erreichter Gesamtpunktzahl betrachtet (s. Tabelle 8). Wie erwartet, fanden sich durchweg hohe Zusammenhänge zwischen den drei Notenblöcken, wobei die höchsten Korrelationen zwischen Kursblockpunktzahl und Gesamtpunktzahl bestehen. Die Korrelation fällt mit $r = .97$ ($p < .001$) annähernd perfekt aus. Mit Blick auf die Aussagekraft eines Anerkennungsabiturs ist jedoch zu berücksichtigen, dass sich trotz perfekter korrelativer Zusammenhänge dennoch Niveauunterschiede finden können (siehe Ergebnisse zu Fragestellung 2). Die geringste (aber immer noch substanzielle) Korrelation in Höhe von $r = .84$ ($p < .001$) findet sich zwischen Kurs- und Prüfungsblockpunktzahl. Die Ergebnisse in den vier absolvierten Prüfungsfächern (davon zwei Leistungskurse) weisen somit hohe Zusammenhänge mit dem erreichten Kursergebnis über alle in die Gesamtqualifikation eingebrachten Kurse auf. Die Korrelation zwischen Prüfungsblockpunktzahl und Gesamtpunktzahl ist etwas geringer als die Korrelation von Gesamtpunktzahl und Kursblockpunktzahl, fällt jedoch mit $r = .94$ ($p < .001$) ebenfalls sehr hoch aus.

Tabelle 8: Korrelationen zwischen Punktzahl in Kurs- und Prüfungsblock sowie Gesamtpunktzahl

	Kursblock	Prüfungsblock	Gesamtpunktzahl
Kursblock	1	0,84	0,97
Prüfungsblock		1	0,94
Gesamtpunkzahl			1

Anmerkung: alle Korrelationen auf dem p < .001-Niveau statistisch signifikant

Tabelle 9 gibt einen Überblick über die im Mittel erreichten Punktzahlen in Kurs- und Prüfungsblock sowie die Gesamtpunktzahl (Fragestellung 2). Im oberen Teil sind die jeweiligen Werte (nebst zugehöriger Streuungsangaben) in der Originalmetrik ausgewiesen. Im unteren Teil finden sich gewichtete Angaben von Kurs- und Prüfungsblockpunktzahlen, um einen Niveauvergleich mit der erreichten Gesamtpunktzahl (theoretischer Maximalwert = 900 Punkte) zu ermöglichen. Dazu wurden die Kursblockpunktzahlen mit dem Faktor 1,5 (600 x 1,5 = 900 Punkte) und die Prüfungsblockpunktzahlen mit dem Faktor 3,0 (300 x 3,0 = 900 Punkte) gewichtet. Aus den Ergebnissen geht zunächst hervor, dass sich Kurs- und Prüfungsblockpunktzahlen im Mittel um rund 29 Punkte zugunsten des Kursblocks unterscheiden. Übertragen auf die herkömmliche Notenmetrik (Notenstufen 1 bis 6) entspricht dies einem Unterschied von etwa 0,16 Notenstufen (p < .001). Die Differenz zwischen Kursblockpunktzahl und Gesamtpunktzahl fällt mit 9,5 Punkten deutlich geringer aus und entspricht rund 0,05 Notenstufen (p < .001). Das im Mittel erreichte Niveau der herkömmlichen Abiturgesamtnote und das eines möglichen Anerkennungsabiturs lägen also auf Basis der vorliegenden Datengrundlage nur unwesentlich auseinander. Neben feststellbaren Unterschieden in der mittleren Punktzahl deuten die Standardabweichungen auf eine größere Streuung des Prüfungsnotenblocks im Vergleich zum Kursblock hin. Die Ergebnisse im Prüfungsblock variieren also etwas stärker zwischen den Schülerinnen und Schülern als die Ergebnisse im Kursblock.

Tabelle 9: Mittelwerte, Standardabweichungen sowie Angaben zu Minima und Maxima in Kurs- und Prüfungsblock sowie Gesamtpunktzahl

	M	*SD*	*MIN*	*MAX*
Originalmetrik				
Kursblock (200–600)	403,6	68,5	248,0	597,0
Prüfungsblock (100–300)	192,2	45,5	100,0	296,0
Kursblock + Prüfungsblock (= Gesamtpunktzahl) (300–900)	595,9	109,3	372,0	893,0
Vergleichsmetrik				
Kursblock gewichtet (300–900)	605,4	102,7	372,0	895,5
Prüfungsblock gewichtet (300–900)	576,7	136,6	300,0	888,0

Anmerkung: M = Mittelwert, SD = Standardabweichung, MIN = Minimum, MAX = Maximum

Abschließend sollen die Streuungen der Differenzen zwischen den drei Notenblöcken (Kurs/Prüfung/Gesamt) betrachtet werden (Fragestellung 3). Dazu sind in Tabelle 10 zunächst die Mittelwerte und Streuungsangaben für die Differenzen von Kurs- und Prüfungsblockpunktzahlen sowie Gesamtpunktzahl und Kursblockpunktzahl ausgewiesen. Wie den Standardabweichungen zu entnehmen ist, variiert die Differenz der Punktzahlen in Kurs- und Prüfungsblock nicht unerheblich zwischen den Schülerinnen und Schülern, die Variation für die Differenz von Gesamtpunktzahl und Kursblock fällt hingegen geringer aus.

Tabelle 10: Mittelwerte und Streuungen der Differenzen von gewichteter Punktzahl in Kurs- und Prüfungsblock sowie Kursblock und Gesamtpunktzahl

	M	*SD*	*MIN*	*MAX*
Differenz Kursblock-Prüfungsblock	28,67	75,95	-202,50	246,00
Differenz Gesamtpunktzahl-Kursblock	-9,48	25,31	-82,00	67,50

Anmerkung: M = Mittelwert, SD = Standardabweichung, MIN = Minimum, MAX = Maximum

Die Tabellen 11 und 12 veranschaulichen die Streuungen nochmals mittels kategorialer Darstellung. Aus Tabelle 11 geht zunächst hervor, dass die Differenzen zwischen Kurs- und Prüfungsblockpunktzahlen sowohl in die eine (Kursblock > Prüfungsblock) als auch in die andere Richtung (Kursblock < Prüfungsblock) ausfallen können, für den überwiegenden Teil der Schülerinnen und Schüler (d. h. für rund 63 %) jedoch höhere Punktzahlen im Kursblock zu verzeichnen sind. Für einen Teil der Schülerinnen und Schüler in den Randbereichen der Verteilung kann es dabei durchaus zu beträchtlichen Abweichungen zwischen Kurs- und Prüfungsblock kommen.

Tabelle 11: Streuung der Differenz von gewichteten Kursblock- und Prüfungsblockpunkten

	Größe der Differenz	**Anteil in Prozent**
Kursblockpunktzahl < Prüfungsblockpunktzahl (35,75 %)	-100 bis -249,99	4,75
	-75 bis -99,99	4,25
	-50 bis -74,99	5,25
	-25 bis -49,99	9,75
	0 bis -24,99	11,75
Kursblockpunktzahl = Prüfungsblockpunktzahl (1,0 %)	0	1,00
Kursblockpunktzahl > Prüfungsblockpunktzahl (63,25 %)	0 bis 25	13,50
	25 bis 49,99	12,00
	50 bis 74,99	8,75
	75 bis 99,99	10,50
	100 bis 249,99	18,50

Tabelle 12 weist mit Blick auf die übergreifende Leitfrage des Abschnitts abschließend die Anteile der Schülerinnen und Schüler aus, die von einem Anerkennungsabitur profitieren würden, bei dem ausschließlich der Kursblock die Grundlage für die erreichte Gesamtqualifikation bilden würde. Dies wäre auf Basis der vorliegenden Datengrundlage bei rund zwei Dritteln der Schülerinnen und Schüler der Fall. Für knapp 30 Prozent würden sich Vorteile von 25 Punkten (bzw. 0,14 Notenstufen auf der herkömmlichen Notenskala von 1 bis 6) oder mehr ergeben. Gut ein Drittel der Schülerinnen und Schüler würde schlechter abschneiden, jedoch würde der Nachteil bei nur 9 Prozent von ihnen 25 Punkte oder mehr umfassen.

Tabelle 12: Streuung der Differenz von Gesamtpunktzahl und gewichteter Kursblockpunktzahl

	Differenzrange	Anteil in Prozent
Gesamtpunktzahl < Kursblockpunktzahl (63,0 %)	-50 bis -100	5,00
	-25 bis -49.99	23,75
	0 bis -24.99	34,25
Gesamtpunktzahl = Kursblockpunktzahl (1,0 %)	0	1,00
Gesamtpunktzahl > Kursblockpunktzahl (36,0 %)	0 bis 25	27,00
	25 bis 49.99	8,50
	50 bis 100	0,50

3.4 Fazit

Die Befunde zum Verhältnis von Punktzahlen aus Kursblock, Prüfungsblock und Gesamtpunktzahl, die anhand einer Stichprobe von Berliner Abiturientinnen und Abiturienten ermittelt wurden, lassen sich wie folgt zusammenfassen: Erstens bestehen zwischen den drei globalen Ergebnissen, die im Abiturzeugnis ausgewiesen werden, sehr hohe korrelative Zusammenhänge. Dies gilt, wenn auch etwas abgeschwächter, auch für den Zusammenhang von erreichten Kurs- und Prüfungsblockpunktzahlen. Zweitens lässt sich ein im Mittel höheres Niveau des Kursnotenblocks im Verhältnis zum Prüfungsblock konstatieren, was sich gut in die Befundlage zu Differenzen von Kurs- und Prüfungsnoten für einzelne Fächer einfügt (vgl. die Abschnitte 1 und 2 in diesem Beitrag) und dieses Ergebnismuster somit bestätigt. Mit 0,16 Notenstufen auf der herkömmlichen Notenskala von 1 bis 6 erscheint das Ausmaß des festgestellten Unterschieds durchaus bedeutsam. Die praktische Bedeutsamkeit dieses Unterschieds wird allerdings durch einen Blick auf die lediglich 0,05 Notenstufen umfassende Differenz zwischen der Kursblockpunktzahl und der die Prüfungsleistung beinhaltenden Gesamtpunktzahl relativiert: Im Prüfungsblock werden zwar schlechtere Ergebnisse erzielt als im Kursblock, auf die Abiturgesamtnote hat dies jedoch nur einen geringen mittleren Effekt. Drittens differiert der Unterschied zwischen Kurs- und Prüfungsblockpunktzahlen deutlich zwischen den Schülerinnen und Schülern.

Für rund zwei Drittel der Prüflinge ergibt sich ein besseres Ergebnis für den Fall einer Beschränkung auf die Kursergebnisse, für rund ein Drittel hingegen ein schlechteres.

Die Befunde liefern damit empirische Hinweise dafür, dass ein Verzicht auf die Einbeziehung von Abiturprüfungsleistungen in die Gesamtqualifikation für den überwiegenden Teil der Schülerinnen und Schüler Vorteile für die Abiturgesamtnote mit sich bringen könnte. Für einen gewissen Anteil durchaus in nicht unerheblichem Ausmaß, wenn man die Bedeutung von Zehntelnotenstufen für manche Auswahlprozesse beim Übergang ins Studium oder die Berufsausbildung bedenkt. Gleichwohl ist für die Einordnung der Befunde ganz klar auf die vorhandenen Einschränkungen hinzuweisen. So handelt es sich ausschließlich um eine Betrachtung von Berliner Abiturientinnen und Abiturienten eines Schülerjahrgangs aus allgemeinbildenden Gymnasien. Die Befunde sind also nicht ohne Weiteres generalisierbar auf andere Länder, Zeitpunkte oder Schulformen, an denen ebenfalls das Abitur erworben werden kann.

Diskussion und Ausblick

Der vorliegende Beitrag fasst drei empirische Arbeiten zusammen, in denen, anhand jeweils unterschiedlicher Daten und mit unterschiedlichen Schwerpunktsetzungen, die Passung und der Zusammenhang von Abiturprüfungsnoten und Kursnoten untersucht wurden. Insgesamt bestätigen die Ergebnisse der hier beschriebenen Analysen die Annahme, dass die Abiturprüfungsnoten in der Regel schlechter ausfallen als die in der Qualifikationsphase erzielten Kursnoten. Dieses Resümee ist allerdings mit verschiedenen Einschränkungen verbunden. Erstens variiert die Differenz zwischen Prüfungsnoten und Kursnoten deutlich zwischen den Fächern. So weisen insbesondere die Ergebnisse der in Abschnitt 1 dargestellten Metaanalysen darauf hin, dass diese Differenz im Fach Englisch deutlich geringer ausfällt als zum Beispiel im Fach Mathematik. Ferner scheint die Differenz von Prüfungsnoten und Kursnoten zum Teil erheblich zwischen den Ländern und auch zwischen den Schulen eines Landes zu variieren. In nicht wenigen Ländern und Schulen gelingt es den Schülerinnen und Schülern sogar, in den Abiturprüfungen im Mittel bessere Ergebnisse zu erzielen als in der Qualifikationsphase. Plausibel erscheint dabei, dass diese Länder- und Schuleffekte maßgeblich Länder- und Schulunterschiede in den jeweils im Mittel erreichten Kompetenzen widerspiegeln. Bemerkenswert sind in diesem Zusammenhang außerdem die in Abschnitt 2 dargestellten Befunde, die vermuten lassen, dass Länderunterschiede in der Passung von Prüfungsnoten und Kursnoten im Zusammenhang mit Länderunterschieden bei der Gliederung und Organisation der Oberstufe stehen könnten. Schließlich verdeutlichen die in Abschnitt 3 dargestellten Befunde zu den auf der Blockebene eingebrachten Punkten, dass ein Verzicht auf den Prüfungsblock die Abiturgesamtnote zwar anheben würde, der Effekt im Mittel aber vermutlich eher klein ausfiele.

In der Gesamtschau legen die in diesem Beitrag dargestellten Befunde also den Schluss nahe, dass ein pandemiebedingtes Anerkennungsabitur zwar vermutlich in

der Tendenz zu einer geringfügigen Verbesserung der Abiturgesamtnote geführt hätte, allerdings nicht zu einer erheblichen Noteninflation. Mit Blick auf die Übertragbarkeit der Befunde auf den hypothetischen Fall eines Anerkennungsabiturs muss allerdings darauf hingewiesen werden, dass in den Studien dieses Beitrags lediglich die Unterschiede zwischen Prüfungsnoten und Kursnoten betrachtet werden, während mögliche Auswirkungen eines Verzichtes auf gesamtqualifikationsrelevante Abiturprüfungen für das Lernen und Unterrichten völlig unberücksichtigt bleiben. Fasst man die Abiturprüfungen als einen externen Motivationsanreiz auf (wovon für einen großen Teil der Schülerinnen und Schüler sicherlich ausgegangen werden kann), könnte der Verzicht auf abschließende Prüfungen zum Beispiel Einbußen in Lernmotivation und Lernerfolgen nach sich ziehen, die sich auch in den erzielten Kursergebnissen, vor allem aber den tatsächlich erreichten Kompetenzen, niederschlagen dürften.

Vor dem Hintergrund des eingangs skizzierten Diskurses zum Anerkennungsabitur ist allerdings bemerkenswert, dass die Abiturgesamtnoten der Absolventinnen und Absolventen des Prüfungsjahrgangs 2021, die in besonderem Maße von den Auswirkungen der Pandemie auf das schulische Lernen betroffen waren, trotz des Festhaltens an den Abiturprüfungen schlussendlich deutlich besser ausgefallen sind als in den Jahren zuvor.[13] Als Gründe hierfür werden im Allgemeinen die Nachteilsausgleiche angeführt, die den Abiturientinnen und Abiturienten des Prüfungsjahrgangs 2021 eingeräumt wurden. Diese waren je nach Land unterschiedlich und umfassten zum Beispiel eine Erhöhung der Bearbeitungszeit sowie eine Erweiterung der Wahlmöglichkeiten. In Mecklenburg-Vorpommern wurden sogar die Ergebnisse der schriftlichen Abiturprüfungen im Fach Mathematik (auch mit Verweis auf eine im Jahr 2021 neu eingefügte Abiturprüfungsverordnung und ungewohnte Aufgabenstellungen) um zwei Notenpunkte angehoben.

Die drei empirischen Arbeiten, die im vorliegenden Beitrag zusammengefasst sind, erweitern die Befundlage zur Passung und zum Zusammenhang von Abiturprüfungsnoten und Kursnoten. Dennoch bleiben einige Fragen offen. Mit Blick auf die in den Analysen festgestellten Schuleffekte betrifft dies insbesondere die Frage, inwieweit die Passung von Prüfungsnoten und Kursnoten je nach Schulart variiert, also zum Beispiel an Gymnasien geringer ausfällt als an Gesamtschulen oder beruflichen Gymnasien. Vor dem Hintergrund, dass, wie im Beitrag 2 des vorliegenden Bandes dargestellt, die Aufgaben der Gemeinsamen Abituraufgabenpools der Länder zukünftig noch häufiger zum Einsatz kommen werden als bislang und sich mithin die Vergleichbarkeit der Abiturprüfungen der Länder weiter erhöhen wird, wäre es außerdem lohnenswert, die Passung zwischen Prüfungsnoten und Kursnoten auch für die kommenden Prüfungsjahre zu untersuchen. Mit Blick auf die in den Abschnitten 2 und 3 dargestellten Studien stellt sich insbesondere die Frage nach der Generalisierbarkeit der Befunde. Dementsprechend wünschenswert erscheint es, die Analysen zu den Effekten von Oberstufenreformen sowie die Vergleiche zwischen den im Kurs-

13 Vgl. https://www.jmwiarda.de/2021/08/02/das-rekord-abi/ [10.03.2022]

block und im Prüfungsblock eingebrachten Punkten auch für andere Länder, Zeitpunkte und Schularten durchzuführen.

Literatur

Arens, A. K., Marsh, H. W., Pekrun, R., Lichtenfeld, S., Murayama, K. & vom Hofe, R. (2017). Math self-concept, grades, and achievement test scores: Long-term reciprocal effects across five waves and three achievement tracks. *Journal of Educational Psychology, 109*(5), 621–634. https://doi.org/10.1037/edu0000163

Balduzzi, S., Rücker, G. & Schwarzer, G. (2019). How to perform a meta-analysis with R: A practical tutorial. *Evidence-Based Mental Health, 22*(4), 153–160. https://doi.org/10.1136/ebmental-2019-300117

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal for Statistical Software, 67*(1), 48. https://doi.org/10.18637/jss.v067.i01

Blossfeld, H.-P. & Roßbach, H.-G. (Hrsg.). (2019). *Education as a lifelong process: The German National Educational Panel Study* (NEPS). Edition ZfE (2. Aufl.). Springer VS.

Bölling, R. (2010). *Kleine Geschichte des Abiturs*. Paderborn: Schöningh.

Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons.

Brodkorb, M. & Koch, K. (2020). *Der Abiturbetrug. Vom Scheitern des deutschen Bildungsföderalismus. Eine Streitschrift*. Dietrich zu Klampen.

Deeks, J. J., Higgins, J. P. T. & Altman, D. G. (2021). Chapter 10: Analysing data and undertaking meta-analyses. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page & V. A. Welch (Hrsg.), *Cochrane Handbook for Systematic Reviews of Interventions version 6.2 (updated February 2021)*. Cochrane. http://www.training.cochrane.org/handbook [22.06.2021]

Domina, T. & Saldana, J. (2012). Does raising the bar level the playing field? Mathematics curricular intensification and inequality in American high schools, 1982–2004. *American Educational Research Journal, 49*(4), 685–708. https://doi.org/10.3102/0002831211426347

Dreeben, R. & Barr, R. (1988). Classroom Composition and the Design of Instruction. *Sociology of Education, 61*(3), 129–142. https://doi.org/10.2307/2112622

Higgins, J. P. T. & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539–1558.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J. & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*, 557–560.

Hildenbrand, C. (Hrsg.). (2017). Schulabschlüsse. In Behörde für Schule und Berufsbildung, *Bildungsbericht Hamburg 2017* (S. 149–168). Waxmann.

Hildenbrand, C. (Hrsg.). (2020). Schulabschlüsse. In Behörde für Schule und Berufsbildung, *Bildungsbericht Hamburg 2020* (S. 155–174). Waxmann.

Hofbauer, L. (2021). *Umsetzung der KMK Bildungsstandards in Abituraufgaben im Fach Mathematik – Ein Vergleich der Länder Bayern und Berlin* (Masterarbeit). Freie Universität Berlin.

Hoffmann, L., Schröter, P. & Stanat, P. (2018). *Evaluation von Aufgaben der Pools für das Prüfungsjahr 2017. Ergebnisse zur Bewährung der Aufgaben*. https://www.iqb.hu-berlin.de/abitur/evaluation/PoolsfrdasPrfung.pdf [22.06.2021]

Hoffmann, L., Schröter, P., & Stanat, P. (2020). *Evaluation von Aufgaben der Pools für das Prüfungsjahr 2019. Ergebnisse zur Bewährung der Aufgaben*. https://www.iqb.hu-berlin.de/abitur/evaluation/PoolsfrdasPrfung_1.pdf [22.06.2021]

Hox, J. J., Moerbeek, M. & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Hübner, N., Wagner, W., Hochweber, J., Neumann, M., & Nagengast, B. (2020). Comparing apples and oranges: Curricular intensification reforms can change the meaning of students' grades! *Journal of Educational Psychology, 112*(1), 204–220. https://doi.org/10.1037/edu0000351

Hübner, N., Wagner, W., Nagengast, B. & Trautwein, U. (2019). Putting all students in one basket does not produce equality: gender-specific effects of curricular intensification in upper secondary school. *School Effectiveness and School Improvement, 30*(3), 261–285. https://doi.org/10.1080/09243453.2018.1504801

Hübner, N., Wille, E., Cambria, J., Oschatz, K., Nagengast, B. & Trautwein, U. (2017). Maximizing gender equality by minimizing course choice options? Effects of obligatory coursework in math on gender differences in STEM. *Journal of Educational Psychology, 109*(7), 993–1009. https://doi.org/10.1037/edu0000183

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). Waxmann.

KMK – Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2021). *Vereinbarung zur Gestaltung der gymnasialen Oberstufe und der Abiturprüfung* (Beschluss der Kultusministerkonferenz vom 07.07.1972 i. d. F. vom 18.02.2021). https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/1972/1972_07_07-VB-gymnasiale-Oberstufe-Abiturpruefung.pdf [22.06.2021]

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

Kunter, M., Dubberke, T., Baumert, J., Blum, W., Brunner, M., Jordan, A. & Tsai, Y. M. (2006). Mathematikunterricht in den PISA-Klassen 2004: Rahmenbedingungen, Formen und Lehr-Lernprozesse. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, J. Rost & U. Schiefele (Hrsg.), *PISA 2003: Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres* (S. 161–194). Waxmann.

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

Maag Merki, K. (2012). *Zentralabitur*. VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-94023-6

Maag Merki, K. & Holmeier, M. (2015). Comparability of semester and exit exam grades: long-term effect of the implementation of state-wide exit exams. *School Effectiveness and School Improvement, 26*(1), 57–74. https://doi.org/10.1080/09243453.2013.861353

Maaz, K., Baumert, J., Neumann, M., Becker, M. & Dumont, H. (2013). *Die Berliner Schulstrukturreform: Bewertung durch die beteiligten Akteure und Konsequenzen des neuen Übergangsverfahrens von der Grundschule in die weiterführenden Schulen*. Waxmann.

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology, 79*(3), 280–295. https://doi.org/10.1037/0022-0663.79.3.280

Muthén, L. K. & Muthén, B. O. (1998–2017). *Mplus user's guide* (Eighth Edition). Muthén & Muthén.

Nagy, G., Neumann, M., Becker, M., Watermann, R., Köller, O., Lüdtke, O. & Trautwein, U. (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). Waxmann.

Nagy, G., Neumann, M., Trautwein, U. & Lüdtke, O. (2010). Voruniversitäre Mathematikleistungen vor und nach der Neuordnung der gymnasialen Oberstufe in Baden-Württemberg. In U. Trautwein, M. Neumann, G. Nagy, O. Lüdtke & K. Maaz (Hrsg.), *Schulleistungen von Abiturienten. Die neu geordnete gymnasiale Oberstufe auf dem Prüfstand* (S. 147–180). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92037-5_6

NEPS-Netzwerk. (2014). *Nationales Bildungspanel, Scientific Use File der Zusatzstudie Thüringen. Leibniz-Institut für Bildungsverläufe* (LIfBi), Bamberg. https://doi.org/10.5157/NEPS:TH:2.0.0

Neumann, M. (2010). Innovation oder Restauration – Die (Rück-?)Reform der gymnasialen Oberstufe in Baden-Württemberg. In U. Trautwein, M. Neumann, G. Nagy, O. Lüdtke & K. Maaz (Hrsg.), *Schulleistungen von Abiturienten: Die neugeordnete gymnasiale Oberstufe auf dem Prüfstand* (1. Aufl., S. 37–90). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92037-5\textunderscore

Neumann, M., Becker, M., Baumert, J., Maaz, K. & Köller, O. (2017a). *Zweigliedrigkeit im deutschen Schulsystem: Potenziale und Herausforderungen in Berlin*. Waxmann.

Neumann, M., Maaz, K., Baumert, J., Becker, M., Kropf, M., Jansen, M. & Köller, O. (2017b). Anlage der BERLIN-Studie und Fragestellungen des vorliegenden Bandes. In M. Neumann, M. Becker, J. Baumert, K. Maaz & O. Köller (Hrsg.), *Zweigliedrigkeit im deutschen Schulsystem: Potenziale und Herausforderungen in Berlin* (S. 39–54). Waxmann.

Neumann, M., Nagy, G., Trautwein, U. & Lüdtke, O. (2009). Vergleichbarkeit von Abiturleistungen. *Zeitschrift für Erziehungswissenschaft, 12*(4), 691–714. https://doi.org/10.1007/s11618-009-0099-6

Neumann, M. & Trautwein, U. (2019). Zwischen individueller Schwerpunktsetzung und Standardisierung – Reformen in der gymnasialen Oberstufe. In N. Berkemeyer, W. Bos & B. Hermstein (Hrsg.), *Schulreform: Zugänge, Gegenstände, Trends* (Pädagogik, S. 547–558). Beltz.

Neumann, M., Trautwein, U. & Nagy, G. (2011). Do central examinations lead to greater grading comparability? A study of frame-of-reference effects on the University entrance qualification in Germany. *Studies in Educational Evaluation, 37*(4), 206–217. https://doi.org/10.1016/j.stueduc.2012.02.002

Pekrun, R. & Götz, T. (2006). Emotionsregulation: Vom Umgang mit Prüfungsangst. In H. Mandl & H. F. Friedrich (Hrsg.), *Handbuch Lernstrategien* (S. 248–258). Hogrefe.

R Development Core Team. (2021). *R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing*. http://www.R-project.org

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

Stanat, P., Böhme, K., Schipolowski, S. & Haag, N. (Hrsg.). (2016). *IQB-Bildungstrend 2015. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich*. Waxmann.

Stanat, P., Schipolowski, S., Mahler, N., Weirich, S. & Henschel, S. (Hrsg.). (2019). *IQB-Bildungstrend 2018. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich*. Waxmann.

Thüringer Kultusministerium. (2008). *Informationen zur neuen gymnasialen Oberstufe*. Thüringer Kultusministerium.

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

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O. & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology, 98*(4), 788–806. https://doi.org/10.1037/0022-0663.98.4.788

Trautwein, U. & Neumann, M. (2008). Das Gymnasium. In K. S. Cortina, J. Baumert, A. Leschinsky, K. U. Mayer & L. Trommer (Hrsg.), *Das Bildungswesen in der Bundesrepublik Deutschland: Strukturen und Entwicklungen im Überblick* (Bd. 62339, S. 467–501). Rowohlt.

Trautwein, U., Neumann, M., Nagy, G., Lüdtke, O. & Maaz, K. (Hrsg.). (2010). *Schulleistungen von Abiturienten: Die neugeordnete gymnasiale Oberstufe auf dem Prüfstand* (1. Aufl.). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92037-5

Tabellenverzeichnis

10 Mathematische Lernvoraussetzungen für MINT-Studiengänge – eine Delphi-Studie mit Hochschullehrenden

Aiso Heinze, Irene Neumann & Christoph Deeken

Zusammenfassung

Mathematik scheint für Studienanfängerinnen und Studienanfänger im MINT-Bereich eine große Herausforderung darzustellen. Das vorliegende Kapitel nimmt daher die mathematischen Anforderungen in der Übergangsphase von der Schule in ein MINT-Studium genauer in den Blick. Berichtet werden zwei aufeinander aufbauende Studien. In der ersten Studie wurde im Rahmen einer Delphi-Befragung mit Hochschullehrenden erfasst, welche mathematischen Lernvoraussetzungen Hochschulen von Studienanfängerinnen und Studienanfängern im MINT-Bereich erwarten. In einer zweiten Studie, von der erste Teilergebnisse berichtet werden, wurden diese Erwartungen der Hochschulseite mit den normativen Vorgaben der Lehrpläne für das Fach Mathematik in der Sekundarstufe abgeglichen. Damit wird untersucht, inwieweit die Erwartungen der Hochschullehrenden sich in den laut Lehrplänen zu erreichenden mathematischen Kompetenzen bis zum Abitur widerspiegeln. Das Kapitel schließt mit einer Diskussion und Bewertung der Ergebnisse.

1 Einleitung

Abiturientinnen und Abiturienten zu einem Hochschulstudium zu befähigen, ist ein zentrales Ziel der gymnasialen Oberstufe. Der Mathematik wird dabei eine besondere Bedeutung beigemessen und ist ein Pflichtfach für alle Schülerinnen und Schüler (KMK, 2021). Trotz dieser Bedeutung scheint Mathematik jedoch eine große Herausforderung für viele Studienanfängerinnen und Studienanfänger darzustellen, insbesondere zu Beginn eines Studiums im Bereich MINT (Mathematik, Informatik, Naturwissenschaften, Technik). So beklagen nicht nur Hochschullehrende mangelnde Mathematikkenntnisse ihrer Studierenden (z. B. Offener Brief, 2017), auch die Studierenden selbst äußern Probleme mit Mathematik (z. B. Albrecht & Nordmeier, 2011) oder fühlen sich nicht ausreichend durch den schulischen Mathematikunterricht vorbereitet (z. B. Bescherer, 2003; Heublein et al., 2017). Diese berichteten Erfahrungen, und nicht zuletzt die – gerade im MINT-Bereich – konstant hohen Studienabbruchquoten (vgl. Heublein & Wolter 2011; Heublein et al. 2017) sind Symptome eines nicht reibungslosen Übergangs von der Institution Schule in die Institution Hochschule, für dessen Verbesserung auch mathematische Fachverbände Handlungsbedarf sehen (DMV, GMD & MNU, 2019).

Grundsätzlich sind Übergänge zwischen Institutionen Auslöser für „Passungs- und Abstimmungsprobleme [...] auf einer qualitativen Ebene (wie den Anforderungen und Voraussetzungen)“ (Wolter, 2013, S. 46). Mit Blick auf die Mathematik sind die Voraussetzungen für den Übergang an die Hochschule festgelegt durch Standards und Lehrpläne, die die Bildungsziele des schulischen Mathematikunterrichts beschreiben. Dass jedoch die Anforderungen, die die Hochschullehrenden in Form erwarteter Lernvoraussetzungen an Studienanfängerinnen und Studienanfänger stellen, auch den Bildungszielen des Mathematikunterrichts entsprechen, ist nicht geregelt. Dieses Problem ist insbesondere für das Fach Mathematik nicht neu (Behnke, 1965) und eine (staatliche) Regelung aufgrund der grundgesetzlich garantierten Hochschulautonomie nur schwer möglich. Umso wichtiger ist daher die Frage, welche mathematischen Lernvoraussetzungen konkret von Studienanfängerinnen und Studienanfängern im MINT-Bereich seitens der Hochschullehrenden erwartet werden. Denn diese Erwartungen werden die Ausgestaltung der Lehrveranstaltungen prägen und damit einen entscheidenden Faktor darstellen, ob und inwieweit Studienanfängerinnen und Studienanfänger Hürden im hochschulischen Lernen erfahren. Sind diese erwarteten mathematischen Lernvoraussetzungen der Hochschulseite bekannt, dann kann ein Abgleich stattfinden, ob Hochschulen mehr erwarten als Schulen bis zur Hochschulreife vermitteln sollen. In diesem Beitrag werden zwei aufeinander aufbauende Studien vorgestellt, die dieses untersucht haben. In der ersten Studie wird über die Delphi-Befragung MaLeMINT berichtet, in der die von Hochschullehrenden erwarteten mathematischen Lernvoraussetzungen für MINT-Studiengänge in Deutschland ermittelt wurden. In der zweiten, noch laufenden Studie werden erste Ergebnisse eines Abgleichs dieser erwarteten Lernvoraussetzungen mit den laut Lehrplan zu erreichenden mathematischen Kompetenzen bis zum Abitur berichtet. Dazu wird exemplarisch der Lehrplan des eigenen Bundeslands Schleswig-Holstein herangezogen.

2 Anforderungskataloge mit Erwartungen von Hochschullehrenden an Studienanfängerinnen und Studienanfänger

Dass Hochschullehrende Schulwissen im Fach Mathematik für einen wichtigen Teil von Studierfähigkeit erwarten, zeigten bereits die Befragungen von Hochschullehrenden zur (allgemeinen) Studierfähigkeit, die von Heldmann (1984) und Konegen-Grenier (2001) durchgeführt wurden. Bei Konegen-Grenier zeigte sich dabei der Stellenwert mathematischen Schulwissens insbesondere für die MINT-Fächer. Auch wenn diese Studie zwar die Meinung von einer großen Stichprobe von Lehrenden an Hochschulen aus ganz Deutschland widerspiegelte, lässt sie keine Schlüsse zu, welche mathematischen Aspekte genau von den Hochschullehrenden erwartet werden, da lediglich um eine Einschätzung der Relevanz des Schulwissens zu Mathematik im Allgemeinen gebeten wurde.

Detailliertere Einblicke in die Erwartungen von Hochschullehrenden lassen sich aus der Arbeit einzelner Kommissionen und Arbeitsgruppen ableiten. So veröffentlichte beispielsweise die Konferenz der Fachbereiche Physik (KFP) 2012 eine „Empfehlung [...] zum Umgang mit den Mathematikkenntnissen von Studienanfängern der Physik" (KFP, 2012). Darin werden mathematische Inhalte gelistet, die in den ersten Semestern des Physikstudiums relevant sind (z. B. Gleichungen umstellen, Bruchrechnung, komplexe Zahlen, Folgen, rationale und trigonometrische Funktionen, Exponential- und Logarithmusfunktion, Differenziations- und Integrationsregeln, Taylorentwicklung, Differenzialoperatoren, Differenzialgleichungen, Rechenregeln für Vektoren und Matrizen, Gauß-Algorithmus, Zufallsexperiment, Binomialverteilung). Auf Basis der damals gültigen Lehrpläne in den 16 Bundesländern gab die KFP für alle aufgeführte Aspekte an, ob sie zu Beginn eines Physikstudiums von den Studierenden erwartet werden können oder erst im Studium thematisiert werden sollten.

Mit Blick auf den Bereich WiMINT (Wirtschaft und MINT) erarbeiteten die Mitglieder der cooperation schule:hochschule (cosh, 2014) den „Mindestanforderungskatalog Mathematik (Version 2.0)" (cosh, 2014) für Studierende an einer Hochschule in Baden-Württemberg. Der sogenannte cosh-Katalog beinhaltet einerseits allgemeine mathematische Kompetenzen, die zu Studienbeginn vorausgesetzt werden (z. B. Sachverhalte mathematisch modellieren, Größenordnungen abschätzen, Fachsprache und Fachsymbolik verstehen und verwenden, Fallunterscheidungen vornehmen). Andererseits werden mathematische Inhalte als notwendige Lernvoraussetzungen gelistet (z. B. Terme umformen, Rechnen mit Brüchen, Potenzen und Wurzeln, Gleichungen und Ungleichungen lösen, Differenzial- und Integralregeln, geometrische Interpretation von linearen Gleichungssystemen, Addition und Skalarmultiplikation von Vektoren). Darüber hinaus werden Beispielaufgaben angeführt, in denen diese Lernvoraussetzungen konkretisiert wurden.

Für die Ingenieurwissenschaften wurde von der Europäischen Gesellschaft für Ingenieur-Ausbildung SEFI ein „Framework for Mathematics Curricula in Engineering Education" (SEFI, 2013) vorgelegt. Das Framework benennt, welche allgemeinen mathematischen Kompetenzen (z. B. mathematisch argumentieren oder modellieren) und welche mathematischen Inhalte (u. a. auch auf dem Niveau, wie es zu Studienbeginn erwartet wird, z. B. im Bereich Algebra, Analysis oder Geometrie) für ein Ingenieurwissenschaftliches Studium wichtig sind. Darüber hinaus werden Einstellungen gegenüber dem Fach Mathematik als notwendige Voraussetzungen angeführt.

Diese drei Beispiele – KFP, cosh, SEFI – zeigen, dass von Hochschullehrenden erwartete mathematische Lernvoraussetzungen deutlich differenzierter formuliert werden können als lediglich die allgemeine Erwartung mathematischen Schulwissens. Ferner zeigt sich ein gewisser Überlapp, ja sogar ein gemeinsamer Kern der drei oben genannten Kataloge. Es sind jedoch auch Unterschiede zu konstatieren. So benennen alle drei genannten Kataloge mathematische Inhalte (wie das Bruchrechnen, Differenzieren oder Integrieren), jedoch werden anwendungsbezogene Lernvoraussetzungen (z. B. Modellieren) nur bei cosh und SEFI aufgeführt, und SEFI geht darüber hinaus sogar noch auf Persönlichkeitsmerkmale ein. Inwieweit ein Konsens über

mathematische Lernvoraussetzungen unter Hochschullehrenden aus verschiedenen Bundesländern und für verschiedene Fächer des MINT-Bereichs angenommen werden kann, ist allein auf Grundlage von Katalogen einzelner Kommissionen und Arbeitsgruppen daher noch unklar.

3 Delphi-Studie MaLeMINT

Vor diesem Hintergrund zielte das Projekt MaLeMINT (Mathematische Lernvoraussetzungen für MINT-Studiengänge) darauf, zu untersuchen, inwieweit ein Konsens unter Hochschullehrenden in Deutschland darüber besteht, welche mathematischen Lernvoraussetzungen von Studienanfängerinnen und Studienanfängern für einen erfolgreichen Einstieg in ein MINT-Studium erwartet werden. Im Falle eines Konsenses hätte ein solcher systematisch erarbeiteter und empirisch abgesicherter Katalog das Potenzial, für einen Abgleich von Erwartungen auf der Hochschulseite einerseits und normativen Vorgaben (wie Abiturbildungsstandards, Lehrpläne für die Oberstufe u. ä.) auf der Schulseite andererseits richtungsweisend zu sein. Im Folgenden wird die MaLeMINT-Studie in ihrer Anlage, der Methodik und den Ergebnissen im Überblick vorgestellt. Für weitere Details, insbesondere den Katalog an mathematischen Lernvoraussetzungen, sei auf die zugehörigen Publikationen verwiesen (Deeken et al., 2020; Neumann et al., 2017).

Um dieses Ziel zu erreichen wurde eine bundesweit angelegte Befragung von Expertinnen und Experten durchgeführt, in der die Delphi-Methode genutzt wurde (vgl. Häder, 2014; Webler et al., 1991). Diese zeichnet sich dadurch aus, dass über mehrere Befragungsrunden hinweg die Einschätzungen einer Gruppe – hier zu Erwartungen an mathematische Lernvoraussetzungen für das MINT-Studium seitens der Hochschullehrenden – eingeholt wird. Nach jeder Runde werden die individuellen Angaben der Expertinnen und Experten vom Projektteam zusammengefasst und strukturiert. Diese Ergebnisse werden dann den einzelnen Expertinnen und Experten in einer weiteren Befragungsrunde zu einer erneuten Bewertung vorgelegt. Dabei werden lediglich aggregierte Ergebnisse mitgeteilt, ohne offenzulegen, wer welche Meinung eingebracht hat. So kann über mehrere Runden sukzessiv ein potenzieller Konsens ermittelt werden, ohne dass dabei Effekte der sozialen Beeinflussung zum Tragen kommen (z. B. Meinungsführerschaft von Einzelpersonen) wie diese beispielsweise in Gruppendiskussionen auftreten können (Häder, 2014). Mit dieser Methode lässt sich auch feststellen, ob es überhaupt einen Konsens unter den Expertinnen und Experten gibt, wie weitreichend dieser ist, und ob er gegebenenfalls nur für Subgruppen festzustellen ist.

3.1 Teilnehmende

Für die Aussagekraft der Ergebnisse von Delphi-Studien ist die angemessene Auswahl von Expertinnen und Experten von besonderer Relevanz. Mit Blick auf die Fragestellung, die in MaLeMINT beantwortet werden sollte, wurden daher zu Projektbeginn im Jahr 2015 Hochschullehrende als Expertinnen und Experten ausgewählt, die in

MINT-Studiengängen im Zeitraum 2010–2015 Mathematikveranstaltungen für das erste Semester angeboten hatten. Diese Hochschullehrenden wurden aus mehreren Gründen als geeignete Expertinnen und Experten angesehen. Zunächst ist davon auszugehen, dass Hochschullehrende bei der Gestaltung ihrer Lehrveranstaltungen (z. B. hinsichtlich des angelegten Anforderungsniveaus, der Auswahl von Beispielen etc.) Erwartungen über Lernvoraussetzungen bei ihren Studierenden zugrunde legen. Da in der Regel Mathematiklehrveranstaltungen im ersten Semester in MINT-Studiengängen verpflichtend zu belegen sind und ihr Bestehen obligatorisch für den weiteren Studienverlauf ist, kann man also annehmen, dass die Erwartungen von Hochschullehrenden in diesen Lehrveranstaltungen hinsichtlich eines erfolgreichen Übergangs von der Schule in ein MINT-Studium bedeutsam sind. Um nicht nur das Bild für ein bestimmtes Jahr widerzuspiegeln und dennoch die Aktualität zu wahren, wurde ein Zeitraum von fünf Jahren vor Projektbeginn gewählt, in dem die Hochschullehrenden ihre Veranstaltung durchgeführt haben mussten. Da die allgemeine Hochschulreife gleichermaßen für die Aufnahme an einer (Fach-)Hochschule und einer Universität befähigt, wurden Hochschullehrende an beiden Typen von Hochschulen einbezogen. Um Hochschullehrende zu identifizieren, die obiges Kriterium erfüllten, wurden online Vorlesungsverzeichnisse, Modulhandbücher und Stundenpläne gesichtet. So konnten 2233 Hochschullehrende ermittelt werden, die als Expertinnen und Experten zur Delphi-Befragung eingeladen wurden[1].

3.2 Anlage der Befragung

Die Delphi-Studie war als Online-Befragung angelegt und umfasste insgesamt drei Runden.

Befragungsrunde 1

Die erste Runde wurde als explorative Befragung angelegt (vgl. Häder, 2014). Das bedeutet, dass seitens des Projektteams keine Lernvoraussetzungen vorgegeben wurden, die lediglich hätten bewertet werden müssen. Vielmehr sollten die Erwartungen seitens der Hochschullehrenden möglichst unbeeinflusst erfasst werden. Dazu wurden in Anlehnung an die Critical Incident Technique nach Flanagan (1954) drei erzählgenerierende Fragen als Impulse verwendet. Die Impulse adressierten:

(1) Die Studierfähigkeit im Bereich Mathematik: Was kennzeichnet mathematikbezogene Studierfähigkeit für die MINT-Studiengänge? Was sollten Erstsemesterstudierende für ein MINT-Studium mitbringen? Welches Fähigkeitsniveau sollten Studierende zu Beginn des ersten Semesters mindestens aufweisen (in Hinblick auf eine für den Mathematikunterricht typische und exemplarisch ausgewählte Aufgabe)?
(2) Die Konzeptionierung von mathematischen Orientierungstests für das MINT-Studium: Stellen Sie sich bei der Beantwortung der folgenden Unterfragen bitte vor,

1 Von allen eingeladenen Hochschullehrenden konnten 2138 erreicht werden. Es gab in der Gruppe einzelne Hochschullehrende, die keine eigene Lehrveranstaltung in Mathematik angeboten hatten, aber aufgrund besonderer Funktionen (u. a. Modulverantwortung) aufgenommen wurden.

dass Sie bei der Entwicklung eines mathematischen Orientierungstests für Interessierte an einem MINT-Studium mitwirken. a) Welche mathematischen Aspekte sollten Ihrer Erfahrung nach in diesem Test unbedingt abgedeckt werden? b) Formulieren Sie 1–2 Aufgaben (bzw. Aufgabenideen) zur Illustrierung der 1–2 aus Ihrer Sicht bedeutendsten Aspekte aus Teil (a). c) Beschreiben Sie bitte, welches Wissen und welche Fähigkeiten für die Lösung der von Ihnen formulierten Aufgaben bzw. Aufgabenideen benötigt werden.

(3) Die Unterschiede von erfolgreichen und nicht erfolgreichen Erstsemesterstudierenden in Mathematikvorlesungen: Denken Sie bitte an typische erfolgreiche und nicht erfolgreiche Studierende in einer von Ihnen angebotenen mathematischen Erstsemestervorlesung. In welchen Merkmalen unterscheiden sich diese beiden Personengruppen Ihrer Erfahrung nach bereits in der Studieneingangsphase?

Zu allen drei Impulsen wurden die Einschätzungen der Hochschullehrenden zu erwarteten mathematischen Lernvoraussetzungen in offenen Textfeldern erfasst. Da eine Auswertung dieser offenen Antworten von einer großen Stichprobe technisch nur schwer umzusetzen gewesen wäre, wurde in diese erste explorative Befragungsrunde nur eine kleine Teilstichprobe einbezogen. Aus der Gesamtstichprobe wurden dafür Expertinnen und Experten ausgewählt, die einen besonderen Bezug zur Lehre haben (z. B. eine besondere Verantwortung als Studiendekanin oder Studiendekan oder eine langjährige Lehrerfahrung). Bei der Auswahl wurde berücksichtigt, dass in dieser Expertenteilgruppe alle Bundesländer, die verschiedenen MINT-Studienrichtungen und die unterschiedlichen Hochschularten abgebildet sind. 82 Hochschullehrende, die diese Kriterien erfüllten, wurden zur Teilnahme an der ersten Befragungsrunde eingeladen, von denen sich 36 beteiligten.

Die Antworten dieser Hochschullehrenden zu allen drei erzählgenerierenden Fragen wurden angelehnt an die qualitative Inhaltsanalyse (Mayring, 2003) ausgewertet. Dabei wurden Kategorien von notwendigen mathematischen Lernvoraussetzungen herausgearbeitet (die Interrater-Reliabilität 85–97% im paarweisen Vergleich, Cohen's κ .60–.94 war zufriedenstellend). Die genannten Lernvoraussetzungen wurden hierarchisiert und zu einem Katalog zusammengestellt. Oft verwiesen die Hochschullehrenden in ihren Antworten auf Bildungsdokumente wie beispielsweise die KFP-Empfehlungen, den cosh-Katalog oder die Bildungsstandards Mathematik. Daher wurden die direkt genannten Lernvoraussetzungen ergänzt um solche aus diesen Dokumenten. Abschließend wurde der gesamte Katalog erneut mit den Originaläußerungen der Hochschullehrenden abgeglichen, um die Vollständigkeit und Korrektheit sicherzustellen. Insgesamt lag damit als Ergebnis der ersten Befragungsrunde ein Katalog von 152 mathematischen Lernvoraussetzungen vor. Diese Lernvoraussetzungen gliederten sich in vier Kategorien: *Mathematische Inhalte, Mathematische Arbeitstätigkeiten, Wesen der Mathematik* und *persönliche Merkmale*.

Befragungsrunde 2

In der zweiten Runde wurden die in Runde 1 explorativ ermittelten und zu den Kategorien zusammengefassten Lernvoraussetzungen der Gesamtstichprobe vorgelegt, von denen sich 952 Hochschullehrende beteiligten (44 %). Für jede zuvor ermittelte Lernvoraussetzung wurden die Hochschullehrenden gebeten zu bewerten, inwieweit sie diese als notwendig für einen erfolgreichen Start in das Studium jenes Faches erachten, in dem sie die mathematische Lehrveranstaltung unterrichten. Damit zielte die zweite Runde darauf, zu validieren, welche Lernvoraussetzungen aus Sicht einer breiten Stichprobe von Hochschullehrenden seitens der Studienanfängerinnen und Studienanfänger aus der Schule mitgebracht werden sollten.

In den Kategorien *Mathematische Inhalte* und *Wesen der Mathematik* sollten die Hochschullehrenden darüber hinaus angeben, auf welchem Niveau sie die einzelnen Lernvoraussetzungen als notwendig erachteten. Niveau 1 bezog sich dabei auf ein eher grundlegendes, Niveau 2 auf ein eher komplexeres Verständnis. Auf eine solche Differenzierung wurde bei den *Mathematischen Arbeitstätigkeiten* zunächst verzichtet, da durch die Bewertung einzelner Teilprozesse bereits die Möglichkeit einer abgestuften Bewertung bestand. Lernvoraussetzungen im Bereich *persönlicher Merkmale* sollten auf einer 4-stufigen Likert-Skala differenziert bewertet werden (unwichtig, eher unwichtig, eher wichtig, wichtig). Nach diesen geschlossenen Items zur Bewertung der Lernvoraussetzungen hatten die Hochschullehrenden in allen Kategorien die Möglichkeit, in offenen Textfeldern die zuvor genannten Lernvoraussetzungen zu präzisieren, zu verändern oder durch weitere Aspekte zu ergänzen.

Bei der Auswertung der Daten in dieser Befragungsrunde stand insbesondere die Frage im Fokus, inwieweit sich ein Konsens unter den Hochschullehrenden feststellen lässt. Dazu wurden konservative Kriterien angelegt (vgl. Abschnitt 3.3), bei denen das Meinungsbild in der Gesamtstichprobe, gleichzeitig aber auch in Teilstichproben (d. h. Hochschullehrende verschiedener Hochschultypen und verschiedener Studiengangsgruppen) berücksichtigt ist. Für die Aufnahme von Ergänzungen, Änderungen und Präzisierungen wurden ebenfalls Kriterien festgelegt, sodass dadurch neue Ideen aufgenommen, jedoch eine Beeinflussung durch Einzelmeinungen möglichst ausgeschlossen wurde. Bei der Auswertung der Anmerkungen seitens der Hochschullehrenden zeigte sich dabei außerdem, dass eine Unterscheidung in Anforderungsniveaus auch für Lernvoraussetzungen der Kategorie *Mathematische Arbeitstätigkeiten* gewünscht wird. So lag am Ende der Befragungsrunde 2 ein Katalog von 179 Lernvoraussetzungen vor.

Befragungsrunde 3

Das sich in Befragungsrunde 2 abzeichnende Meinungsbild sollte in der Befragungsrunde 3 schließlich konsolidiert und so die Stabilität der Ergebnisse sichergestellt werden. Dazu wurden die Lernvoraussetzungen zusammen mit Angaben über die Zustimmungsraten der Gesamtstichprobe ($N = 664$; 30 % Rücklaufquote) zu einer erneuten Bewertung vorgelegt. Auch in dieser Runde hatten die Hochschullehrenden in

offenen Textfeldern die Möglichkeit, Präzisierungen, Änderungen und Ergänzungen zu machen.

Die in Runde 2 positiv bestätigten Lernvoraussetzungen wurden unter Angabe des jeweiligen Niveaus aufgelistet. Zur besseren Übersichtlichkeit wurden dazu entsprechend der Kategorien Blöcke thematisch zusammengehöriger Lernvoraussetzungen gebildet. Den Wünschen der Hochschullehrenden entsprechend wurden dabei für die Kategorie *Mathematische Arbeitstätigkeiten* zwei Anforderungsniveaus eingeführt, um eine differenziertere Bewertung der Notwendigkeit vornehmen zu können. Für jeden Block an Lernvoraussetzungen sollten die Hochschullehrenden auf einer 6-stufigen Likert-Skala (1 = „stimme gar nicht zu“ bis 6 = „stimme voll zu“) angeben, inwiefern sie zustimmen, dass die dargestellten Lernvoraussetzungen für den Studieneinstieg in MINT-Studiengänge notwendig sind.

Die Lernvoraussetzungen, bei denen sich in Befragungsrunde 2 kein Konsens ergeben hatte, wurden als einzelne Aspekte zu einer erneuten Bewertung vorgelegt, um eine individuelle Bewertung einholen zu können. Für jeden Aspekt mit uneinheitlichem Meinungsbild wurden dabei die Zustimmungsraten aus der Vorrunde angegeben. Gleichermaßen wurden die Lernvoraussetzungen, die nach Angaben der Hochschullehrenden in der Befragungsrunde 2 geändert worden waren, oder die ihrer Ansicht nach noch ergänzt werden sollten, in der Befragungsrunde 3 als einzelne Aspekte zur Bewertung vorgelegt. Dazu wurde wie schon in der Vorrunde um eine Einschätzung der Aspekte gebeten hinsichtlich der Notwendigkeit als Lernvoraussetzung für einen MINT-Studiengang gebeten. Bei Lernvoraussetzungen in den Kategorien *Mathematische Inhalte*, *Mathematische Arbeitstätigkeiten* und *Wesen der Mathematik* sollte im Falle einer Notwendigkeit auch das notwendige Anforderungsniveau angegeben werden, bei Lernvoraussetzungen zu *persönlichen Merkmalen* sollte die Notwendigkeit auf einer 4-stufigen Likert-Skala eingeschätzt werden.

Lernvoraussetzungen, die in der Befragungsrunde 2 zwar die Kriterien für notwendige Lernvoraussetzungen erfüllten, jedoch nicht eindeutig einem der beiden Niveaus zugeordnet werden konnten, wurden den Hochschullehrenden in der Befragungsrunde 3 ebenfalls zu einer erneuten individuellen Bewertung vorgelegt.

Zur Auswertung der Daten wurden erneut die konservativen Konsenskriterien angelegt (vgl. Abschnitt 3.3). Dabei zeichnete sich für einen Großteil der Lernvoraussetzungen ein Konsens ab. Für die Bestätigung der bereits als notwendig identifizierten Lernvoraussetzungen ergaben sich sehr hohe Zustimmungsraten (mittlere Zustimmung 92.4 %; minimale Zustimmung 87.1 %, wobei Zustimmung als eine Bewertung von 4 bis 6 auf der sechsstufigen Likert-Skala definiert wurde). Darüber hinaus zeigten sich keine weiteren substanziellen Änderungen oder Ergänzungen in den offenen Kommentaren der Hochschullehrenden. Insgesamt konnte damit von einem stabilen Ergebnis ausgegangen werden. Von der Durchführung einer weiteren Befragungsrunde wurde daher abgesehen, da gegeben diese Stabilität einerseits und die zu beobachtende Stichprobenmortalität andererseits keine weitere Erhöhung der Aussagekraft erwartet werden konnte.

3.3 Auswertungskriterien

Bei der Auswertung der geschlossenen Items in den Befragungsrunden 2 und 3 war eine zentrale Frage, welche mathematischen Lernvoraussetzungen die Hochschullehrenden als notwendig für einen Einstieg in ein MINT-Studium erachten und inwieweit dabei von einem Konsens ausgegangen werden kann. Zur Beurteilung eines Konsenses mussten Kriterien festgelegt werden. Diese Kriterien sollten konservativ angesetzt sein, um das Meinungsbild der Hochschullehrenden nicht zu verzerren und sicherzustellen, dass einerseits über alle Befragten hinweg ein Konsens anzunehmen ist und andererseits der Konsens auch in den einzelnen Studiengangsgruppen und Hochschularten gilt. Daher orientierten sich die Kriterien an in Deutschland üblichen Normen, die bei Fragen eines gesellschaftlichen Konsenses angelegt werden, wie die beispielsweise zur Änderung der Verfassung notwendige Zweidrittelmehrheit.

So wurde in MaLeMINT eine Lernvoraussetzung als notwendig angesehen, wenn (1) mindestens zwei Drittel aller Befragten und (2) mindestens die Hälfte der Lehrenden in jeder Studiengangsgruppe (Mathematik, MINT oder INT) und (3) mindestens die Hälfte der Lehrenden in jeder Hochschulart (Universität, (Fach-)Hochschule) die Lernvoraussetzung als notwendig ansahen. Da eine einheitliche Einschätzung einer Lernvoraussetzung als nicht notwendig eine deutlich schwerwiegendere Konsequenz zur Folge hätte, wurden dafür härtere Kriterien angelegt. So wurde eine Lernvoraussetzung als nicht notwendig gesehen, wenn (1) mindestens drei Viertel aller Befragten und (2) mindestens zwei Drittel der Lehrenden in jeder Studiengangsgruppe (Mathematik, MINT oder INT) und (3) mindestens zwei Drittel der Lehrenden in jeder Hochschulart (Universität, (Fach-)Hochschule) die Lernvoraussetzung als nicht notwendig ansahen.

Bei der Auswertung der Äußerungen in den offenen Textfeldern stand vor allem die Frage im Vordergrund, inwieweit die Anmerkungen in der Folgerunde berücksichtigt werden sollten. Um eine Beeinflussung durch Einzelmeinungen zu vermeiden und dennoch neue Aspekte angemessen zu berücksichtigen, wurde als Auswahlkriterium eine Nennung von mindestens drei Hochschullehrenden festgelegt. Dazu wurden die Anmerkungen zu Präzisierungen, Verbesserungen und Ergänzungen der Lernvoraussetzungen zunächst strukturiert und zusammengefasst. Aspekte, die dann das Kriterium der drei Nennungen erfüllten, wurden in der Folgerunde für eine Bewertung aufgenommen.

3.4 Ergebnisse: der MaLeMINT-Katalog

Ein zentrales Ergebnis der MaLeMINT-Studie ist der breite Konsens, der sich über die drei Befragungsrunden hinweg abzeichnete. So erfüllten 144 von 179 (80 %) der Lernvoraussetzungen die gesetzten Konsenskriterien (s. Details in Tabelle 1). 140 davon wurden von den Hochschullehrenden als notwendig angesehen, vier als nicht notwendig (diese bezogen sich vor allem auf abstrakte und formale Aspekte wie beispielsweise die axiomatische Definition eines Vektorraumes). 35 Lernvoraussetzungen erfüllten nicht die Konsenskriterien. Dabei kam der fehlende Konsens nicht durch

systematische Unterschiede zwischen den einzelnen Teilstichproben zustande, sondern war eher innerhalb der Teilstichproben festzustellen. Insgesamt deuten diese Ergebnisse darauf hin, dass die Hochschullehrenden für die MINT-Fächer vergleichbare Erwartungen hinsichtlich der mathematischen Lernvoraussetzungen ihrer Studienanfängerinnen und Studienanfänger haben und dies über die verschiedenen Studienfächer, Hochschularten und Bundesländer hinweg.

Im Folgenden werden einige Ergebnisse strukturiert nach den vier Kategorien von Lernvoraussetzungen hervorgehoben.

Tabelle 1: Übersicht über die ermittelten Lernvoraussetzungen

Kategorie	Notwendig	Nicht notwendig	Kein Konsens	Gesamt
A) Mathematische Inhalte				
A1) Grundlagen	46	0	4	50
A2) Analysis	20	0	10	30
A3) Lineare Algebra und Analytische Geometrie	7	3	6	16
A4) Stochastik und bereichsübergreifende Inhalte	4	1	5	10
B) Mathematische Arbeitstätigkeiten				
B1) Grundlagen (Rechnen, Hilfsmitteleinsatz, Darstellungen)	9	0	0	9
B2) Mathematisches Argumentieren und Beweisen	8	0	1	9
B3) Mathematisches Kommunizieren	5	0	0	5
B4) Mathematisches Definieren	3	0	1	4
B5) Problemlösen	7	0	1	8
B6) Mathematisches Modellieren	4	0	2	6
B7) Recherche	1	0	0	1
C) Wesen der Mathematik	7	0	2	9
D) Persönliche Merkmale				
D1) Einstellungen und Arbeitsweisen	11	0	0	11
D2) Kognitive Fähigkeiten und Kenntnisse	5	0	2	7
D3) Soziale Fähigkeiten	3	0	1	4
Gesamt	140	4	35	179

Lernvoraussetzungen im Bereich *Mathematische Inhalte*

Aspekte mathematischer Inhalte machten den Großteil der erwarteten Lernvoraussetzungen aus. Dabei adressierten die Erwartungen der Hochschullehrenden überwiegend Inhalte, die üblicherweise im Mathematikunterricht bis zum Ende der Sekun-

darstufe I behandelt werden (Unterkategorie Grundlagen, vgl. Tab. 1). Hinzu kamen noch zentrale Inhalte der Sekundarstufe II.

Im Bereich der Grundlagen erwarteten die Hochschullehrenden unter anderem ein Verständnis der Bruchrechnung, Proportionalität, Prozentrechnung, linearen und quadratischen Gleichungen und Funktionen, linearen Gleichungssystemen, Betrags-, Exponential- und Logarithmusgleichungen oder Berechnung von Flächen und Volumina bekannter Figuren und Körper.

Für das Gebiet der Analysis erstrecken sich die erwarteten Lernvoraussetzungen von Folgen und Reihen über Stetigkeit bis hin zu Differenzial- und Integralrechnung. So werden beispielsweise Kenntnisse der Differenziations- und Integrationsregeln und Vorstellungen von Ableitung und Integral (z. B. als lokale lineare Approximation respektive als orientierter Flächeninhalt) erwartet. Interessant ist dabei, dass lediglich intuitive Vorstellungen zu zentralen Begriffen der Analysis wie Stetigkeit und Differenzierbarkeit erwartet werden (z. B. Stetigkeit als „durchgezogener Graph" bzw. Differenzierbarkeit als „kein Knick im Graph"), während formale Vorstellungen (z. B. die ε-δ-Definition der Stetigkeit oder die Differenzierbarkeit und Integrierbarkeit mit formalem Grenzwertkonzept auf Basis von Folgen) zu einem uneinheitlichen Meinungsbild führten.

Im Bereich der linearen Algebra und analytischen Geometrie zeigte sich ein ähnliches Bild. Die Hochschullehrenden erwarten hier beispielsweise eine Vorstellung von Vektoren als Pfeilklassen, ein Verständnis von Kollinearität, sowie elementare Operationen von Vektoren inklusive Skalarprodukt oder die Fähigkeit, Objekte wie Punkte, Gerade und Ebene in Ebene und Raum analytisch zu beschreiben. Inhalte, die ein eher formales, abstraktes Verständnis erfordern, führten wie auch im Bereich der Analysis zu einem uneinheitlichen Meinungsbild, wie beispielsweise Linearkombinationen und lineare Abhängigkeit von Vektoren über Kollinearität hinaus oder Matrizenrechnung.

Schließlich wurden mathematische Inhalte zur Stochastik sowie bereichsübergreifende Inhalte genannt. Im Bereich der Stochastik werden nur wenige Lernvoraussetzungen erwartet, wie beispielsweise die abzählende Kombinatorik (Permutationen, Variationen, Kombinationen) und ein Verständnis von Wahrscheinlichkeit sowie Binomial- und Normalverteilung. Die bereichsübergreifenden Inhalte umfassen zum Beispiel die Aussagenlogik inklusive Verknüpfung und Umkehrung von sowie Rechnen mit Aussagen, oder ein Verständnis übergeordneter Begriffe wie *Definition*, *Satz* oder *Beweis*.

Lernvoraussetzungen im Bereich *Mathematische Arbeitstätigkeiten*

Neben Kenntnissen mathematischer Inhalte erwarten die Hochschullehrenden auch Fähigkeiten, für die Mathematik typische Arbeitstätigkeiten anwenden zu können. Ein Viertel der Lernvoraussetzungen ist den grundlegenden Arbeitstätigkeiten zuzuordnen. Diese umfassen beispielsweise das schnelle und korrekte Ausführen von bekannten Verfahren ohne elektronische Hilfsmittel, den sicheren Umgang mit sowie Wechsel zwischen Standarddarstellungen wie auch den sicheren Umgang mit Ta-

schenrechnern und Computern inklusive kritischer Reflexion der Ergebnisse. Darüber hinausgehend werden folgende Arbeitstätigkeiten als Lernvoraussetzungen erwartet:

- mathematisches Argumentieren und Beweisen (z. B. Verstehen und Explorieren von mathematischen Behauptungen und Sätzen, das Verstehen und Prüfen mathematischer Beweise),
- mathematisches Kommunizieren (z. B. schriftliche mathematische Formulierungen mit Fachsprache und Fachsymbolik verstehen, Mathematik in präziser mathematischer Notation mit Fachsprache und Fachsymbolik darzustellen),
- mathematisches Definieren (z. B. mathematische Begriffe anhand ihrer Definition erklären, mathematische Definitionen bekannter Begriffe angemessen zu formulieren),
- mathematisches Problemlösen (z. B. allgemeine heuristische Prinzipien sicher und flüssig verwenden, Probleme mit mindestens drei Lösungsschritten zu lösen),
- mathematisches Modellieren (z. B. Kontrolle von Ergebnissen einer mathematischen Modellierung mit Blick auf die Realsituation),
- Recherchierens (z. B. Informationen in Nachschlagewerken, dem Internet oder anderen Quellen ermitteln).

Auch im Bereich der mathematischen Arbeitstätigkeiten wird deutlich, dass anspruchsvollere Arbeitstätigkeiten, die eher den abstrakten und formalen Charakter der Mathematik widerspiegeln, ein uneinheitliches Meinungsbild hervorrufen (z. B. das Entwickeln und Formulieren mathematischer Beweise zu einer gegebenen Behauptung oder das Entwickeln eigener Definitionen zu mathematischen Begriffen).

Lernvoraussetzungen im Bereich *Wesen der Mathematik*

Neben Wissen und Fähigkeiten zu Inhalten und Arbeitstätigkeiten erwarten die Hochschullehrenden auch ein Meta-Wissen über Mathematik als wissenschaftliche Disziplin und Anwendungsdisziplin. Zum Beispiel sollten Studienanfängerinnen und Studienanfänger im MINT-Bereich ein Verständnis davon mitbringen, dass sich die Mathematik durch die spezielle Art des Beweisens von vielen anderen Disziplinen abgrenzt oder dass die für das Beweisen notwendige Präzision eine Strenge in der Begriffsdefinition und Argumentation erfordert. Auch sollte Mathematik verstanden werden als Schulung des präzisen und abstrakten Denkens, das über das schablonenartige Anwenden von Routinen hinausgeht, und als offenes System, das weit mehr und qualitativ Anderes enthält als die Schulmathematik. Die Hochschullehrenden halten es für ausreichend, dass die Studienanfängerinnen und Studienanfänger diese Vorstellungen als abstraktes Meta-Wissen vorliegen haben (es also in irgendeiner Form mitgeteilt bekommen haben). Es wird nicht erwartet, dass bereits eigene Erfahrungen mit den jeweiligen Wesenszügen der Mathematik gemacht wurden (also dass beispielsweise die Studienanfängerinnen und Studienanfänger bereits eigene Beweise geführt und über diese Vorgehensweise reflektiert haben).

Lernvoraussetzungen im Bereich *Persönliche Merkmale*
Schließlich erwarten die Hochschullehrenden auch persönliche Merkmale seitens der Studienanfängerinnen und Studienanfänger, die für das Mathematiklernen relevant sind. So sollten sie Einstellungen und Arbeitsweisen mitbringen, wie beispielsweise eine Offenheit gegenüber Mathematik als wissenschaftliche Disziplin und dem Mathematiklernen an der Hochschule, die Bereitschaft zum Herleiten neuer Zusammenhänge, die Bereitschaft und Fähigkeit zum selbstständigen Arbeiten, oder Durchhaltevermögen, Ausdauer und Frustrationstoleranz gegenüber mathematikbezogenen Aufgaben. Erwartet werden auch kognitive Fähigkeiten und Kenntnisse, wie beispielsweise ein schnelles Auffassungsvermögen, die Bereitschaft und Fähigkeit zu konzentriertem Arbeiten über einen längeren Zeitraum, oder Kreativität und Vorstellungsvermögen (z. B. für die Übertragung und Weiterentwicklung von Methoden oder die Entwicklung von Problemlöseideen). Außerdem sollten sie soziale Fähigkeiten mitbringen, wie beispielsweise die Bereitschaft zum Austausch über Mathematik mit Lehrenden und anderen Lernenden oder auch Teamfähigkeit, um beispielsweise in kleineren Übungsgruppen gemeinsam mathematische Probleme zu bearbeiten.

4 Die Schnittstelle: Anschlussfähigkeit von Erwartungen an normative Vorgaben

Mit den Ergebnissen der MaLeMINT-Studie liegt ein detaillierter Katalog an mathematischen Lernvoraussetzungen vor, die Hochschullehrende von Studienanfängerinnen und Studienanfängern in MINT-Studiengängen erwarten. Durch den Konsens unter den Hochschullehrenden werden die mathematischen Anforderungen für den Einstieg in ein MINT-Studium in Deutschland nicht nur transparent. Der vorliegende Katalog erlaubt auch, mögliche Passungsprobleme (vgl. Wolter, 2013) zwischen den Erwartungen der Hochschullehrenden einerseits und den normativen Zielvorgaben schulischen Mathematikunterrichts andererseits im Detail zu untersuchen. Insbesondere kann damit analysiert werden, ob die von Hochschullehrenden beklagten mangelnden Mathematikkenntnisse der Studienanfängerinnen und Studienanfänger auf unzureichende Ziele des Mathematikunterrichts zurückzuführen sind, die durch die Bildungsstandards für die gesamte Schulzeit bis zum Abitur festgelegt werden (z. B. Offener Brief, 2017).

Ein naheliegender Zugang zur Untersuchung dieser Frage wäre ein Abgleich zwischen dem MaLeMINT-Katalog und den Bildungsstandards Mathematik für den Mittleren Schulabschluss und für die Allgemeine Hochschulreife (KMK, 2004, 2015), da die Bildungsstandards einen bundesweiten Rahmen für die Ziele des Mathematikunterrichts bis zum Abitur vorgeben. Hier zeigt allerdings ein Blick in die Dokumente sehr schnell, dass die Bildungsstandards einen deutlich geringeren Auflösungsgrad aufweisen als die Erwartungen der Hochschullehrenden im MaLeMINT-Katalog. Somit wäre bei der Analyse sehr viel Interpretation notwendig und es besteht die Gefahr, dass man zu keinen validen Aussagen kommt. Daher braucht es Bildungsdokumente mit einem höheren Auflösungsgrad, wie beispielsweise die Lehrpläne der Bundeslän-

der, die auf Basis der Bildungsstandards ausgearbeitet wurden. Im Folgenden werden erste Ergebnisse einer laufenden Studie vorgestellt, in der die von Hochschullehrenden erwarteten mathematischen Lernvoraussetzungen mit den Lehrplänen mehrerer Bundesländer abgeglichen werden. Begonnen wurde mit dem eigenen Bundesland Schleswig-Holstein, dessen Fachanforderungen Mathematik für allgemeinbildende Schulen bereits ausgewertet wurden und im Folgenden berichtet werden können.

4.1 Die Fachanforderungen Mathematik in Schleswig-Holstein

Die Fachanforderungen Mathematik für die Sekundarstufen I und II der allgemeinbildenden Schulen in Schleswig-Holstein wurden 2014 vom Ministerium für Schule und Berufsbildung eingeführt (IQSH, 2014). Sie basieren auf den Bildungsstandards Mathematik der Kultusministerkonferenz und sind in einen für alle Fächer geltenden sogenannten allgemeinen Teil und einen für das Fach Mathematik geltenden fachlichen Teil gegliedert. „Die Fachanforderungen beschreiben [...] den spezifischen Beitrag der Fächer zur allgemeinen und fachlichen Bildung. Darauf aufbauend legen sie fest, was Schülerinnen und Schüler jeweils am Ende der Sekundarstufe I beziehungsweise am Ende der Sekundarstufe II wissen und können sollen“ (ebd., S. 6). Wie in anderen Bundesländern hat der Wechsel vom klassischen Lehrplan zu Lehrplanbeschreibungen, die an den Bildungsstandards orientiert sind, zwar substanzielle Änderungen in der Darstellung der curricularen Inhalte und Ziele mit sich gebracht (z. B. keine feste Zuordnung von Inhalten zu Jahrgangsstufen), hinsichtlich der curricularen Inhalte wird die bundeslandspezifische Tradition aber zumeist fortgeführt und inhaltliche Neuerungen oder Umstellungen im Curriculum werden eher vermieden. Für den Abgleich mit den Erwartungen der Hochschullehrenden gemäß MaLeMINT-Katalog ergibt sich aufgrund der Struktur der Fachanforderungen, dass die Lernvoraussetzungen der Kategorien *Mathematische Inhalte, Mathematische Arbeitstätigkeiten* und *Wesen der Mathematik* vornehmlich im fachlichen Teil der Fachanforderungen zu finden sind, während die Lernvoraussetzungen der Kategorie *Persönliche Merkmale* vor allem im allgemeinen (fächerübergreifenden) Teil erwähnt werden.

4.2 Methodisches Vorgehen

Für die Bewertung, inwieweit die Fachanforderungen Mathematik der Sekundarstufen I und II für Schleswig-Holstein die von den Hochschullehrenden erwarteten mathematischen Lernvoraussetzungen für MINT-Studiengänge abdecken, wurden die 140 Lernvoraussetzungen zugrunde gelegt, die bei der MaLeMINT-Studie einen Konsens als notwendige Lernvoraussetzungen erreicht hatten (vgl. Abschnitt 3.4). Für die Durchführung des Abgleichs waren zwei Prozessschritte notwendig. Zuerst wurde für jede der 140 MaLeMINT-Lernvoraussetzungen der Text der Fachanforderungen Mathematik nach Stellen durchsucht, die die jeweilige Lernvoraussetzung inhaltlich abbilden könnte. Im zweiten Prozessschritt wurden die Textstellen anschließend danach bewertet, in welchem Umfang die jeweilige Lernvoraussetzung abgedeckt wurde. Beide Schritte wurden von Kontrollmaßnahmen begleitet, um mögliche Zusammenhänge nicht zu übersehen. Im Folgenden wird das Vorgehen im Detail beschrieben.

Um die MaLeMINT-Lernvoraussetzungen im Dokument der Fachanforderungen zu identifizieren, wurde der Text der Fachanforderungen im ersten Prozessschritt genau gelesen und Textstellen, die einer oder mehreren Lernvoraussetzungen zugeordnet werden können, ermittelt. Dabei ging es jeweils um eine inhaltliche und nicht um eine wortwörtliche Übereinstimmung. So wird beispielsweise die MaLeMINT-Lernvoraussetzung der Analysis „Anschauliche/grafische Beziehung zwischen Funktions- und Ableitungsgraph" in den Fachanforderungen durch die Beschreibung „[Die Schülerinnen und Schüler] entwickeln Ableitungsgraphen aus dem Funktionsgraphen und umgekehrt" (IQSH, 2014, S 59) abgedeckt. Um sicherzustellen, dass in den Fachanforderungen alle dort abgedeckten MaLeMINT-Lernvoraussetzungen auch identifiziert worden sind, wurde anschließend mit einer Suchfunktion die pdf-Datei mit den Fachanforderungen nach zentralen Begriffen der noch nicht zugeordneten Lernvoraussetzungen durchsucht. Als letzter Schritt des Suchprozesses wurden schließlich die wenigen Lernvoraussetzungen, die keiner Textstelle in den Fachanforderungen zugeordnet werden konnten, einem Experten des IQSH vorgelegt, der an der Erstellung der Fachanforderungen beteiligt war. Darüber sollte abgesichert werden, dass die Lernvoraussetzungen, für die kein Hinweis in den Fachanforderungen gefunden werden konnte, auch tatsächlich nicht durch die Fachanforderungen abgebildet werden.

Für den zweiten Prozessschritt der Bewertung wurden alle MaLeMINT-Lernvoraussetzungen nach dem Grad ihrer Abdeckung in den Fachanforderungen kategorisiert. Dabei wurden die vier Stufen „gar nicht abgedeckt", „teilweise abgedeckt" (d. h. nur Teile der Lernvoraussetzung oder ein geringeres Niveau war abgedeckt), „optional abgedeckt" (d. h. die Lernvoraussetzung ist als Wahloption oder nur für Kurse mit erhöhtem Niveau vorgesehen) oder „vollständig abgedeckt" unterschieden. Zur Absicherung dieser Kategorisierung wurden alle Bewertungsschritte von einer zweiten Person durchgeführt. Es ergab sich eine durchschnittliche Übereinstimmung von 92 % (Cohen's Kappa .85), was einer guten Übereinstimmung entspricht. Insgesamt konnte mit dem Vorgehen eine reliable und valide Bewertung der Abdeckung der MaLeMINT-Lernvoraussetzungen in den Fachanforderungen Schleswig-Holstein erreicht werden.

4.3 Ergebnisse

In diesem Abschnitt wird zunächst der Gesamtüberblick auf der Basis von allen 140 Lernvoraussetzungen berichtet und anschließend im Detail die Teilergebnisse für die Lernvoraussetzungen der vier Kategorien aus der MaLeMINT-Studie betrachtet.

Abbildung 1 zeigt, in welchem Umfang die 140 MaLeMINT-Lernvoraussetzungen in den Fachanforderungen Mathematik in Schleswig-Holstein abgedeckt werden. Knapp zwei Drittel dieser Lernvoraussetzungen (91 von 140) sind auch als Ziele des Mathematikunterrichts in Schleswig-Holstein vorgesehen. Während eine Lernvoraussetzung als optional abgedeckt identifiziert wurde, sind 33 teilweise abgedeckt und 15 kommen in den Fachanforderungen nicht vor. Wie im Folgenden zu sehen sein wird, unterscheiden sich die Abdeckungen der Lernvoraussetzungen aus den vier Kategorien der MaLeMINT-Studie (Tab. 1) deutlich voneinander.

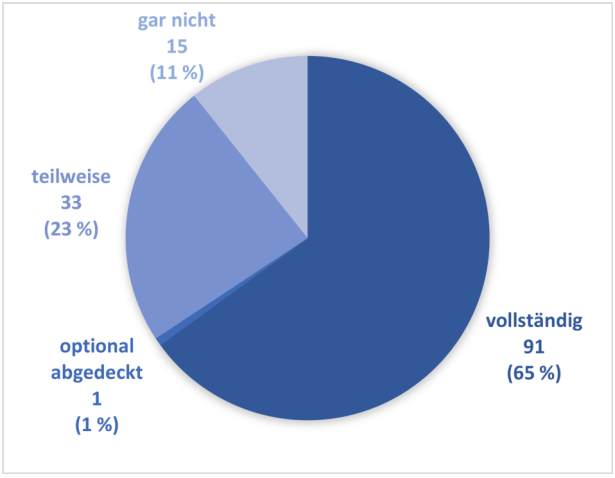


Abbildung 1: Abdeckung der 140 als notwendig erwarteten MaLeMINT-Lernvoraussetzungen in den Fachanforderungen Mathematik Schleswig-Holstein (Anzahlen und prozentuale Anteile)

Eine Interpretation der dargestellten Ergebniswerte erfolgt in der anschließenden Diskussion (Abschnitt 5). Es soll an dieser Stelle aber bereits angemerkt werden, dass eine nicht vorhandene Berücksichtigung einer von Hochschulseite erwarteten Lernvoraussetzung in den Fachanforderungen Mathematik zunächst auf eine Lücke in der Passung zwischen Schule und Hochschule hinweist. Die Ursache für dieses Passungsproblem kann auf zwei unterschiedliche Arten gedeutet werden: Entweder gibt es einen Mangel bei den normativ gesetzten Zielen des Mathematikunterrichts der Schule oder die von den Hochschulen gesetzten mathematischen Anforderungen zu Beginn der MINT-Studiengänge sind zu hoch. Je nach Ursachenzuschreibung kommen dann unterschiedliche Maßnahmen in Betracht, um das Passungsproblem zu reduzieren (vgl. Abschnitt 5).

Lernvoraussetzungen im Bereich *Mathematische Inhalte*

Die von Hochschulen erwarteten Lernvoraussetzungen im Bereich *Mathematische Inhalte* machen mit 55 % (77 von 140) den Großteil der MaLeMINT-Lernvoraussetzungen aus (vgl. Tab. 1). Im Sinne der Bildungsstandards Mathematik und auch der Fachanforderungen entsprechen sie den inhaltsbezogenen Kompetenzen zu den mathematischen Leitideen und damit insbesondere dem Wissen über mathematische Begriffe und Aussagen. Die Fachanforderungen Mathematik beschreiben diese Aspekte vor allem in umfangreichen Tabellen (Abschnitte II.2.2 und III.2.2 in IQSH, 2014).

In Abbildung 2 ist ersichtlich, dass 70 % (54 von 77) der MaLeMINT-Lernvoraussetzungen in den Fachanforderungen vollständig berücksichtigt werden, wohingegen 8 % (6 von 77) gar nicht abgedeckt sind. Bei den sechs nicht abgedeckten Lernvoraussetzungen handelt es sich um „Betragsgleichungen“, „Ungleichungen mit Beträgen“, „Polynomdivision“, „Funktionen mit Fallunterscheidungen“, „arithmetische und geometrische Folgen“ sowie „Vektoren als Pfeilklassen“. Die ersten vier dieser sechs Lernvoraussetzungen wurden von den Hochschullehrenden als „Grundlagen“ angesehen.

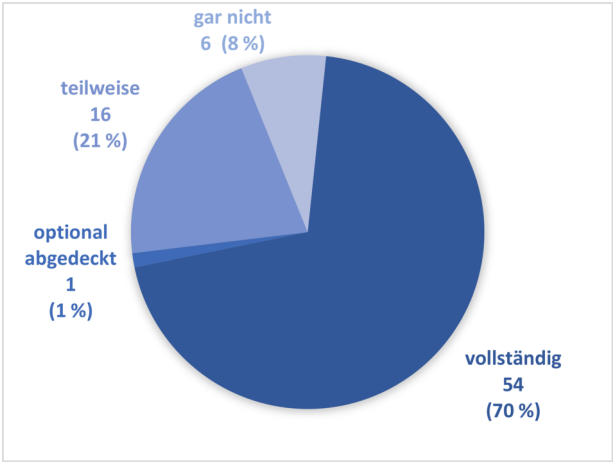


Abbildung 2: Abdeckung der als notwendig erwarteten MaLeMINT-Lernvoraussetzungen der Kategorie *Mathematische Inhalte* in den Fachanforderungen Mathematik Schleswig-Holstein (Anzahlen und prozentuale Anteile)

Knapp ein Fünftel der Lernvoraussetzungen (16 von 77) wurden nur teilweise abgedeckt (Abbildung 2). In mehreren Fällen handelt es sich hier um Lernvoraussetzungen, bei denen die Fachanforderungen sich auf die Berücksichtigung „einfacher" Beispiele beschränken. So findet sich etwa in der MaLeMINT-Studie die Lernvoraussetzung „Lineare und quadratische Ungleichungen", zu der in den Fachanforderungen als verbindlicher Inhalt nur „einfache Ungleichungen" im Kontext linearer und quadratischer Gleichungen erwähnt werden (IQSH, 2014, S. 24). Demgegenüber gibt es aber auch Lernvoraussetzungen, deren teilweise Abdeckung auf fehlende Aspekte hindeutet. So findet sich beispielsweise von der MaLeMINT-Lernvoraussetzung „Produktregel und Quotientenregel (Differenzialrechnung)" in den Fachanforderungen nur die Produktregel wieder, nicht aber die Quotientenregel. In etwas anderer Weise betrifft dies die Lernvoraussetzung „Kongruenz und Ähnlichkeit (und zugehörige Abbildungen)". Hier werden die Aspekte Kongruenz und Ähnlichkeit in den Fachanforderungen berücksichtigt, die zugehörigen Abbildungen sind aber nur optional vorgesehen. So dürfen die Schulen in Schleswig-Holstein selbst entscheiden, welchen Zugang sie zur Geometrie wählen: „Die Fachkonferenz entscheidet, ob entweder Kongruenzgeometrie oder Abbildungsgeometrie behandelt wird. Bei der Entscheidung für die Abbildungsgeometrie sind Achsenspiegelung, Drehung, Punktspiegelung, Translation an dieser Schule verbindliche Inhalte" (IQSH, 2014, S. 29). Diese Wahlfreiheit verursacht auch die einzige Lernvoraussetzung, die in Abbildung 2 als „optional abgedeckt" kategorisiert wurde. Hier handelt es sich um die Lernvoraussetzung „Strahlensätze", zu der in den Fachanforderungen angemerkt wird: „Die Fachkonferenz entscheidet, ob die Strahlensätze oder die zentrische Streckung behandelt werden." (IQSH, 2014, S. 25).

Zu erwähnen ist an dieser Stelle noch, dass es auch einige Lernvoraussetzungen aus der MaLeMINT-Studie gibt, die Hochschulen zu Beginn eines MINT-Studiums

nicht für notwendig halten bzw. für die es keinen Konsens gab, die aber in den Fachanforderungen teilweise oder vollständig berücksichtigt werden. Dazu gehören z. B. das Rotationsvolumen, einfache numerische Methoden, das Integral als aus der Änderungsrate rekonstruierter Bestand oder Beweisverfahren (direkter und indirekter Beweis, vollständige Induktion). An dieser Stellen gehen die Fachanforderungen über die Erwartungen der Hochschulen hinaus.

Lernvoraussetzungen im Bereich *Mathematische Arbeitstätigkeiten*

Neben dem Wissen über mathematische Inhalte erwarten die Hochschullehrenden auch, dass die Studienanfängerinnen und Studienanfänger dieses Wissen mittels typischer Arbeitstätigkeiten der Mathematik anwenden können. Im Bereich *Mathematische Arbeitstätigkeiten* findet sich etwa ein Viertel (36 von 140) der MaLeMINT-Lernvoraussetzungen (vgl. Tab. 1). Im Sinne der Bildungsstandards Mathematik entsprechen diese Arbeitstätigkeiten der Idee der allgemeinen mathematischen Kompetenzen, die in den Fachanforderungen Mathematik vor allem in den Abschnitten II.2.1 und III.2.1 (IQSH, 2014) erläutert werden. Erwähnenswert ist dabei, dass die Hochschullehrenden auch Arbeitstätigkeiten genannt haben, die in den Bildungsstandards und Fachanforderungen nicht explizit als Kompetenzen formuliert wurden (z. B. mathematisches Definieren oder Recherche, vgl. Tab. 1).

In Abbildung 3 ist ersichtlich, dass 28 von 36 (78 %) dieser MaLeMINT-Lernvoraussetzungen in den Fachanforderungen vollständig berücksichtigt werden, wohingegen 5 von 36 (14 %) teilweise und 3 von 36 (8 %) gar nicht abgedeckt sind.

Die drei nicht berücksichtigten Lernvoraussetzungen betreffen zwei von den Hochschullehrenden als „Grundlagen“ angesehene Aspekte der Darstellung („Sprachliche Fähigkeiten in Englisch zum Verstehen von Aufgabenstellungen oder Texten zur Mathematik“ und „Sicherer Umgang mit dem Summen- und dem Produktzeichen“) sowie einen Aspekt zum mathematischen Modellieren („Reflektieren des Nutzens und der Grenzen mathematischer Modellierungen für reale Problemsituationen“). Während zu den ersten beiden Aspekten keine expliziten Hinweise in den Fachanforderungen zu finden sind, wird für den Bereich des Modellierens die folgende prozessbezogene Kompetenzerwartung benannt: „[Die Schülerinnen und Schüler] reflektieren die Abhängigkeit einer Lösung von den getroffenen Annahmen des Modells“ (IQSH, 2014, S. 53). Diese Kompetenzerwartung für den Mathematikunterricht deutet zwar auch auf eine Reflexion von Modellen hin, stellt aber nicht sicher, dass tatsächlich Nutzen und Grenzen mathematischer Modellierungen für reale Problemsituationen im übergreifenden Sinne reflektiert werden, so wie es die Hochschullehrenden als Lernvoraussetzung erwarten. Selbstverständlich ist an dieser Stelle nicht ausgeschlossen, dass Lehrkräfte dies im Unterricht tun und über die Fachanforderungen hinausgehen (ebenso wie sie vielleicht auch das Summen- und Produktzeichen thematisieren). Die hier dargestellte Analyse hält sich aber eng an den Text der Fachanforderungen.

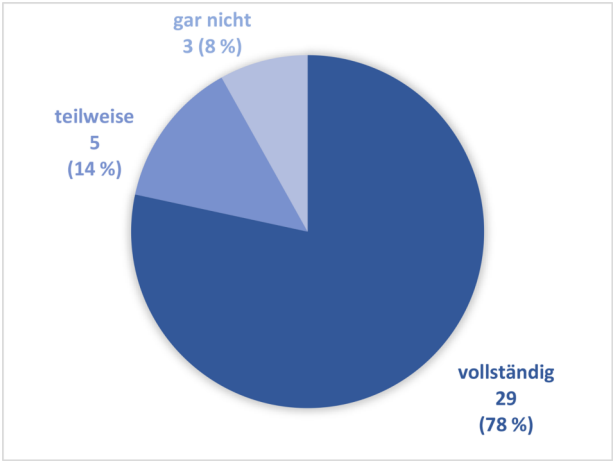


Abbildung 3: Abdeckung der als notwendig erwarteten MaLeMINT-Lernvoraussetzungen der Kategorie *Mathematische Arbeitstätigkeiten* in den Fachanforderungen Mathematik Schleswig-Holstein (Anzahlen und prozentuale Anteile)

Bei den fünf MaLeMINT-Lernvoraussetzungen, die nur teilweise abgedeckt sind, handelt es sich zumeist um Aspekte, die in den Fachanforderungen nicht konkret benannt werden, die aber in allgemeineren Beschreibungen angesprochen werden. Dies lässt sich exemplarisch an der Lernvoraussetzung „Mathematische Begriffe anhand ihrer Definition erklären" verdeutlichen, zu der sich in den Fachanforderungen folgende Aspekte finden: „Die Schülerinnen und Schüler erfassen, strukturieren und formalisieren Informationen aus zunehmend komplexen mathematikhaltigen Texten und Darstellungen, aus authentischen Texten, mathematischen Fachtexten, [...] erläutern mathematische Begriffe in theoretischen und in Sachzusammenhängen" (IQSH, 2014, S. 54). Demnach können Schülerinnen und Schüler einerseits Informationen aus mathematischen Fachtexten entnehmen und anderseits mathematische Begriffe in theoretischen Zusammenhängen erläutern. Damit kann die Fähigkeit, mathematische Begriffe anhand ihrer Definition zu erklären, zumindest als teilweise abgedeckt interpretiert werden, auch wenn mathematische Definitionen in ihrer Darstellung oft einen besonders kompakten Charakter haben.

Insgesamt kann festgestellt werden, dass die MaLeMINT-Lernvoraussetzungen zu den *Mathematischen Arbeitstätigkeiten* weitgehender abgedeckt sind als die Lernvoraussetzungen der Kategorie *Mathematische Inhalte* und nur wenige Aspekte in den Fachanforderungen nicht adressiert werden.

Lernvoraussetzungen im Bereich *Wesen der Mathematik*

Die Kategorie *Wesen der Mathematik* beschreibt Lernvoraussetzungen, die ein Meta-Wissen über Mathematik als wissenschaftliche Disziplin und Anwendungsdisziplin umfassen. Hier geht es konkret darum, dass Studienanfängerinnen und Studienanfänger die charakteristischen Merkmale der Mathematik in Abgrenzung zu anderen Disziplinen kennen. Laut Tabelle 1 umfasst diese Kategorie sieben Lernvoraussetzun-

gen, die Hochschullehrende für den Einstieg in ein MINT-Studium als notwendig ansehen. In den Fachanforderungen Mathematik finden sich Informationen zu diesen Lernvoraussetzungen vor allem in den Abschnitten II.1.2 und III.1.2 (IQSH, 2014), in denen der Beitrag des Faches Mathematik zur allgemeinen und fachlichen Bildung erläutert wird.

In Abbildung 3 ist ersichtlich, dass bis zum Abitur 3 von 7 der MaLeMINT-Lernvoraussetzungen in den Fachanforderungen teilweise berücksichtigt werden, wohingegen 4 von 7 gar nicht vorkommen. Dies heißt insbesondere, dass keine der sieben MaLeMINT-Lernvoraussetzungen vollständig durch die Fachanforderungen abgedeckt wird.

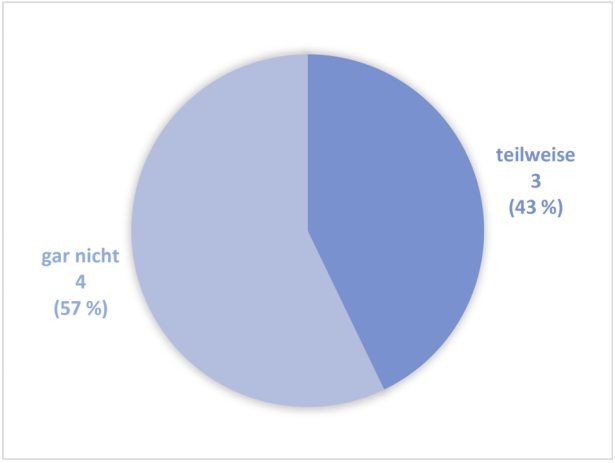


Abbildung 4: Abdeckung der als notwendig erwarteten MaLeMINT-Lernvoraussetzungen der Kategorie *Wesen der Mathematik* in den Fachanforderungen Mathematik Schleswig-Holstein (Anzahlen und prozentuale Anteile)

Wie schon bei Lernvoraussetzungen der Kategorie *Mathematische Arbeitstätigkeiten* zeigt sich bei den drei nicht vollständig berücksichtigten Lernvoraussetzungen, dass diese nicht explizit in den Fachanforderungen erwähnt werden. Stattdessen können sie durch allgemeine Beschreibungen oder durch eine Erwähnung bei konkreten Inhaltsbereichen nur als teilweise abgedeckt angesehen werden. So wird etwa in der Lernvoraussetzung „Begriffe werden in der Mathematik vollständig durch definierende Eigenschaften charakterisiert und auf Basis dieser Eigenschaften werden mithilfe deduktiver Schlussfolgerungen weitere Aussagen abgeleitet und bewiesen“ zwischen definierenden und abgeleiteten Eigenschaften bei Begriffen unterschieden und es soll bekannt sein, dass mathematische Begriffe über Eigenschaften definiert werden. In den Fachanforderungen finden sich dazu allgemeine Angaben wie „Kennzeichen mathematischer Arbeitsweise sind präziser Sprachgebrauch, Entwicklung klarer Begriffe, folgerichtige Gedankenführung und Argumentation“ (IQSH, 2014, S. 13), aus denen man die Bedeutung einer präzisen Begriffsdefinition ableiten kann. Dies wird ergänzt durch Angaben in der Leitidee Raum und Form, in der für bestimmte

Dreiecks- und Viereckssklassen die Unterscheidung von definierenden und abgeleiteten Eigenschaften vorgesehen ist, z. B.: „[Die Schülerinnen und Schüler] benennen, zeichnen und charakterisieren besondere Dreiecke und unterscheiden definierende und abgeleitete Eigenschaften." (IQSH, 2014, S. 29). Zusammengenommen kann die vorgesehene MaLeMINT-Lernvoraussetzung damit als teilweise abgedeckt angesehen werden.

Nicht durch die Fachanforderungen abgedeckt sind fünf Lernvoraussetzungen. Dazu gehören beispielsweise zentrale Vorstellungen zur Mathematik als wissenschaftliche Disziplin wie „Die spezielle Art des Beweisens grenzt die Mathematik von vielen anderen Disziplinen ab" oder „Mathematik sollte als ein offenes System angesehen werden, das viel mehr und qualitativ anderes enthält, als in der Schulmathematik thematisiert wird".

Lernvoraussetzungen im Bereich *Persönliche Merkmale*

Die von Hochschullehrenden erwarteten mathematikbezogenen persönlichen Merkmale von Studienanfängerinnen und Studienanfängern umfassen 19 Lernvoraussetzungen (Tab. 1). Darunter fallen sowohl kognitive Fähigkeiten und spezifische Kenntnisse, wie auch Einstellungen und Arbeitsweisen sowie soziale Fähigkeiten. Wie bereits in Abschnitt 4.1 erwähnt, werden diese Lernvoraussetzungen in den Fachanforderungen zumeist in den allgemeinen (fächerübergreifenden) Teilen adressiert. Dabei werden 8 der 19 vollständig abgedeckt, 9 teilweise abgedeckt und 2 werden nicht berücksichtigt (Abb. 5).

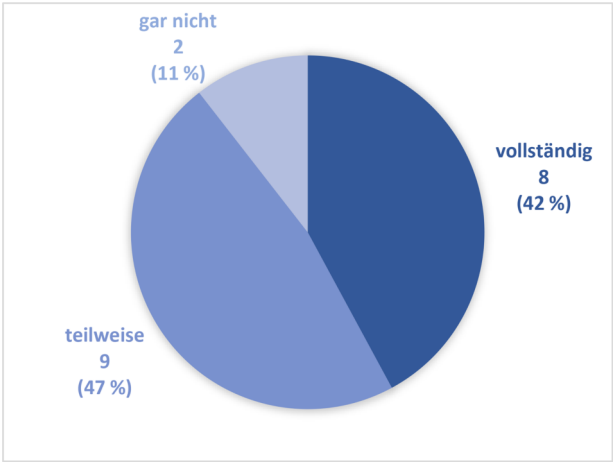


Abbildung 5: Abdeckung der als notwendig erwarteten MaLeMINT-Lernvoraussetzungen der Kategorie *Persönliche Merkmale* in den Fachanforderungen Mathematik Schleswig-Holstein (Anzahlen und prozentuale Anteile)

Die beiden Merkmale, die in den Fachanforderungen nicht explizit adressiert werden, sind „Konzentrationsfähigkeit (Bereitschaft und Fähigkeit zu konzentriertem Arbeiten über einen längeren Zeitraum)" sowie „Kenntnisse über den Aufbau und die Ziele

des zu wählenden Studiengangs“. Während bei der zweiten Lernvoraussetzung fraglich ist, ob sie zum Bildungsauftrag der Schule gehört (s. Abschnitt 5.2 Limitationen), kann die erste Lernvoraussetzung durchaus mit dem Mathematikunterricht in Verbindung gebracht werden. In den Fachanforderungen finden sich dazu nur sehr allgemeine Aussagen, die nicht spezifisch genug sind, z. B. „Guter Unterricht [...] fördert nicht allein die intellektuellen und kognitiven Kompetenzen der Schülerinnen und Schüler, sondern auch ihre sozialen und emotionalen, kreativen und körperlichen Potenziale“ (IQSH, 2014, S. 9).

Wie bei den beiden vorherigen Kategorien zeigt sich auch bei den *Persönlichen Merkmalen*, dass sich Aspekte der teilweise abgedeckten Lernvoraussetzungen in allgemeinen Beschreibungen wiederfinden. So finden sich Aspekte der Lernvoraussetzungen „Offenheit gegenüber der Mathematik als wissenschaftliche Disziplin und dem Mathematiklernen an der Hochschule“ beispielsweise in den folgenden beiden Zielen wieder: „[Die Schülerinnen und Schüler] lernen verschiedene Formen mathematischer Betrachtungs- und Vorgehensweisen kennen, wodurch sich geistige Beweglichkeit und Offenheit für unterschiedliche Fragestellungen und Sichtweisen weiterentwickeln“ und „Mit dem Erwerb der Kompetenz, komplexe Sachverhalte zu analysieren und dabei fachliche Methoden der Mathematik in angemessener Weise anzuwenden sowie die Ergebnisse strukturiert darzustellen, werden Schülerinnen und Schüler auf das wissenschaftliche Arbeiten in der Hochschule vorbereitet“ (IQSH, 2014, S. 48).

5 Diskussion

In diesem Kapitel sollen zunächst die Ergebnisse diskutiert und insbesondere bewertet und einordnet werden. Anschließend werden die Limitationen der Studie diskutiert und am Ende ein Ausblick auf weitere Studien im Kontext der Passungsprobleme beim Übergang Schule-Hochschule im Bereich Mathematik gegeben.

5.1 Diskussion der Ergebnisse

Fasst man die in Abschnitt 4.3 dargestellten Ergebnisse zusammen, so deutet sich auf den ersten Blick an, dass es in Schleswig-Holstein zwischen den Zielen des Mathematikunterrichts bis zum Abitur und den Erwartungen der Hochschullehrenden für den Einstieg in ein MINT-Studium eine Lücke gibt. Nach Abbildung 1 sind nur 91 (65 %) der 140 von Hochschulseite als notwendig angesehenen mathematischen Lernvoraussetzungen vollständig in den Fachanforderungen Mathematik abgedeckt und 49 (35 %) werden nicht vollständig oder sogar gar nicht im Sinne von Lernzielen explizit adressiert. Betrachtet man die Kategorien *Mathematische Inhalte* und *Mathematische Arbeitstätigkeiten*, so stellt sich der Anteil der vollständig abgedeckten Lernvoraussetzungen hier mit 73 % höher dar. Erwähnenswert ist die relativ hohe Abdeckung der Lernvoraussetzungen zu *Mathematische Arbeitstätigkeiten*, was zweifellos auf die starke Betonung der allgemeinen mathematischen Kompetenzen in den Fachanforderungen – und damit auch der Bildungsstandards – zurückzuführen ist.

Wenn man davon ausgeht, dass vor allem die gar nicht in der Schule berücksichtigten Lernvoraussetzungen ein gravierendes Problem darstellen, so bleibt insgesamt eine Lücke von 15 der 140 Lernvoraussetzungen (11 %). Betrachtet man diese geringe Anzahl gemeinsam mit den teilweise abgedeckten Lernvoraussetzungen, bei denen bereits eine Grundlage für das Weiterlernen vorhanden ist, so stellt sich das Passungsproblem beim Übergang von der Schule in ein MINT-Studium keinesfalls als unüberwindbar dar und sollte durch Anpassungen auf beiden Seiten sowie mithilfe von Unterstützungsmaßnahmen (z. B. Vor- oder Brückenkurse) grundsätzlich lösbar sein.

Da Herausforderungen für Studienanfängerinnen und Studienanfänger umso geringer ausfallen sollten, je kleiner die Lücke zwischen Zielen der Schule und Anforderungen der Hochschulen sind, stellt sich die Frage, inwieweit Schulen und Hochschulen durch Anpassung der von ihnen intendierten Ziele bzw. Anforderungen zur Verringerung dieser Lücke beitragen können. Der einfachste Weg aus Sicht der Hochschulen wäre die Erweiterung der Fachanforderungen Mathematik in dem Sinne, dass alle fehlenden Lernvoraussetzungen aufgenommen und während der Schulzeit bis zum Abitur adressiert werden. Dieser Wunsch wird allerdings dadurch begrenzt, dass die Schule nicht das alleinige Ziel hat, bei Schülerinnen und Schülern in einer beschränkten Zeit für die von Hochschulseite definierte Studierfähigkeit für MINT-Studiengänge zu sorgen. Zum einen bezieht sich die Studierfähigkeit auch auf andere Studiengänge, in denen beispielsweise höhere Anforderungen im Bereich der Stochastik und Statistik gestellt werden als in MINT-Studiengängen (z. B. Psychologie, vgl. Neumann et al., 2021), sodass auch diese Aspekte zu adressieren sind. Zum anderen hat die gymnasiale Oberstufe auch auf die berufliche Ausbildung vorzubereiten bzw. strebt eine vertiefte Allgemeinbildung als Bildungsziel an (KMK, 2021). Zusammen genommen bedeutet dies, dass im Mathematikunterricht auch der Erwerb von Wissen und Kompetenzen angestrebt werden muss, die nicht unbedingt für den Einstieg in ein MINT-Studium relevant sind und für die entsprechend zeitliche Ressourcen vorzuhalten sind. Dies zeigt sich ansatzweise auch schon, wenn alle 179 MaLe-MINT-Lernvoraussetzungen mit den Fachanforderungen Mathematik abgeglichen werden. So waren von den 39 Lernvoraussetzungen, die von den Hochschullehrenden als nicht notwendig angesehen wurden oder für die es keinen Konsens gab (vgl. Tab. 1), auch mehrere vollständig bzw. teilweise durch die Fachanforderungen abgedeckt.

Eine notwendige Konsequenz aus dem breiteren Auftrag der Schule wäre, dass nicht nur Schulen den Bereich der anzustrebenden Ziele erweitern, sondern auch Hochschulen ihre Anforderungen in einigen Bereichen verringern. Betrachtet man beispielsweise die in Abschnitt 4.3 erwähnten Lernvoraussetzungen der Kategorien *Mathematische Inhalte* und *Mathematische Arbeitstätigkeiten*, die nicht oder nur teilweise in den Fachanforderungen abgedeckt sind, so könnte man hier überlegen, dass Aspekte wie Gleichungen und Ungleichungen mit Beträgen, Funktionen mit Fallunterscheidungen, die Quotientenregel aber auch die Diskussion von Nutzen und Grenzen mathematischer Modelle beim Modellieren in die Fachanforderungen aufgenommen und damit im Mathematikunterricht adressiert werden sollten. Dagegen

könnten die arithmetischen und geometrischen Folgen, Vektoren als Pfeilklassen sowie der Umgang mit dem Summen- und Produktzeichen von den Hochschulen zu Beginn des Studiums behandelt werden. Für die Schulseite wäre zu fragen, ob durch Differenzierungsmaßnahmen eine passgenauere Vorbereitung auf ein MINT-Studium leistbar wäre. Auf Hochschulseite wäre zu diskutieren, inwieweit fehlende Lernvoraussetzungen durch Brückenkurse gezielt adressiert werden können.

Blickt man über die Kategorien *Mathematische Inhalte* und *Mathematische Arbeitstätigkeiten* hinaus, so stellt sich die Situation weniger klar dar. Bei den neun Lernvoraussetzungen zu Vorstellungen zum *Wesen der Mathematik* weisen die Fachanforderungen große Lücken auf. Bei diesen Lernvoraussetzungen handelt es sich um individuelle Vorstellungen von Schülerinnen und Schülern, die vermutlich nicht in kurzer Zeit aufgebaut werden können. Insbesondere zentrale Vorstellungen zur Mathematik als wissenschaftlicher Disziplin wie etwa „Die spezielle Art des Beweisens grenzt die Mathematik von vielen anderen Disziplinen ab“ oder „Mathematik sollte als ein offenes System angesehen werden, das viel mehr und qualitativ anderes enthält, als in der Schulmathematik thematisiert wird“ dürften für ein Verständnis der Lehrinhalte gleich zu Studienbeginn notwendig sein. Entsprechend wäre anzustreben, diese Lernvoraussetzungen in den Fachanforderungen explizit herauszuarbeiten und differenzierter darzustellen, als es bisher gemacht wird. Die derzeitige Darstellung weist eher einen Präambelcharakter auf und ist noch zu wenig angebunden an die konkreten Listen von curricularen Inhalten. Die explizite Aufnahme eines solchen Meta-Wissens über die Mathematik würde dabei nicht nur den Erwartungen von Hochschullehrenden entgegenkommen, sondern auch einem allgemeinen Bildungsziel der Oberstufe, der Wissenschaftspropädeutik (KMK, 2021), besser Rechnung tragen.

Ähnliches gilt für einige der nicht abgedeckten Lernvoraussetzungen der Kategorie *Persönliche Merkmale*. Allerdings werden in dieser Kategorie auch Lernvoraussetzungen von den Hochschullehrenden erwartet, die im Allgemeinen nicht zum Aufgabenbereich der Schule gehören und im Rahmen der Studienberatung von der Hochschulseite zu adressieren wären (z. B. „Kenntnisse über den Aufbau und die Ziele des zu wählenden Studiengangs“).

Der in diesem Beitrag dargestellte Abgleich zwischen den Erwartungen an mathematische Lernvoraussetzungen der Hochschulseite und den Fachanforderungen für den Mathematikunterricht basiert auf den intendierten Zielen der gymnasialen Oberstufe im Bereich der mathematischen Bildung. In diesem Sinne stellen die Fachanforderungen als Grundlage für den Abgleich in Abschnitt 4 ein Optimalniveau dar, das realistisch gesehen nur von einem Teil der Schülerinnen und Schüler erreicht werden dürfte. Betrachtet man die Schulleistungen im Fach Mathematik in Deutschland, so zeigte sich in den letzten 25 Jahren deutlich, dass die intendierten Ziele nur von wenigen Schülerinnen und Schülern der gymnasialen Oberstufe erreicht werden (im Überblick Rolfes et al., 2021). Für das Bundesland Schleswig-Holstein ergaben sich beispielsweise im Prüfungsfach Mathematik in den Abiturjahren 2019 und 2020 nur geringe Durchschnittsnoten von 6,9 bzw. 7,8 Punkten (MBWK, 2021). So

gesehen ist davon auszugehen, dass ein nicht unbeträchtlicher Teil der Studienanfängerinnen und Studienanfänger in MINT-Studiengängen geringere mathematische Lernvoraussetzungen mitbringt als in den Ergebnissen im Abschnitt 4.3 dargestellt. Hier zeigt sich entsprechend eine deutlich größere Herausforderung bei der Bewältigung des Passungsproblems zwischen mitgebrachten und erwarteten mathematischen Lernvoraussetzungen zu Beginn des Studiums. Fraglich ist an dieser Stelle, welche Wirksamkeit eine Steuerung über Lehrplanänderungen für dieses Problem haben kann.

5.2 Limitationen der Studie

Wie in den Abschnitten 4.2 und 4.3 erläutert und an einzelnen Lernvoraussetzungen illustriert, basiert der Abgleich zwischen den Lernvoraussetzungen des MaLeMINT-Katalogs und den Fachanforderungen Mathematik Schleswig-Holstein auf einer Analyse der in den Fachanforderungen explizit benannten Ziele des Mathematikunterrichts. Am Ende des vorherigen Abschnitts 5.1 wurde bereits angesprochen, dass diese Ziele intendierte Ziele darstellen und somit nicht mit den realisierten Zielen im Sinne von erreichten Schulleistungen verwechselt werden dürfen, die niedriger ausfallen können. Umgekehrt ist aber ebenfalls zu beachten, dass die Interpretation der Fachanforderungen durch Mathematiklehrkräfte und entsprechend auch das implementierte Curriculum über die Fachanforderungen hinausgehen können. So ist es Mathematikkollegien in Gymnasien nicht verboten, weiter gehende Lernangebote zu machen. Konkret kann dies bedeuten, dass ein Teil der Mathematiklehrkräfte sehr wohl Lernvoraussetzungen wie Gleichungen und Ungleichungen mit Beträgen, Funktionen mit Fallunterscheidungen oder die Quotientenregel im Rahmen eines differenzierenden Unterrichts behandeln und damit mehr Lernvoraussetzungen des MaLeMINT-Katalogs abdecken, als im Abschnitt 4.3 dargestellt wurden.

Neben den Fachanforderungen Mathematik kann auch die Validität des MaLeMINT-Katalogs als Basis für einen Abgleich von Hochschulanforderungen und Zielen der Schule hinterfragt werden. Wie in Abschnitt 3 dargestellt, basiert der Katalog auf einer Befragung von Hochschullehrenden. Inwieweit die Lernvoraussetzungen tatsächlich als Mindestanforderungen angesehen werden können, ist eine empirische Frage und bedarf einer empirischen Untersuchung (s. u. Abschnitt 5.3).

Als weitere Limitation ist zu nennen, dass die Fachanforderungen Mathematik für allgemeinbildende Schulen formuliert wurden, die den Bildungsweg über die gymnasiale Oberstufe bis zur allgemeinen Hochschulreife anbieten. Der MaLeMINT-Katalog umfasst dagegen Lernvoraussetzungen für MINT-Studiengänge an Universitäten und (Fach-)Hochschulen und betrifft damit insbesondere auch Studienanfängerinnen und Studienanfänger, die mit anderen Hochschulzugangsberechtigungen ein MINT-Studium antreten. So ist an dieser Stelle offen, welche Lernvoraussetzungen beispielsweise Studierende, die im Berufsschulsystem eine Fachhochschulreife erworben haben, an Fachhochschulen mitbringen. Es kann vermutet werden, dass die untersuchten Fachanforderungen Mathematik eher das Optimum an intendierten mathematischen Bildungszielen der Schulen in Schleswig-Holstein darstellen.

Schließlich ist zu erwähnen, dass der Abgleich nur auf den Lehrplanvorgaben aus einem Bundesland (Schleswig-Holstein) basiert. Es ist nicht unwahrscheinlich, dass ein Abgleich des MaLeMINT-Katalogs mit Lehrplänen aus anderen Bundesländern zu etwas anderen Ergebnissen führt. Vor diesem Hintergrund können die berichteten Ergebnisse nur auf das Land Schleswig-Holstein bezogen werden.

5.3 Ausblick

Mit dem vorgestellten Abgleich von erwarteten mathematischen Lernvoraussetzungen für MINT-Studiengänge und den Zielen des Mathematikunterrichts ist es exemplarisch für ein Bundesland gelungen, das Passungsproblem zwischen von Hochschulen erwarteten und von Schulen intendierten Lernvoraussetzungen für den Eintritt in MINT-Studiengänge zu analysieren. Als Ergebnis ließ sich das Passungsproblem auf eine Liste von konkreten Lernvoraussetzungen zurückführen, die von der Bildungsadministration und von Hochschulen in Schleswig-Holstein genutzt werden kann, um die Fachanforderungen oder die Unterstützungsmaßnahmen an Hochschulen zu ergänzen. Einen ersten Ansatz zur Lösung des Passungsproblems wurde im Rahmen eines Kooperationsprojekts vom Ministerium für Bildung, Wissenschaft und Kultur, dem Institut für Qualitätsentwicklung an Schulen Schleswig-Holstein (IQSH) und dem IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik im Anschluss an die MaLeMINT-Studie verfolgt. Im Rahmen des Projekts *MaLeMINT-Implementation* wurden von Mathematiklehrkräften aus Schulen und Mathematiklehrenden aus Hochschulen miteinander die im MaLeMINT-Katalog gelisteten mathematischen Inhalte und Arbeitstätigkeiten in Form von Aufgaben konkretisiert. Der dabei entstandene Aufgabenkatalog zeigt, dass sich Beteiligte aus Schulen und Hochschulen aufeinander einlassen und konstruktiv an der Übergangsproblematik Schule-Hochschule arbeiten können. Ein wichtiger Punkt in diesem Prozess war insbesondere das gemeinsame Aushandeln dessen, was seitens der Hochschulen konkret erwartet wird und seitens der Schulen – auf Basis gegebener normativer Rahmenbedingungen für allgemeinbildende Gymnasien – erwartet werden kann.

Bisher beschränkt sich der Lehrplanabgleich auf ein Bundesland. Wie in der Einleitung dieses Beitrags erwähnt, stellen die Befunde im Abschnitt 4 erste Ergebnisse einer laufenden Studie dar, in der noch vier weitere Bundesländer einbezogen worden sind. Die weiteren Analysen werden zeigen, welche Varianz es zwischen den insgesamt fünf Bundesländern gibt und welche Profile die jeweils identifizierten Passungsprobleme aufweisen. Die Erfahrungen aus MaLeMINT-Implementation, insbesondere die Gelingensbedingungen für einen erfolgreichen Austausch zwischen Schulen und Hochschulen, können für Akteure in den anderen Bundesländern interessante und hilfreiche Hinweise geben, ähnliche Arbeitsprozesse auch dort anzuregen.

Alle in diesem Beitrag berichteten Ergebnisse beziehen sich auf den Übergang von der Schule in ein MINT-Studium. Anzumerken ist, dass es noch viele weitere Studienfächer gibt, in denen mathematische Lernvoraussetzungen erwartet werden. In der Erweiterungsstudie MaLeMINT-E (Neumann et al., 2021) wurden diese Studienfächer ermittelt und ebenfalls Hochschullehrende befragt. Dabei zeigte sich, dass

in diesen Studienfächern in der Regel geringere Lernvoraussetzungen erwartet werden als in den MINT-Studiengängen. Ausnahmen waren einige Studienfächer, für die im Bereich Stochastik und Statistik im geringen Umfang zusätzliche Lernvoraussetzungen als notwendig angesehen werden. In der Grundtendenz kann aber davon ausgegangen werden, dass für die Studienfächer außerhalb des MINT-Bereichs keine größeren Passungsprobleme zwischen Hochschulerwartungen und Bildungszielen der Schule auftreten dürften als für die in diesem Beitrag adressierten MINT-Studienfächer.

Literatur

Albrecht, A. & Nordmeier, V. (2011). Ursachen des Studienabbruchs in Physik. Eine explorative Studie. *die hochschule* 2/2011, 131–145.

Behnke, H. (1965). Die Pflichten der Universität gegenüber dem Gymnasium. *Mathematisch-physikalische Semesterberichte*, XI, 1–19.

Bescherer, C. (2003). *Selbsteinschätzung mathematischer Studierfähigkeit von Studienanfängerinnen und -anfängern – Empirische Untersuchung und praktische Konsequenz*. Dissertation PH Ludwigsburg. https://phbl-opus.phlb.de/frontdoor/deliver/index/docId/4/file/bescherer.pdf

cosh – Cooperation Schule:Hochschule (2014). *Mindestanforderungskatalog Mathematik (Version 2.0) der Hochschulen Baden-Württembergs für ein Studium von WiMINT-Fächern (Wirtschaft, Mathematik, Informatik, Naturwissenschaft und Technik)*. http://www.mathematik-schule-hochschule.de/images/Aktuelles/pdf/MAKatalog_2_0.pdf

Deeken, C., Neumann, I. & Heinze, A. (2020). Mathematical prerequisites for STEM programs: What do university instructors expect from new STEM undergraduates? *International Journal of Research in Undergraduate Mathematics Education*, *6*, 23–41.

DMV, GDM & MNU (2019) = Deutsche Mathematiker-Vereinigung, Gesellschaft für Didaktik der Mathematik & Verband zur Förderung des MINT-Unterrichts (2019). *Mathematik: 19 Maßnahmen für einen konstruktiven Übergang Schule – Hochschule*. Stellungnahme S8 der Mathematik-Kommission Übergang Schule-Hochschule. http://www.mathematik-schule-hochschule.de/images/2019_02_18_PI-Handlungsempfehlungen.pdf

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, *51*(4), 327–358.

Häder, M. (2014). *Delphi-Befragungen: Ein Arbeitsbuch* (3. Aufl.). Wiesbaden: Springer.

Heldmann, W. (1984). *Studierfähigkeit: Ergebnisse einer Umfrage ; Thesen zur Studierfähigkeit und zum Hochschulzugang*. Schriften des Hochschulverbandes: Vol. 29. Göttingen: Schwartz.

Heublein, U., Ebert, J., Hutzsch, C., Isleib, S., König, R., Richter, J. & Woisch, A. (2017). *Zwischen Studienerwartungen und Studienwirklichkeit. Ursachen des Studienabbruchs, beruflicher Verbleib der Studienabbrecherinnen und Studienabbrecher und Entwicklung der Studienabbruchquote an deutschen Hochschulen*. Forum Hochschule 1|2017. http://www.dzhw.eu/pdf/pub_fh/fh-201701

Heublein, U. & Wolter, A. (2011). Studienabbruch in Deutschland: Definition, Häufigkeit, Ursachen, Maßnahmen. *Zeitschrift für Pädagogik, 57*(2), 214–236.

IQSH – Institut für Qualitätsentwicklung an Schulen Schleswig-Holstein (2014). *Fachanforderungen Mathematik. Allgemein bildende Schulen Sekundarstufe I Sekundarstufe II*. Kiel: Ministerium für Schule und Berufsbildung des Landes Schleswig-Holstein.

KFP – Konferenz der Fachbereiche Physik (2012). *Empfehlung der Konferenz der Fachbereiche Physik zum Umgang mit den Mathematikkenntnissen von Studienanfängern der Physik*. http://www.kfp-physik.de/dokument/KFP-Empfehlung-Mathematikkenntnisse.pdf

KMK – Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2004). *Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss Beschluss vom 4.12.2003*. München: Luchterhand.

KMK – Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2015). *Bildungsstandards im Fach Mathematik für die Allgemeine Hochschulreife Beschluss vom 18.10.2012.* Köln: Wolters Kluwer.

KMK – Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2021). *Vereinbarung zur Gestaltung der gymnasialen Oberstufe und der Abiturprüfung. Beschluss der Kultusministerkonferenz vom 07.07.1972 i. d. F. vom 18.02.2021.* https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/1972/1972_07_07-VB-gymnasiale-Oberstufe-Abiturpruefung.pdf

Konegen-Grenier, C. (2001). *Studierfähigkeit und Hochschulzugang.* Köln: Deutscher Instituts-Verlag.

Mayring, P. (2003). *Qualitative Inhaltsanalyse. Grundlagen und Techniken*. Weinheim: Beltz.

MBWK – Ministerium für Bildung, Wissenschaft und Kultur (2021). *Abitur-Ergebnisse*. https://za.schleswig-holstein.de/?view=100&path=1%20Abitur|Ergebnisse

Neumann, I., Pigge, C. & Heinze, A. (2017). *Welche mathematischen Lernvoraussetzungen erwarten Hochschullehrende für ein MINT-Studium? Eine Delphi-Studie*. Kiel: IPN.

Neumann, I., Rohenroth, D. & Heinze, A. (2021). *Studieren ohne Mathe? Welche mathematischen Lernvoraussetzungen erwarten Hochschullehrende für Studienfächer außerhalb des MINT-Bereichs? Eine Delphi-Studie*. Kiel: IPN.

Offener Brief (2017) = Mathematikunterricht und Kompetenzorientierung – ein offener Brief. https://www.tagesspiegel.de/downloads/19549926/2/offener-brief.pdf

Rolfes, T., Lindmeier, A. & Heinze, A. (2021). Mathematikleistungen von Schülerinnen und Schülern der gymnasialen Oberstufe in Deutschland: Ein Review und eine Sekundäranalyse empirischer Daten aus Schulleistungsstudien seit 1995. *Journal für Mathematik-Didaktik, 42(*2), 395–429.

SEFI (2013). *A Framework for Mathematics Curricula in Engineering Education. A Report of the Mathematics Working Group*. Brussels: European Society for Engineering Education (SEFI). https://www.sefi.be/publication/a-framework-for-mathematics-curricula-in-engineering-education/

Webler, T., Levine, D., Rakel, H., & Renn, O. (1991). The Group Delphi: A Novel Attempt at Reducing Uncertainty. *Technological Forecasting and Social Change, 3*, 253–263.

Wolter, A. (2013). Übergang aus dem Schulsystem heraus Übergänge zwischen Schule, beruflicher Bildung und Hochschule – Entwicklungen und Herausforderungen aus der Sicht der empirischen Bildungsforschung. In G. Bellenberg & M. Forell (Hrsg.), *Bildungsübergänge gestalten. Ein Dialog zwischen Wissenschaft und Praxis* (S. 45–61). Münster: Waxmann

Abbildungsverzeichnis

Tabellenverzeichnis

Schlusswort

Zur Vergleichbarkeit von Abiturnoten: Eine kritische Zwischenbilanz[1]

Eckhard Klieme

Seit mehr als 20 Jahren wird im deutschen Schulsystem daran gearbeitet, die Noten über einzelne Klassen, Schulen, Schularten (soweit sie die gleichen Bildungsgänge anbieten) und Bundesländer hinweg besser vergleichbar zu machen. Wichtige Instrumente hierfür sind Bildungsstandards, Vergleichsuntersuchungen, zentrale Prüfungen sowie Lehrkräftebildung zum Thema „Diagnostik". PISA-Daten weisen darauf hin, dass auf diesem Weg in der Sekundarstufe I Fortschritte gemacht wurden: 2012 war bundesweit der Zusammenhang zwischen Testleistungen und Schulnoten vor allem in Hauptschulen und Realschulen deutlich enger als im Jahr 2000 (Autorengruppe Bildungsberichterstattung, 2016, S. 94). Noten sind also insofern vergleichbarer geworden, als sie nun besser mit dem Vergleichsmaßstab der PISA-Skala übereinstimmen.

In der allgemeinbildenden Sekundarstufe II, d. h. in den Oberstufen von Gymnasien und Gesamtschulen, war die Ausgangslage vor 20 Jahren im Prinzip sogar besser, weil Gymnasialpädagogik und Abiturregeln – insbesondere die Einheitlichen Prüfungsanforderungen, EPA – schon lange teils implizite, teils explizite Normierungen vornahmen. Zusätzliche Anstrengungen wurden in den vergangenen Jahren mit der Angleichung der Kurssysteme und nicht zuletzt mit dem länderübergreifenden Abituraufgabenpool unternommen (vgl. Beitrag 2 in diesem Band). Ein Grund für diese Anstrengungen ist sicherlich, dass das Abitur unter ganz besonderem Druck der Öffentlichkeit steht. Schon dreimal war die Hochschulzulassung in der Medizin Gegenstand von Entscheidungen des Bundesverfassungsgerichts (1972, 1977 und 2017). Aus dem letzten Urteil des Jahres 2017, aber auch aus stetig wiederholten politischen Forderungen nach Chancengleichheit über Bundesländer hinweg leitet sich die große Aufmerksamkeit der Bildungspolitik für das Thema ab, die letztlich auch jene Anstrengungen des IQB initiiert hat, über die im vorliegenden Buch in einzelnen Beiträgen berichtet wird. Das Buch zeigt, wie intensiv konzeptionell, gestaltend und evaluierend an der Vergleichbarkeit des Abiturs gearbeitet wird. Jeder Schritt in Richtung auf

1 Der Verfasser engagiert sich seit langem zu dem Thema. Als Schüler half er, eine Demonstration gegen den Numerus Clausus zu organisieren, und schrieb 1971 in der Schülerzeitung über Benotungsmaßstäbe in der Oberstufe: „Es wird behauptet: Verglichen mit anderen Schulen, werden in Altenkirchen gute Noten für schlechtere Leistungen gegeben. Die Problematik dieser Behauptung liegt darin, dass niemand sie je objektiv beweisen konnte". Als Wissenschaftler arbeitete er am Test für Medizinische Studiengänge und an der TIMSS-Oberstufenstudie mit und erforschte Kriterien der Notengebung. Dieser Essay ist daher sowohl eine Zwischenbilanz aus persönlicher Sicht als auch ein Nachwort zu dem vorliegenden Buch.

das Ideal eines transparenten und gerechten Bildungssystems ist zu begrüßen. Aber der Band zeigt auch, wie schwierig es ist, sichtbare und nachhaltige Fortschritte zu machen. Der gemeinsame Abituraufgabenpool als solcher, auch die Vorgabe gemeinsamer Bewertungshinweise für die Lehrkräfte, scheint noch nicht auszureichen.

Am Schluss des Bandes soll im vorliegenden Beitrag noch einmal grundsätzlich gefragt werden, welche Strategien und Maßnahmen für eine Stärkung der Vergleichbarkeit überhaupt infrage kommen, wie man eine belastbare empirische Grundlage für diese Arbeit herstellt und welche Erwartungen an die Vergleichbarkeit der Abiturnoten realistisch sind. Bevor diese Fragen in den Abschnitten III bis V des Beitrags diskutiert werden, wird zunächst skizziert, warum Vergleichbarkeit eine fundamentale Bedeutung für die Logik der Notengebung hat (I) und wie sich der Disput darüber im Spannungsfeld von Politik, Recht, Pädagogik und empirischer Forschung darstellt (II).

I Vergleichbarkeit von Abschlussnoten als notwendige, aber kontrafaktische Annahme

Noten geben idealerweise eine Rückmeldung an die Lernenden, informieren Lehrkräfte wie Eltern über den Leistungsstand und unterstützen so den pädagogischen Prozess in Schulen. Unabhängig davon, ob sich dieses Ideal *im Verlauf* von Lehr-Lern-Prozessen und Bildungsgängen erfüllen lässt[2], können Noten *am Ende* eines Bildungsgangs, als Abschlussnoten, solche pädagogischen Zwecke sicherlich nicht mehr erfüllen. Hier werden Erfolge und Misserfolge vielmehr abschließend bilanziert und für Dritte (weiterführende Bildungseinrichtungen, Arbeitgeber und potenzielle Förderer) aufbereitet. Das hat Rückwirkung in die Lehr-Lern-Praxis und die Familien hinein, weil allen Beteiligten dieses summative Urteil als Chance oder Bedrohung vor Augen steht. Vor allem aber erfüllen Abschlussnoten eine prospektive Aufgabe: Sie sind gewissermaßen die „Währung", in der man erworbenes Wissen und Können in verdichteter, möglichst quantifizierter Form ausdrücken, vergleichen und im wahrsten Sinne des Wortes „vermarkten" kann[3].

In den Worten des Soziologen Niklas Luhmann geht es im Erziehungssystem der Gesellschaft (organisatorisch getragen von Schulen) nicht nur um die Vermittelbarkeit von Wissen und Fähigkeiten, sondern auch um die vergleichende Bewertung der Ergebnisse solcher Vermittlung als „besser" oder „schlechter". Dieser „Selektionscode", materialisiert in Zensuren und Zertifikaten, kann nach Luhmann (2002, S. 65)

2 Diese Frage kann und muss im vorliegenden Kontext nicht aufgearbeitet werden. Sicherlich gibt es informativere, für Lernende und Lehrende hilfreichere Formen der lernprozessbegleitenden Leistungserfassung und -rückmeldung als Ziffernnoten. Aber im vorliegenden Beitrag geht es allein um die Leistungsbewertung am Ende des gymnasialen Bildungsgangs, nicht um diagnostische Praktiken in dessen Verlauf.

3 Mit der Reduktion von Bildungsergebnissen auf eine Zahl unterscheiden sich Noten grundlegend von Kompetenzmodellen, wie sie in den Bildungsstandards verwendet werden. Diese wurden ja eingeführt, um qualitativ diagnostizieren zu können, was Lernende schon wissen und können und wie die nächste Stufe ihrer Kompetenzentwicklung im betreffenden Fach aussehen kann.

auch von Unbeteiligten nachvollzogen und so in andere gesellschaftliche Systeme (z. B. Wirtschaft, Recht, Wissenschaft, aber auch Familie) hinein kommuniziert und dort genutzt werden, um Entscheidungen zu fällen und soziale Chancen zu vergeben.

Dieser für unsere Gesellschaft fundamentale Prozess funktioniert nur, sofern man den Noten – speziell den Abschlussnoten – unterstellt, über Individuen, Organisationen (Schulen), Länder und sogar über Kohorten (Geburtsjahrgänge) hinweg eine einheitliche Bedeutung zu haben, mit anderen Worten: „vergleichbar" zu sein. Im wörtlichen Sinne sind Noten natürlich immer vergleichbar: eine 2 als Berliner Physiknote aus dem Jahr 2020 ist besser als eine 4 als Chemienote aus Köln im Jahr 2010, so wie 2 Pfund Äpfel leichter sind als 4 Pfund Birnen. Was mit der Frage der „Vergleichbarkeit" gemeint ist, bezieht sich jedoch auf das Problem, ob die Notengebung, analog zum Wiegen, widerspruchsfrei als Messoperation auf einer allgemeinen, pädagogisch-inhaltlich interpretierbaren Skala verstanden werden kann. Dazu müssten zumindest Rangordnungen von Notenwerten (Note x ist kleiner oder größer als Note y) übersetzbar sein in Beziehungen inhaltlicher Art (das mit x Bewertete ist besser oder schlechter als das mit y Bewertete). In unserem Beispiel würde die gemeinsame Skala so etwas wie „naturwissenschaftliche Schulleistung" erfassen. Wenn die beiden naturwissenschaftlichen Fächer bei der Berechnung von Abitur-Gesamtergebnissen austauschbar sind, ist tatsächlich auch die fächerübergreifende Vergleichbarkeit ihrer Noten mitgedacht. Je nach Abiturregelung dazu, welche Fächer in die Gesamtnote eingebracht werden können, kann die Vergleichbarkeit über Fächer hinweg noch breiter angelegt sein.

Wenn man die Forderung nach Vergleichbarkeit von Noten so fundamental und allgemeingültig ausdrückt, ist offensichtlich, dass sie in der Realität gar nicht erfüllbar ist. Für pädagogische Prozesse ist das kein Problem: den pädagogischen Zweck können Noten prinzipiell auch dann erfüllen, wenn sie nur für einen bestimmten Jahrgang, ein Fach, eine Klasse oder gar eine Person auf derselben Skala abgebildet werden können. Aber für Abschlussnoten ist die Annahme, dass sie im breiteren Sinne vergleichbar seien, unerlässlich. Diese Annahme wird, wie Klemm in seinem historischen Abriss zeigt (vgl. Beitrag 1 in diesem Band), in länderübergreifenden Regularien vom Deutschen Bund 1834 bis zur KMK 2021 implizit unterstellt, jedoch gar nicht erst explizit angesprochen, also weder belegt noch hinterfragt. Föderale Vereinbarungen betrafen und betreffen die formalen Rahmungen für Bildungsgänge und Prüfungen, erst seit wenigen Jahrzehnten (mit Einführung der Einheitlichen Prüfungsanforderungen, EPA, sowie später der Bildungsstandards) auch das Curriculum, aber nicht die Details konkreter Bewertungsprozesse.

Legitimation hatten Noten trotzdem, weil gesellschaftlich akzeptiert war, dass sie von Personen verantwortet werden, die einen Beurteilungsspielraum besitzen, wenn sie sich im Rahmen der Normen ihrer Profession und der Regelungen der Kultusbürokratie bewegen. Das juristische Fundament dafür bilden innerhalb der KMK abgestimmte Verordnungen zur Notengebung, welche die Notenstufen definieren und dabei alle Lehrkräfte auf eine sogenannte kriteriale, absolute Bezugsnorm verpflichten: Noten sollen nicht etwa die individuelle Lernentwicklung („ipsative Bezugsnorm")

oder den Vergleich mit Mitschülerinnen und Mitschülern derselben Lerngruppe („soziale Bezugsnorm") ausdrücken, sondern den Grad der Erfüllung verbindlicher Leistungserwartungen. Die Notenstufen sind „unbestimmte Rechtsbegriffe", müssen aber fair und nach „allgemein anerkannten Bewertungsmaßstäben" ausgelegt werden (Avenarius & Hanschmann, 2019). Somit wird juristisch einerseits unterstellt, dass Noten einen objektiven, kriterialen Bezug haben, andererseits aber anerkannt, dass jede konkrete Notenvergabe auf Entscheidungen beruht, die extern – auch gerichtlich – nur sehr bedingt nachprüfbar sind.

Forschungsergebnisse sprechen tatsächlich dafür, dass die strenge Annahme der Vergleichbarkeit empirisch nicht haltbar ist. Im Schulalltag überwiegt die soziale Bezugsnormorientierung, mitunter auch die – für prozessbegleitende Leistungsbewertung pädagogisch-psychologisch empfohlene – ipsative Norm (Rheinberg, 2001; Bürgermeister, 2014). Schon seit den 1960er-Jahren haben Karlheinz Ingenkamp und seine Mitstreiter in Studien der Pädagogischen Diagnostik wieder und wieder gezeigt, dass Noten fehlerbehaftet, verzerrt, durch Einstellungen und Urteilstendenzen beeinflusst und somit nur begrenzt vergleichbar waren, vor allem über Schulklassen hinweg betrachtet (Ingenkamp, 1989). Neuere Forschungen analysieren den Urteilsprozess im Detail (Brookhart et al., 2016). Noten kombinieren, gewichten und bewerten eine Vielzahl von Einzelbeobachtungen und spiegeln unterschiedliche fachliche und fachunabhängige Kriterien (Klieme, 2003; Rakoczy et al., 2008, Westphal et al., 2016). Die „diagnostische Kompetenz" der Lehrkräfte ist dabei begrenzt (Südkamp & Praetorius, 2017). Im vorliegenden Band bestätigt Beitrag 9 erneut die Probleme der Notengebung, indem er bedeutsame, nach Fach, Land und Einzelschule schwankende Abweichungen zwischen Kursnoten und Ergebnissen zentraler Abiturprüfungen belegt.

Einzuräumen ist, dass sich nur eine Minderheit der erwähnten Forschungsarbeiten auf Abschlussnoten bezieht. Hinsichtlich der Messqualität und Vergleichbarkeit von Abiturnoten bestehen große Forschungslücken (vgl. Abschnitte IV und V). Es wäre aber verwunderlich, wenn Noten im Abitur plötzlich ganz andere Qualität hätten als im Schulalltag. Selbst mit analytischen Bewertungsregeln lassen sich die Unterschiede zwischen Beurteilenden nicht wesentlich verringern, wie Beitrag 8 dokumentiert.

Die Vergleichbarkeit von Abiturnoten stellt also einerseits, streng genommen, eine kontrafaktische Annahme dar. Andererseits belegen Studien immer wieder die „ausgezeichnet Validität der Durchschnittsnote der Hochschulzugangsberechtigung" zur Prognose des Studienerfolges (Hell, Trapmann & Schuler 2008, S. 45). Das wäre unmöglich, wenn man Noten gar nicht vergleichen könnte. Es kann also nicht um ein Entweder-Oder gehen, sondern um den *Grad* der Vergleichbarkeit: Wie lässt sich dieser beziffern? Wieviel Vergleichbarkeit ist notwendig, um Studienplätze fair zu vergeben, und wie viel Nicht-Vergleichbarkeit ist gesellschaftlich wie individuell akzeptabel? Über diese Fragen wird in unterschiedlichen Kontexten gestritten.

II Debatten zur Vergleichbarkeit im Spannungsfeld von Politik, Recht, Pädagogik und Bildungsforschung

Bereits Ingenkamp (1969, S. 412) brachte die Benotungsprobleme auf den Punkt: „Der eigentliche diagnostische Prozess in der Schule vollzieht sich methodisch ungenau und unkontrolliert und führt zu unzuverlässigen Ergebnissen"; „Übermäßig ausgeprägte soziale Selektion, Überforderung eines Teils der Schüler und mangelnde pädagogische Förderung eines anderen sind die Folge" (ebd., S. 427). Er entwarf auch eine Therapie des Missstands in Gestalt von Vergleichstests: „Zu bestimmten Zeitpunkten wird die Anwendung von Leistungs- und Eignungstests vorgeschrieben, damit der Lehrer einen besseren Vergleichsmaßstab und ein Korrektiv für sein Urteil erhält", und „diese Arbeiten schaffen die Grundlagen für Schulplanung, Lehrplanforschung und langfristige Leistungsvergleiche" (S. 428). Bemerkenswert ist, dass Tests nicht – wie Kritiker aus der pädagogischen Zunft schon damals behaupteten – die Urteile der Lehrkräfte ersetzen oder einem ökonomischen Kalkül unterwerfen sollten, sondern diesen wie auch der Bildungsadministration und der Forschung eine transparente Basis für professionelles Urteilen und Handeln geben sollten.

Nach Abschaffung des Deutschen Bildungsrates brauchten die Länder sich ab 1975 mit solchen Debatten nicht mehr zu beschäftigen. Zugleich wurde aber auf politischer und juristischer Ebene das Problem des Numerus Clausus übermächtig. Wenn Zehntel oder Hundertstel einer Notenstufe über die Zulassung zum Medizinstudium entschieden, reichte es nicht aus, auf die professionelle Seriosität der Notengebung zu vertrauen. Als Reaktion auf ein Urteil des Bundesverfassungsgerichts aus dem Jahr 1972 führten die Länder die Einheitlichen Prüfungsanforderungen für das Abitur ein (vgl. Beitrag 2) und veränderten die Hochschulzulassung, wobei jedoch zwei Ersatz-Annahmen bemüht wurden (Fay, 1982): (1) Zunächst wurden Abiturnoten mit länderspezifischen Bonus-/Malus-Punkten „adjustiert", später Landes-Quotierungen eingeführt. Beides in der (auch wieder kontrafaktischen) Annahme, dass mangelnde Vergleichbarkeit nur zwischen den Ländern, aber nicht innerhalb der Länder problematisch sei. (2) Sodann wurden Abiturnoten durch weitere Zulassungskriterien wie Studieneingangstests, Wartezeit und Motivationsgespräche ergänzt – wobei viele Jahre lang kaum hinterfragt wurde, dass solche Korrektive ja das Vergleichbarkeitsproblem der Noten nicht aufheben konnten. Die Annahmen wurden sogar dadurch geschützt, dass man am Vergabeverfahren beteiligten Wissenschaftlerinnen und Wissenschaftlern untersagte, Noten und deren Zusammenhänge mit Testleistungen im Ländervergleich zu untersuchen. Studienbewerber:innen wurden damit vertröstet, dass das Verfahren unterschiedliche, auch wiederholte Chancen zum Studieneinstieg erlaube. Dies reichte viele Jahre lang auch vor Gericht aus. Ende der 1990er-Jahre wurde der verpflichtende Medizinertest abgeschafft – nicht wegen technischer oder juristischer Probleme, sondern wegen des Rückgangs der Bewerberzahlen (Klieme, 1987; Trost, 1998) – und die Hochschulen erhielten das Recht zur eigenständigen Auswahl eines Teils der Studienplätze.

Um die Annahme der Vergleichbarkeit von Noten, speziell Abschlussnoten, gründlich und nachhaltig zu erschüttern, bedurfte es neuer Erkenntnisse aus der Bildungsforschung. Jürgen Baumert und Kollegen zeigten, dass Schüler:innen aus der Oberstufe von Integrierten Gesamtschulen in NRW bei gleicher Mathematiknote sehr viel niedrigere Leistungen in einem standardisierten Vergleichstest aufwiesen als Gymnasiastinnen und Gymnasiasten desselben Jahrgangs (Köller, Baumert & Schnabel, 1999). Hier schrillten politische Alarmglocken, ging es doch einerseits um die jahrzehntelang umkämpfte Existenz von Gesamtschulen, andererseits um unfaire Behandlung bei der Studienplatzvergabe zu*un*gunsten von Schichten, die traditionell privilegiert und einflussreich waren. Es half nichts, dass die Autorengruppe die positive Bedeutung der Gesamtschulen für einen alternativen, tatsächlich niederschwelligeren Einstieg ins Studium für bildungsferne Schichten hervorhob. Stattdessen wurde die ursprünglich nur in einer Fußnote angedeutete Forderung nach einheitlichen „Standards" der Benotung zu einem Leitbild der Bildungsreform. Noch vor dem PISA-Schock wurde damit begonnen, Ingenkamps Idee umzusetzen – zunächst in der Primarstufe und der Sekundarstufe I. Aber auch für Oberstufe und Abitur wurde zunehmend ein besserer Vergleichsmaßstab eingefordert, wie der vorliegende Band dokumentiert, und man leitete verschiedene „Standardisierungs- und Annäherungsprozesse" ein (vgl. Beitrag 2).

Juristische Auseinandersetzungen um die Hochschulzulassung setzten sich unabhängig davon fort. Sie werden ja weder durch Analysen der Wissenschaft noch durch politische Reformprojekte stimuliert, sondern durch die Bearbeitung strittiger Fälle auf der Basis geltenden Rechts – hier des Hochschulrahmenrechts und zugehöriger Länderregelungen. Als das Verwaltungsgericht Gelsenkirchen 2014 einen solchen Fall zu entscheiden hatte, rief es das Bundesverfassungsgericht (BVerfG) an, weil fraglich sei, ob die betreffenden Regelungen grundgesetzkonform seien. Eine entscheidende Rolle spielte hierbei die Vergleichbarkeit der Abiturnoten. Das BVerfG fasste die entsprechenden Argumente des Verwaltungsgerichts folgendermaßen zusammen: „Die Abiturnote sei für die Einschätzung der Qualifikation des einzelnen Bewerbers nur bedingt zuverlässig und gewährleiste eine nur eingeschränkte Vergleichbarkeit der Bewerber. In der Notengebung existierten nicht nur statistisch auffällige Niveauunterschiede zwischen den Ländern [...], sondern es schlügen auch die Unterschiede in den konkreten Umständen der Unterrichtung und Benotung an den einzelnen Schulen sowie individuelle Besonderheiten des jeweiligen Schülers zu Buche und stünden so einer uneingeschränkten Vergleichbarkeit der Bewerber allein anhand der Abiturnote entgegen" (BVerfG, 2017, Rn. 53). Tatsächlich sprach auch das höchste Gericht in seinem Urteil von einem „länderübergreifenden Vergleichbarkeitsdefizit der Abiturnoten", führte dazu aber aus: „Dieses beruht nicht auf Detailunschärfen, die jedem Vergleich von Prüfungsnoten innewohnen [genannt werden der Bewertungsspielraum der Lehrer:innen und die Abhängigkeit der Bewertung vom Lernumfeld, E. K.], sondern ist in den länderspezifisch unterschiedlichen Bildungs- und insbesondere auch Bewertungssystemen angelegt (...). Solange derartige Bewertungsdifferenzen bestehen, bedarf es der Ausgleichsmechanismen, die zumindest

eine annähernde Vergleichbarkeit der Noten ermöglichen" (BVErfG, 2017, Rn. 182). Der ausschließliche Blick auf Länderunterschiede zeigt sich auch daran, dass die zentrale Auswahl nach dem „Grad der Qualifikation", d. h. der Abiturleistung, gar nicht beanstandet wurde, weil dort Landesquoten gebildet werden. Als Problem galt allein die Verrechnung von Abiturnoten bei der hochschulspezifischen Auswahl. Der Gesetzgeber wurde zur Herstellung einer „annähernden Vergleichbarkeit der Abiturnoten über die Ländergrenzen hinweg" bis Ende 2019 verpflichtet; als mögliches Mittel wurde eine Art Umrechnungstabelle erwähnt: „eine Relationierung der Noten auf Zentralebene, auf die die Hochschulen dann zurückgreifen können" (ebd., Rn. 188).

Hochinteressant ist, wie das Bundesverfassungsgericht im Dezember 2017 mit bildungspolitischen, -praktischen und -wissenschaftlichen Argumenten umging. Der Hinweis der Länder auf laufende Reformen wurde zur Kenntnis genommen („Im Gegensatz zu früher seien die Leistungsanforderungen in der Abiturprüfung deutlich stärker bundeseinheitlich angeglichen und Mindeststandards zur Sicherung der länderübergreifenden Vergleichbarkeit vereinbart"; BVerfG, 2017, Rn. 64), aber deren tatsächliche oder mögliche Wirkungen wurden nicht angesprochen und schon gar nicht im Urteil berücksichtigt. Anders als beim Verwaltungsgericht Gelsenkirchen zählte für das höchste Gericht auch nicht die pädagogische Praxis bzw. Benotungspraxis an einzelnen Schulen; es sah die alleinige (bzw. allein justiziable) Ursache der Probleme in „länderspezifisch unterschiedlichen Bildungs- und Bewertungssystemen" (ebd., Rn. 182). Diese starke Hypothese wurde an der entsprechenden Stelle mit Hinweis auf zwei Fachaufsätze aus dem Wissenschaftsrecht (Haug 2006; Hailbronner 1996) unterlegt, die aber länderspezifische Regelungen nur beschreiben und deren behauptete Kausalität für Vergleichbarkeitsdefizite in keiner Weise stützen können. An anderer Stelle (BVerfG, 2017, Rn. 178) wird auf „den Vergleich der Länderabiturdurchschnitte sowie empirische Studien" verwiesen. Der in KMK-Statistiken ausgewiesene Niveau-Unterschied der Noten-Durchschnitte wird vom Gericht fälschlich als Beleg für mangelnde Vergleichbarkeit angesehen (vgl. im Detail unten, Abschnitt IV). Die „empirischen Studien" bestehen aus einem einzigen Aufsatz, der Schulnoten und standardisierte Testergebnisse für zwei Fächer in zwei Bundesländern untersucht (Neumann et al., 2009). Dessen Autoren konstatieren in der Tat, „dass Bedenken hinsichtlich der Vergleichbarkeit von Abiturleistungen zwischen den Bundesländern durchaus berechtigt sind" (ebd., S. 707). Allerdings erwiesen sich die Noten im Fach Englisch als „weitgehend vergleichbar" (ebd.), und für das Fach Mathematik wurde der Länderunterschied in Benotungsmaßstäben zu einem bedeutenden Teil (Kursnote) bzw. vollständig (Zentralabitur) durch Referenzgruppeneffekte erklärt, also durch die Schülerzusammensetzung an der jeweiligen Einzelschule, und eben *nicht* durch formale Regelungen auf Landesebene. Diese differenzierten Befunde werden vom Verfassungsgericht verkürzt interpretiert.[4] Zugleich zeigen Neumann und Kollegen übrigens, dass zentrale Abiturprüfungen, selbst wenn sie in den Ländern getrennt organi-

4 Das BVerfG bezieht sich hingegen explizit auf die „Betrachtung möglicher Auswirkungen auf die Studienzulassung" am Ende des Aufsatzes, wo hypothetische, recht unrealistische Szenarien durchgespielt werden: Es wird eine Hochschule simuliert, mit Bewerbungen ausschließlich aus Hamburg und Baden-Württemberg, die nur ein einzelnes Kriterium (Test oder Note) verwendet.

siert sind, die Vergleichbarkeit erhöhen – ein für mögliche Reformen des Abiturs wichtiger Befund.

Als Fazit des hier skizzierten (Nicht-)Diskurses kann festgehalten werden, dass Politik, Wissenschaft, Justiz und Schulpraxis ihre Argumente wechselseitig zur Kenntnis nehmen, aber eher punktuell, oberflächlich, in legitimatorischer oder skandalisierender Absicht. Im Kern funktionieren die Systeme gerade so, wie es die Soziologie Niklas Luhmanns beschreibt: selbstreferenziell. Zu den Folgen gehört es, wenn Aussagen und Vorschläge der Wissenschaft in der Praxis missverstanden (z. B. Bereitstellung von Vergleichsmaßstäben als „Ökonomisierung von Pädagogik") und in der Politik mit drei Jahrzehnten Verzögerung umgesetzt werden – oder wenn Fehl- bzw. Überinterpretationen empirischer Befunde in einem Urteil des Bundesverfassungsgerichts unbemerkt und unbeantwortet bleiben.

III Wie kann die Vergleichbarkeit erhöht werden?

Viele Maßnahmen, die in Deutschland etabliert wurden und teilweise im vorliegenden Band angesprochen und dokumentiert sind, betreffen einheitlichere Rahmenbedingungen der zum Abitur führenden Bildungsgänge (z. B. Wahlmöglichkeiten im Kurssystem, Stundentafeln, Anrechnung verschiedener Vorleistungen; vgl. insbes. Beitrag 3 in diesem Band). Inwieweit diese Art von Maßnahmen tatsächlich die Noten, die Schüler:innen im Abitur erhalten, vergleichbarer macht, ist jedoch fast nie geprüft worden. Hübner und Kollegen (2020) haben anhand von Daten aus Thüringen und Baden-Württemberg nachgewiesen, dass eine Verpflichtung, Mathematik auf erhöhtem Anforderungsniveau zu belegen, den Zusammenhang zwischen Noten und standardisierten Leistungstests in den betroffenen Jahrgängen grundlegend verändert. Die Autoren folgern daraus, dass Änderungen solcher Regulierungen, wenn sie nicht einheitlich in allen Ländern erfolgen, die Vergleichbarkeit der Noten gefährden, und fordern ein sorgfältiges Monitoring von Reformen im Kurssystem.

Mit der Einführung landesweit zentral gestellter Prüfungsaufgaben und den bundesweiten gemeinsamen Abituraufgabenpools wurde erstmals die Abiturprüfung selbst zum Gegenstand vereinheitlichender Maßnahmen. Wichtig ist, dass die Notengebung auch dabei in der Hand von Lehrkräften bleibt, welche die Erst- und Zweitkorrektur vornehmen. Der Aufgabenpool und seine Implementation (z. B. durch verbindliche Vorgaben zur Struktur von Aufgaben und Bewertungshinweisen) belassen die Entscheidungshoheit für jeden einzelnen Fall in der Profession, sorgen aber dafür, dass professionelle Standards des Prüfens klarer festgelegt und umgesetzt werden. Auf diesem Weg sind weitere Schritte denkbar:

- Als Inputkontrolle: Eine intensive Fortbildung im Hinblick auf die Vertrautheit mit einheitlichen Prüfungsanforderungen und Bildungsstandards, das Wissen um den diagnostischen Urteilsprozess und die Reflexionsfähigkeit bestärkt Lehrkräfte in ihrer Rolle als Prüfende. Solche Maßnahmen werden etwa in Beitrag 8 des vorliegenden Bandes empfohlen. Allerdings sind mit Beitrag 6 auch die

Grenzen dieser Vorgehensweise sichtbar: Alle Vorgaben unterliegen der Rekontextualisierung in Institutionen und durch einzelne Akteure. In den Worten des Beitrags 6: „Auf Intermediärer Ebene entscheidet sich, wenn nicht ausschließlich, dann doch maßgeblich, ob durch gemeinsame Poolaufgaben bundesweit vergleichbare Abituranforderungen entstehen (können) oder nicht". Die Autorin und der Autor rechnen z. B. mit der Möglichkeit einer „Konformitätsfassade".

- Als Prozesskontrolle: Inspektorinnen/Inspektoren bzw. Aufsichtsbeamtinnen/Aufsichtsbeamte können von außen in die Schule kommen, um an Prüfungen teilzunehmen. Dies ist schon jetzt stichprobenartige Praxis. Prüfungssituationen könnten zudem videogestützt aufgearbeitet, Zweitkorrekturen systematisch ausgewertet werden (vgl. entsprechende Untersuchungsansätze in Beitrag 8). Beispiel hierfür wären auch die Ansätze zum „gemeinsamen Prüfen" in der Schweiz (vgl. Beitrag 4 in diesem Band).
- Als Outputkontrolle: Statistiken über Notenverteilungen, vor allem aber Analysen zum Zusammenhang der standardisierten Prüfungsteile mit nicht standardisierten Prüfungsteilen und Vornoten könnten regelmäßig an die Lehrenden und Prüfenden zurückgemeldet werden. Grundlage hierfür liefern die evaluativen Arbeiten des vorliegenden Bandes, insbesondere Beitrag 9.

Wenn hier von „Kontrolle" die Rede ist, mag dies als Einschränkung der pädagogischen Autonomie verstanden werden. Aber aus meiner Sicht wird die Urteilskompetenz der Lehrkräfte auf diesen Wegen professionalisiert, d. h. gestärkt, und nicht etwa abgebaut. Wichtig wäre, dass solche Maßnahmen länderübergreifend eingeführt und gemeinsam evaluiert werden.

Hiervon zu unterscheiden sind Maßnahmen, die das Urteil der Lehrkräfte ersetzen und somit faktisch entwerten. Sie sind nicht in Deutschland, sondern vor allem in den angelsächsischen Ländern zu finden. In den USA galten Schulnoten der High School, deren Curricula von Distrikt zu Distrikt große Unterschiede aufweisen können, schon in den 1920er-Jahren als wenig vergleichbar, sodass für den Zugang zum College ein landesweit standardisierter Leistungstest eingeführt wurde, der SAT (Scholastic Achievement/Aptitude/Assessment Test). Er wird vom College Board verantwortet und seit 1947 von einer professionellen, nicht-kommerziellen Einrichtung erstellt, dem Educational Testing Service (ETS). ETS betreut auch andere Testsysteme wie den Englischtest TOEFL und einen standardisierten Test zur Zulassung bei Master- und Promotionsstudiengängen, die Graduate Record Examination (GRE). Inzwischen gibt es viele konkurrierende Testprogramme, die innerhalb von Schulen und an Übergängen im Bildungssystem zum Einsatz kommen. Jeder Test besteht aus einer Serie von Testaufgaben (Items, manchmal thematisch zusammengefasst zu „Testlets"), die nach komplexen Verfahren erprobt und skaliert werden. Mit solchen „psychometrischen" Methoden – standardisierte Tests und mathematisierte Messmodelle – arbeiten u. a. auch Australien und die Niederlande (vgl. Beitrag 4 in diesem Band), wo mit Cito in Arnheim und ACER in Melbourne ebenfalls professionelle Testinstitute etabliert sind; alle drei genannten Testinstitute waren oder sind mit derselben Methodologie auch an der Gestaltung internationaler Schulleistungsstudien wie TIMSS,

PIRLS/IGLU und PISA beteiligt. England, Irland (vgl. Beitrag 4 in diesem Band) und viele andere vom englischen Schulsystem geprägte Staaten hingegen arbeiten mit zentralen Prüfungen ohne psychometrische Fundierung, die in traditionellen Bewertungsrunden von trainierten „Markern" (häufig Lehrkräfte im Nebenjob) benotet werden; auch diese „A-Level"-Prüfungen werden oft an nicht staatliche Anbieter vergeben. Ein sehr lesenswerter Bericht des englischen „Office of Qualifications and Examinations Regulation" (Ofqual, 2012) vergleicht dieses System im internationalen Maßstab mit 17 weiteren Abschlussprüfungen, u.a. bezüglich der Rolle von schulinternen vs. externen Prüfungskomponenten und der Verwendung von gebundenen Aufgabenformaten (Multiple Choice), die in England nicht vorkommen.

Es darf bezweifelt werden, dass in Deutschland eine der angelsächsischen Lösungsvarianten als Ersatz für das Abitur Akzeptanz fände. Zum einen widersprechen sie dem hier etablierten Verständnis von professioneller pädagogischer Verantwortung: Wo grundsätzlich angenommen wird, dass Lehrkräfte verpflichtet, willens und in der Lage sind, Schülerleistungen nach kriterialen Bezugsnormen zu bewerten, wäre die Delegation dieser Aufgabe an Organisationen außerhalb des Schulwesens ein Systembruch. Zum zweiten zeigen sich bei näherem Hinsehen zu viele Fragen und Widersprüche, die kaum weniger problematisch sind als die Vergleichbarkeits-Annahme hierzulande:

- Die nicht-psychometrischen Bewertungsverfahren sind sehr aufwändig und letztlich mit vielerlei problematischen Annahmen zur Vergleichbarkeit der Ergebnisse behaftet – sei es, dass man Fachnoten ohne jede Prüfung ihrer Gleichwertigkeit verrechnet oder die Selektivität der Kurswahlen zu berücksichtigen versucht, indem man die Noten auf der Basis der Durchschnittswerte aller Personen, die den betreffenden Kurs gewählt haben, adjustiert („Moderation"; vgl. auch Cuff, 2007). In England werden immer wieder Schwankungen in Aufgabeninhalten, Aufgabenschwierigkeit und Auswertungsprozesse moniert, die einen Vergleich zwischen den Beurteilerteams (Examination Boards) und zwischen verschiedenen Jahrgängen fragwürdig erscheinen lassen (Newton et al., 2007).
- Psychometrische Verfahren wie der SAT erlauben es prinzipiell, die Annahme der Vergleichbarkeit während der Testentwicklung und anhand der Prüfungsdaten systematisch, auf der Basis probabilistischer bzw. statistischer Modelle zu prüfen (vgl. Abschnitt IV unten). Aber dies wird damit erkauft, dass offene, authentische oder mündliche Prüfungsformate nur schwer integrierbar sind, dass auf Items bzw. Testlets nach rein formalen Kriterien verzichtet werden muss, auch wenn sie wichtige Lernziele repräsentieren, und dass wesentliche Teile des Verfahrens intransparent bleiben – sei es, weil „Ankeritems" über Jahre hinweg geheim gehalten werden müssen oder weil Aspekte des Verfahrens ohne sehr spezielle psychometrische Kenntnisse nicht nachvollzogen werden können. Testkritische Untersuchungen in den USA haben auch immer wieder gezeigt, dass es zu Verzerrungen der Testergebnisse kommen kann, etwa wenn verschiedene Gruppen von Prüflingen unterschiedlich gut mit den Tests vertraut sind, weil manche von ihnen nicht an Vorbereitungskursen teilnehmen können, die mit hohen Kosten verbunden sind.

Interessant ist, dass eine „Zentralmatura“ nach den Regeln psychometrischer Tests, die in Österreich auf den Weg gebracht worden war, 2014 nach wiederholten Problemen im Entwicklungsprozess und Einsprüchen eines wissenschaftlichen Beirats politisch gekappt wurde. Und die Schweiz hat sich, wie in Beitrag 4 dokumentiert, grundsätzlich gegen zentralisierte Verfahren entschieden. Deutschsprachige und angelsächsische Traditionen scheinen demnach nur schwer vereinbar zu sein.

Aus diagnostischer Sicht spricht aber vieles dafür, dass eine Hybridlösung am besten geeignet ist, eine kluge Balance zwischen Transparenz und Professionalität des Urteils, Fairness und Offenheit von Chancen zu halten. Die Hinzunahme bundesweit einheitlicher Aufgaben zu einer Abiturregelung, die neben landesweiten Zentralprüfungen sowie Vornoten auch lokal verantwortete Prüfungen einbezieht und alle Teilergebnisse nach transparenten, pragmatisch gesetzten Regeln kombiniert, könnte eine gute Basis bilden (vgl. ähnlich den Vorschlag von Blossfeld et al., 2011). Die Regeln müssten länderübergreifend politisch ausgehandelt und sorgfältig empirisch evaluiert werden.

IV Wie kommt man zu belastbaren empirischen Befunden?

Will man umfassende Vergleichbarkeit im Sinne der Verankerung aller Abiturnoten auf einer gemeinsamen Skala empirisch absichern, muss man mit testbasierten Systemen auf psychometrischer Grundlage arbeiten. Diese erlauben es, für jede einzelne Testaufgabe statistisch zu prüfen, ob sie in unterschiedlichen Konstellationen (z. B. in verschiedenen Ländern, in unterschiedlichen Jahrgängen, nach Geschlecht und Herkunft, unter unterschiedlichen Prüfungsbedingungen) denselben Bezug zur Fähigkeitsdimension hat, die insgesamt gemessen werden soll, z. B. „naturwissenschaftliche Kompetenz“. Die Gleichförmigkeit des Zusammenhangs zwischen Leistungen bei Einzelaufgaben und Gesamtfähigkeit[5] gilt in der Psychometrie als Beleg für die sogenannte Messinvarianz, d. h. die technisch präzise Fassung dessen, was intuitiv mit „Vergleichbarkeit“ gemeint ist. Auf der Basis psychometrischer Modelle gelingt die Prüfung der Messinvarianz auch dann, wenn die einzelnen Teilnehmer:innen unterschiedliche Teilmengen von Aufgaben bearbeiten, etwa weil sie verschiedene Kursarten besucht haben, weil verschiedene Testversionen das „Abschreiben“ verhindern, weil jährlich neue Aufgaben eingefügt werden oder weil das Schwierigkeitsniveau an das Leistungsniveau der Schule angepasst wird.

Wie in Abschnitt III argumentiert, ist aber nicht damit zu rechnen, dass testbasierte Prüfungssysteme wie in den USA bei uns als Regelverfahren eingeführt würden. Um die Vergleichbarkeit herkömmlicher Schul- oder Abschlussnoten zu bewerten, braucht man daher einen Umweg: Man entwickelt einen standardisierten, psychometrischen Test, der wichtige Anforderungen des Abiturs widerspiegelt, setzt

5 Betrachtet wird dabei nicht die Korrelation, weil sie von der Streuung der Fähigkeiten in der jeweiligen Gruppe beeinflusst ist, sondern die mathematische Funktion, welche das Bearbeitungsergebnis bei der Einzelaufgabe (0 = falsch, 1 = richtig) mit der geschätzten Gesamtfähigkeit verknüpft.

ihn flächendeckend oder zumindest in großen repräsentativen Gruppen ein und erhebt parallel die Noten. Wenn Noten und Testergebnisse über alle Gruppen hinweg eng korrelieren[6] und wenn der Zusammenhang zwischen Note und Test möglichst in allen Gruppen die gleiche Form hat, dann kann man die Vergleichbarkeit der Noten – bezogen auf den Vergleichsmaßstab dieses Tests – als „hoch" bezeichnen. Wenn aber beispielsweise in Bundesland B bei gleicher Testleistung schlechtere Noten erzielt wurden als in Bundesland A, so muss die Vergleichbarkeit hinterfragt werden, denn in Land B würde strenger benotet als in Land A.

Bislang gibt es für Deutschland genau einen Datensatz, der solche Auswertungen ansatzweise erlaubt: die TIMSS-Oberstufenstudie aus dem Jahr 1995, die internationale Tests für Mathematik und Naturwissenschaften verwendete. Die bei Baumert und Watermann (2000) publizierten Befunde werden von Neumann et al. (2009, S. 696) folgendermaßen zusammengefasst: „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".

Alle weiteren Publikationen zum Thema waren auf ein oder zwei Bundesländer konzentriert, z. B. Köller et al. (1999) auf Nordrhein-Westfalen, Maag Merkis Analysen zu Effekten der Zentralabiturs (2012) auf Hessen und Bremen, sowie die Publikationen der Arbeitsgruppe um Köller, Neumann und Trautwein (z. B. Hübner et al. 2020; Köller et al. 2004; Neumann et al. 2009; Trautwein et al. 2007, 2010) auf Vergleiche Baden-Württembergs mit Thüringen oder Hamburg. 25 Jahre nach den Konstanzer Beschlüssen der KMK, die eine „empirische Wende" der Bildungspolitik eingeläutet haben, muss man feststellen, dass es ausgerechnet zum Thema „Vergleichbarkeit von Abiturnoten" keinen einzigen bundesweit repräsentativen Datensatz und dementsprechend keine einzige Studie gibt, die beziffern könnte, wie groß die Diskrepanzen der Benotungsmaßstäbe in der Sekundarstufe II zwischen Ländern, zwischen Einzelschulen innerhalb der Länder und auch zwischen einzelnen Lehrkräften sind. Ein Appell von Bildungsforscherinnen und Bildungsforschern, Deutschland möge an der erneuten TIMSS-Oberstufenstudie 2015 teilnehmen, auch um genau diese Frage zu beantworten, wurde in der Bildungspolitik nicht aufgegriffen.

Dementsprechend unterlag das Bundesverfassungsgericht einem Fehlschluss, als es in seinem Urteil von 2017 meinte, die mangelnde Vergleichbarkeit der Abiturnoten anhand der KMK-Notenstatistik begründen zu können (vgl. Abschnitt II). Das Gericht verwechselte Notenunterschiede mit Bewertungsunterschieden: Die Tatsache,

6 Vgl. das Beispiel zum Zusammenhang zwischen Noten und PISA-Testleistungen zu Beginn des Kapitels.

dass sich die Verteilung der Abiturnoten (sei es der Mittelwert, die Streuung oder etwa der Anteil der 1,0-Abschlüsse) zwischen den Ländern unterscheidet, ist eben *kein* Beleg für mangelnde Vergleichbarkeit im Sinne von unterschiedlichen Benotungsmaßstäben. Im Gegenteil: Wenn die Abiturnoten wirklich „vergleichbar" wären, also Wissen oder Kompetenzen oder schulischen Lernerfolg auf derselben Skala abbildeten, würde man aller Wahrscheinlichkeit nach deutlich stärkere Unterschiede zwischen Ländern feststellen, weil die Selektion zum Abitur und vermutlich auch die tatsächlichen Leistungsniveaus sich von Land zu Land unterscheiden (vgl. dazu Beitrag 2 in diesem Band). Nicht die zu *starke* Schwankung von Notendurchschnitten zwischen den Ländern ist das Problem, wie vom Gericht behauptet, sondern die vermutlich zu *schwache* Schwankung – wenn man sie denn an einem objektiven Vergleichsmaßstab messen könnte. Explizit belegen lässt sich diese These, wie gesagt, aber nicht, weil es einen solchen empirischen Vergleichsmaßstab bis heute nicht gibt.

Es gibt ihn hingegen in der Sekundarstufe I, in den IQB-Bildungstrends und früher bei PISA-E. Dort liegen für alle Länder große, repräsentative Datensätze vor. Sie wurden jedoch im Hinblick auf die Frage der Vergleichbarkeit von Noten noch nicht ausreichend ausgewertet. In früheren Aufsätzen (z. B. Klieme 2003) und Vorträgen (z. B. Klieme 2006) habe ich anhand von PISA-E 2000 den Zusammenhang zwischen Durchschnittsnoten der Schüler:innen einerseits und ihren PISA-Testleistungen andererseits untersucht und festgestellt, dass dieser Zusammenhang in den Ländern, aber auch je nach Schulen innerhalb der Länder, unterschiedlich ausfiel. Die Abweichung der tatsächlichen Noten von dem Wert, der auf Basis des PISA-Tests zu erwarten wäre, zeigt an, inwieweit ein spezifischer, vom Vergleichskriterium PISA abweichender Benotungsmaßstab verwendet wurde. Diese Abweichung – und nicht eine beobachtete Notendifferenz – ist ein angemessener Indikator für das *Ausmaß der Bewertungsunterschiede*, anders gesagt: für den *Grad der Nicht-Vergleichbarkeit* von Noten. Die Varianz dieses Indikators in deutschen Gymnasien konnte zu 9,1 % durch Länderunterschiede erklärt werden, während weitere 7 % (je nach Land zwischen 5,3 % und 10,9 %) auf Unterschiede zwischen Einzelschulen innerhalb der Länder zurückzuführen waren. Dieses Beispiel aus PISA-E 2000 illustriert:

(i) Tatsächlich ist ein bedeutsamer Anteil der Varianz in Benotungsmaßstäben auf der Ebene der Länder zu verorten.
(ii) Es ist aber nicht ausreichend, nur auf die Länderebene zu schauen, wenn man institutionelle Unterschiede in der Notengebung betrachtet: Die Ebene der Einzelschulen kann etwa gleich wichtig sein.
(iii) Institutionelle Effekte machen insgesamt nur einen Bruchteil (hier unter 20 %) der Unterschiede in Benotungsmaßstäben aus.

Das Beispiel darf allerdings nicht überinterpretiert werden, weil es veraltet ist, mit PISA-Tests einen für das deutsche Schulsystem nur begrenzt geeigneten Vergleichsmaßstab verwendet, und vor allem, weil es sich nicht direkt auf die gymnasiale Oberstufe und das Abitur bezieht. Es ist sehr zu hoffen, dass das IQB in naher Zukunft bessere Daten zur Verfügung stellen kann, indem es nicht nur detaillierte Analysen

zu Kurs- und Abiturnoten vorlegt (wie in Kapitel 9 dieses Bandes), sondern auch eine bundesweite Statistik von Abiturnoten im Vergleich zu standardisierten Bewertungskriterien. Um Auswirkungen von Reformen des Abiturs einzuschätzen, bräuchte man solch eine bundesweite Statistik sogar als regelmäßiges Monitoring.

V Wie viel Vergleichbarkeit ist realistisch?

Warum also nicht doch an der Annahme der Vergleichbarkeit von Abiturnoten zumindest innerhalb einzelner Fächer festhalten, die ja letztlich auf einer prinzipiell akzeptierten Kombination von Rationalität, pragmatischer Umsetzbarkeit und Vertrauen in professionelles Urteilen beruht? Der Blick auf die Schweiz, in der man sehr bewusst diesen Weg geht, ist hier hilfreich (vgl. Beitrag 4 in diesem Band). Man kann das bestehende System durch stringenten Einsatz der ländergemeinsamen Abituraufgabenpools, Fortbildung der beteiligten Aufgabenentwickler:innen und Lehrkräfte sowie laufendes Monitoring besser machen, wie es KMK und IQB weiter beabsichtigen, aber es wäre unredlich, den Betroffenen und der Öffentlichkeit zu versprechen, dass Abiturnoten jemals im strikten Sinne vergleichbar werden könnten. Wer als Politiker:in oder Interessengruppe oder Journalist:in die Einheitlichkeit des Abiturs auf seine Fahnen schreibt, sollte sich diesen Realitäten stellen. Zumindest wäre es politisch klug, wissenschaftlich redlich und pädagogisch verantwortungsbewusst, wenn man keine Veränderungen des Prüfungssystems über das derzeit in Deutschland geplante Maßnahmenarsenal hinaus unternehmen würde, bevor nicht belastbare Daten vorliegen, die das Ausmaß des Problems einschätzbar und Effekte möglicher Veränderungen evaluierbar machen.

Bei den Abwägungen, die hier zu leisten sind (und die vom Bundesverfassungsgericht 2017 zumindest angedeutet werden, wenn es „*annähernd* vergleichbare Leistungsbewertungen“ fordert), sollte schließlich berücksichtigt werden, dass „Vergleichbarkeit“ nicht automatisch „Validität“ oder/und „Fairness“ einschließt. Von den sogenannten Testgütekriterien der Psychometrie betrifft Vergleichbarkeit nur die Objektivität der Testung (ihre Unabhängigkeit von der Person, die Benotungen vornimmt, und deren subjektiven Maßstäben) und die Reliabilität (Präzision, z. B. ausgedrückt in der Übereinstimmung wiederholter Messungen), ggfs. auch die sogenannte konkurrente Validität, d. h. die Korrelation mit einem externen Vergleichskriterium wie z. B. PISA oder standardbezogene Tests. Die Validität der Benotung im umfassenden Sinne würde u. a. eine gute Prognose des weiteren Bildungserfolgs einschließen. Sie kann, muss aber nicht mit der Vergleichbarkeit einhergehen. In diesem Zusammenhang sei an ein wichtiges Ergebnis der Evaluierung des „Tests für medizinische Studiengänge (TMS)“ in den 1980er- und 1990er-Jahren erinnert: Der standardisierte Test konnte die Ergebnisse in schriftlichen Examina des Studiums etwas besser vorhersagen, die Abitur-Durchschnittsnote hingegen mündliche Examensleistungen (Trost et al., 1998). Und während der Test männliche Bewerber bevorzugte, waren Bewerberinnen bei der Abiturnote im Vorteil. Diese – statistisch recht komplex definierte – Fairness

lässt sich eben nicht an geschlechtsspezifischen Mittelwerten in Tests bzw. Noten ablesen, sondern an der in aufwändigen Längsschnittstudien ermittelten geschlechtsspezifischen Prognosekraft (vgl. auch Fischer, Schult & Hell, 2015). Analoge Untersuchungen zur Validität und Fairness verschiedener Abitur-Varianten liegen, wie oben ausgeführt, derzeit gar nicht vor.

Bringt man die Erkenntnisse des vorliegenden Buches einerseits, die Grundsatzfragen zum Grad der Vergleichbarkeit und seiner empirischen Bestimmung andererseits zusammen, so muss die Schlussfolgerung lauten: IQB und KMK sind mit den Arbeiten am gemeinsamen Aufgabenpool und dessen möglichst einheitlicher Implementierung auf einem sinnvollen Weg, weil sie sich für die Stärkung der pädagogischen Professionalität entschieden haben und nicht für deren Ersatz durch Tests. Um das Ausmaß des Problems und mögliche Fortschritte empirisch untersuchen zu können bedarf es aber noch großer Anstrengungen. Politik, Öffentlichkeit und auch Justiz – bis hin zum Bundesverfassungsgericht – wären vorerst gut beraten, pragmatische Ziele zu verfolgen und sich genauer als etwa im Nachgang zum BVerfG-Urteil von 2017 zu vergewissern, wo Fallstricke und Missverständnisse liegen. Ganz ohne kontrafaktische oder empirisch ungeprüfte Annahmen wird man auch in Zukunft nicht auskommen.

Literatur

Autorengruppe Bildungsberichterstattung (Hrsg.) (2016). *Bildung in Deutschland 2016. Ein indikatorengestützter Bericht mit einer Analyse zu Bildung und Migration*. Bielefeld: W. Bertelsmann.

Avenarius, H. & Hanschmann, F. (2019). *Schulrecht*. C. Link Verlag.

Baumert, J. & Watermann, R. (2000). Institutionelle und regionale Variabilität und die Sicherung gemeinsamer Standards in der gymnasialen Oberstufe. In Baumert, J., Bos, W. & Lehmann, R. (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*. Opladen: Leske + Budrich, S. 317–372.

Blossfeld, H.-P., Bos, W., Daniel, H.-D., Hannover, B., Lenzen, D., Prenzel, M., Roßbach, H.-G., Tippelt, R. & Wößmann, L. (2011). *Gemeinsames Kernabitur. Zur Sicherung von nationalen Bildungsstandards und fairem Hochschulzugang. Gutachten*. Münster: Waxmann.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., . . . Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86,* 803–848.

Bürgermeister, Anika (2014). *Leistungsbeurteilung im Mathematikunterricht*. Münster: Waxmann.

BVerfG – Bundesverfassungsgericht (2017). Urteil des Ersten Senats vom 19. Dezember 2017–1 BvL 3/14,1 BvL 4/14 – Rn. 1–253, http://www.bverfg.de/e/ls20171219_1bvl000314.html

Cuff, B. M. P. (2007) *International approaches to the moderation of non-examination assessments in secondary education*. Coventry: Office for Qualifications and Examinations Regulations.

Fay, E. (1982). *Der „Test für medizinische Studiengänge" (TMS) – Ausgewählte Aspekte seiner Genese*. Braunschweig: Agentur Pedersen.

Fischer, F., Schult, J. & Hell, B. (2015). Unterschätzung der Studienleistungen von Frauen durch Studierfähigkeitstests: Erklärbar durch Persönlichkeitseigenschaften? *Diagnostica, 61,* 34–46.

Hailbronner, K. (1996). Verfassungsrechtliche Fragen des Hochschulzugangs. *Zeitschrift für Wissenschaftsrecht, 29,* 1 ff.

Haug, V. M. (2006). Hochschulauswahlrecht im Vergleich. *Zeitschrift für Wissenschaftsrecht, 39,* 96–113.

Hell, B., Trapmann, S. & Schuler, H. (2008). Synopse der Hohenheimer Metaanalysen zur Prognostizierbarkeit des Studienerfolgs und Implikationen für die Auswahl- und Beratungspraxis. In Schuler, H. et al. (Hrsg.), *Studierendenauswahl und Studienentscheidung* (S. 43–54). Göttingen: Hogrefe.

Hübner, N., Wagner, W., Hochweber, J., Neumann, M. & Nagengast, B. (2020). Comparing apples and oranges: Curricular intensification reforms can change the meaning of students' grades! *Journal of Educational Psychology, 112,* 202–220.

Ingenkamp, K. (1969). Möglichkeiten und Grenzen des Lehrerurteils und der Schultests. In H. Roth (Hrsg.), *Begabung und Lernen. Gutachten und Studien der Bildungskommission des Deutschen Bildungsrates* (S. 407–447). Stuttgart: Klett.

Ingenkamp, K. (1989). *Diagnostik in der Schule*. Weinheim und Basel: Beltz.

Klieme, E. (1987). Auswahlverfahren mit Chancenausgleich - Simulation von Auswirkungen der neuen Zulassungsregelungen für medizinische Studiengänge. *Empirische Pädagogik, 1,* 209–229.

Klieme, E. (2003). Benotungsmaßstäbe an Schulen: Pädagogische Praxis und institutionelle Bedingungen. Eine empirische Analyse auf der Basis der PISA-Studie. In H. Döbert, B. von Kopp, R. Martini & M. Weiß (Hrsg.), *Bildung vor neuen Herausforderungen. Historische Bezüge, rechtliche Aspekte, Steuerungsfragen, internationale Perspektiven. Hermann Avenarius zum 65. Geburtstag gewidmet* (S. 195–210). Neuwied: Luchterhand.

Klieme, E. (2006). Bildungsstandards als Instrumente zur Harmonisierung von Leistungsbewertungen und zur Weiterentwicklung didaktischer Kulturen. In F. Eder, A. Gastager & F. Hofmann (Hrsg.), *Qualität durch Standards? Beiträge zum Schwerpunktthema der 67. Tagung der AEPF* (S. 55–70). Münster: Waxmann.

Köller, O. (2013). Wege zur Hochschulreife und Sicherung von Standards. In: D. Bosse, F. Eberle, B. Schneider-Taylor (Hrsg.) *Standardisierung in der gymnasialen Oberstufe*. Wiesbaden: Springer VS.

Köller, O., Baumert, J. & Schnabel, K. (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.

Luhmann, N. (2002). *Das Erziehungssystem der Gesellschaft*. Frankfurt a. M.: Suhrkamp.

Maag Merki, K. (Hrsg.). (2012). *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland*. Wiesbaden: Springer VS.

Neumann, M., Nagy, G., Trautwein, U. & Lüdtke, O. (2009). Vergleichbarkeit von Abiturleistungen: Leistungs- und Bewertungsunterschiede zwischen Hamburger und BadenWürttemberger Abiturienten und die Rolle zentraler Abiturprüfungen. *Zeitschrift für Erziehungswissenschaft, 12,* 691–714

Newton, P., Baird, J.-A., Goldstein, H., Patrick, H. & Tymms, P. (Hrsg.). (2007). *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Office for Qualifications and Examinations Regulations (Ofqual) (2012). *International Comparisons in Senior Secondary Assessment*. Coventry: Ofqual.

Rakoczy, K., Klieme, E., Bürgermeister, A. & Harks, B. (2008). The interplay between student evaluation and instruction. *Zeitschrift für Psychologie, 2,* 111–124.

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

Südkamp, A. & Praetorius, A.-K. (Hrsg.). (2017). *Diagnostische Kompetenz von Lehrkräften*. Münster : Waxmann.

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

Trautwein, U., Neumann, M., Nagy, G., Lüdtke, O. & Maaz, K. (Hrsg.). (2010): *Schulleistungen von Abiturienten: Die neu geordnete gymnasiale Oberstufe auf dem Prüfstand*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Trost, Günter (Hrsg.). (1998). *Evaluation des Tests für medizinische Studiengänge (TMS): Synopse der Ergebnisse*. Bonn: Institut für Test- und Begabungsforschung.

Westphal, A., Becker, M., Vock, M., Maaz, K., Neumann, M. & McElvany, N. (2016). The link between teacher-assigned grades and classroom socioeconomic composition: The role of classroom behavior, motivation, and teacher characteristics. *Contemporary Educational Psychology, 46,* 218–227.

Autorinnen und Autoren

Christoph Deeken war wissenschaftlicher Mitarbeiter im Projekt MaLeMINT am IPN | Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik in Kiel. Derzeit ist er als Studienrat am Berufsbildungszentrum Rendsburg-Eckernförde in den Fachbereichen Mathematik und Wirtschaft tätig.
Kontakt: christoph.deeken@bbz-rd-eck.de

Alexander Groß ist Akademischer Rat an der Universität Koblenz-Landau, Campus Koblenz im Arbeitsbereich Bildungssystem- und Schulentwicklungsforschung. Im Rahmen seiner Promotion beschäftigt er sich mit der Implementation der Gemeinsamen Abituraufgabenpools der Länder als bildungspolitische Steuerungsmaßnahme.
Kontakt: gross@uni-koblenz.de

Prof. Dr. Aiso Heinze ist Direktor der Abteilung Didaktik der Mathematik am IPN | Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik in Kiel und Professor für Didaktik der Mathematik an der Universität zu Kiel. Seine Forschung beschäftigt sich mit dem Mathematiklernen vom Kindergartenalter bis in das Studium.
Kontakt: heinze@leibniz-ipn.de

Dr. Lars Hoffmann ist Wissenschaftlicher Mitarbeiter am Institut zur Qualitätsentwicklung im Bildungswesen (IQB) an der Humboldt-Universität zu Berlin und dort für die Evaluation des Einsatzes von Aufgaben der Pools in den Abiturprüfungen der Länder zuständig. Seine Arbeits- und Forschungsschwerpunkte sind: Forschung zu Abitur und Abiturprüfungen, Privatschulen, Diagnostische Kompetenz von Lehrkräften sowie Quer- bzw. Seiteneinsteigende in den Lehrerberuf.
Kontakt: lars.hoffmann@iqb.hu-berlin.de

Dr. Nicolas Hübner ist Juniorprofessor am Institut für Erziehungswissenschaft (IfE) an der Eberhard Karls Universität Tübingen. Zentrale Inhalte seiner aktuellen Forschung sind die Untersuchung der Implementation und Effekte von Bildungsreformen sowie weiteren Programmen und Interventionen im Bildungswesen (z. B. Auslandsaufenthalte oder digitale Lernformate). Weitere Arbeiten beschäftigen sich mit der effektiven Gestaltung von Professionalisierungsprozessen von Lehrkräften sowie mit der (differenziellen) Entwicklung von Kompetenzen und der Motivation von Schülerinnen und Schülern, insbesondere in MINT-Fächern.
Kontakt: nicolas.huebner@uni-tuebingen.de

Prof. Dr. Jörg Jost ist Sprachdidaktiker an der Universität zu Köln. Seine Arbeits- und Forschungsschwerpunkte liegen in den Bereichen Diagnostik und Förderung sprachlicher Kompetenzen, auf Bildungsstandards bezogene Kompetenzmessungen, Schreiben, Text und Lesen.
Kontakt: joerg.jost@uni-koeln.de

Prof. Dr. Michael Kämper-van den Boogaart ist seit 1997 Professor für Neuere deutsche Literatur und Fachdidaktik Deutsch an der Humboldt-Universität zu Berlin. Zu seinen Forschungsschwerpunkten zählen neben der Geschichte des Abituraufsatzes curriculare Aspekte des Deutschunterrichts, diachrone und synchrone Perspektiven auf Fachlichkeit und Wissen und die Modellierung literarischer Rezeptionskompetenzen.
Kontakt: michael.kaemper-van.den.boogaart@rz.hu-berlin.de

Prof. i. R. Dr. Klaus Klemm hatte nach einem Lehramtsstudium mit den Fächern Deutsch und Geschichte und einem wirtschaftswissenschaftlichen Ergänzungsstudium von 1977 bis zu seiner Pensionierung 2007 an der Universität Duisburg-Essen eine Professur für empirische Bildungsforschung und Bildungsplanung inne. Seine Arbeitsschwerpunkte liegen in den Arbeitsfeldern Bildungsfinanzierung, regionale Schulplanung, Lehrerbedarfsplanung sowie Inklusion.
Kontakt: kl.klemm@t-online.de

Prof. Dr. Eckhard Klieme ist Research Fellow am DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation, wo er bis 2020 als Direktor tätig war. Zuvor war er von 1983 bis 1997 Referent im Projekt „Test für Medizinische Studiengänge“ am Bonner Institut für Test- und Begabungsforschung, von 1998 bis 2001 Mitarbeiter am Max-Planck-Institut für Bildungsforschung u. a. in der TIMSS-Studie. Seine Arbeitsgebiete umfassen: Pädagogische Diagnostik, Unterrichts- und Schulforschung.
Kontakt: klieme@dipf.de

Prof. Dr. Svenja Mareike Schmid-Kühn ist Leiterin des Arbeitsbereichs Bildungssystem- und Schulentwicklungsforschung im Fachbereich 1: Bildungswissenschaften an der Universität Koblenz-Landau, Campus Koblenz. In der Forschung beschäftigt sie sich mit dem Schulsystem (v. a. mit aktuellen bildungspolitischen Entwicklungen im Gymnasialbereich) und der Steuerung im Bildungswesen, der Schulentwicklungsforschung sowie mit Fragen aus dem Bereich der Lehrer:innenbildung und des Lehrer:innenberufs.
Kontakt: schmid-kuehn@uni-koblenz.de

Dr. Irene Neumann ist Leiterin der Forschungsgruppe „Lehren und Lernen an der Schnittstelle zwischen Physik und Mathematik“ am IPN | Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik in Kiel. Ihre Arbeits- und Forschungsschwerpunkte sind u. a. mathematische Kompetenzen am Übergang von der Schule zur Hochschule, die Rolle mathematischer Kompetenzen für das Physiklernen sowie Nature of Science im Physikunterricht.
Kontakt: ineumann@leibniz-ipn.de

Dr. Marko Neumann ist wissenschaftlicher Mitarbeiter am DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation Frankfurt am Main/Berlin und leitet dort den Arbeitsbereich Bildungsstrukturen und Reformen. Seine Arbeits- und Forschungsschwerpunkte sind: Reformprozesse und Qualitätsentwicklung im Bildungswesen, Einfluss institutioneller Lernumwelten auf die Entwicklung schulischer Leistungen und psychosozialer Merkmale, Bildungsentscheidungen und Übergänge im Bildungssystem, Gymnasiale Oberstufe und Abitur.
Kontakt: marko.neumann@dipf.de

Prof. Dr. Sabine Reh ist Professorin für Historische Bildungsforschung an der Humboldt-Universität zu Berlin und Direktorin der Bibliothek für Bildungsgeschichtliche Forschung (BBF) des DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation. Zu ihren Forschungsschwerpunkten gehören die Geschichte der Schule und der Praktiken des (Fach-)Unterrichts, insbesondere des Prüfens, ebenso wie die Geschichte des pädagogischen und erziehungswissenschaftlichen Wissens, insbesondere nach 1945.
Kontakt: sabine.reh@dipf.de

Anja Riemenschneider ist wissenschaftliche Mitarbeiterin am Institut zur Qualitätsentwicklung im Bildungswesen (IQB) an der Humboldt-Universität zu Berlin und dort im Bereich der Evaluation des Einsatzes von Aufgaben der Pools in den Abiturprüfungen der Länder tätig. Im Rahmen ihrer Promotion beschäftigt sie sich mit der Messung und Bewertung von sprachlicher Komplexität von Textvorlagen und -produktionen im Abitur.
Kontakt: anja.riemenschneider@iqb.hu-berlin.de

Dr. Pauline Schröter ist wissenschaftliche Mitarbeiterin am Institut zur Qualitätsentwicklung im Bildungswesen (IQB) an der Humboldt-Universität zu Berlin und dort für die Evaluation des Einsatzes von Aufgaben der Pools in den Abiturprüfungen der Länder zuständig. Zuvor war sie am Max-Planck-Institut für Bildungsforschung in der Forschungsgruppe „Reading Education and Development“ tätig. Zu ihren Forschungsinteressen gehören die Messung und Bewertung sprachlicher Kompetenzen, die Entwicklung von Mehrsprachigkeit sowie Open Science und Wissenschaftskommunikation.
Kontakt: pauline.schroeter@iqb.hu-berlin.de

Hannelore Söldner ist Koordinatorin für das Fach Deutsch in der Sekundarstufe II am Institut zur Qualitätsentwicklung im Bildungswesen (IQB) an der Humboldt-Universität zu Berlin. Davor war sie als Lehrerin für die Fächer Deutsch und Geschichte an verschiedenen Gymnasien in Bayern sowie als Referentin im Referat „Auslandsschulwesen“ im Sekretariat der Kultusministerkonferenz tätig.
Kontakt: hannelore.soeldner@iqb.hu-berlin.de

Prof. Dr. Petra Stanat, Ph.D. ist wissenschaftlicher Vorstand des Instituts zur Qualitätsentwicklung im Bildungswesen (IQB) an der Humboldt-Universität zu Berlin. Ihre Arbeits- und Forschungsschwerpunkte sind: Bildungsqualität und Bildungsmonitoring, Bedingungen und Förderung des Bildungserfolgs von Heranwachsenden aus zugewanderten und sozial benachteiligten Familien, Zweitsprachförderung und Lesekompetenz.
Kontakt: petra.stanat@iqb.hu-berlin.de

Prof. Dr. Dorothee Wieser ist seit 2014 Professorin für Neueste deutsche Literatur und Didaktik der deutschen Sprache und Literatur an der Technischen Universität Dresden. Ihre Arbeits- und Forschungsschwerpunkte sind u. a.: literaturdidaktische Professionalisierungsforschung, Verstehen von Metaphern in literarischen Texten, Wissensvermittlung im Literaturunterricht sowie Interpretationskulturen.
Kontakt: dorothee.wieser@tu-dresden.de

Kaum ein Bildungsthema wird so anhaltend und intensiv diskutiert wie das Abitur, das aufgrund von Länderunterschieden vielfach als „unvergleichlich" wahrgenommen wird. Dieser Band richtet einen wissenschaftlichen Blick auf den Diskurs und gibt einen Überblick über den aktuellen Stand der Forschung.

In Teil 1 werden historische und aktuelle Entwicklungen der Abiturprüfungen beschrieben und zentrale strukturelle Rahmenbedingungen – auch im internationalen Vergleich – dargestellt. Teil 2 bündelt die Ergebnisse aktueller Forschungsprojekte zum Abitur aus verschiedenen bildungswissenschaftlichen Disziplinen. Eine abschließende Reflexion zu bestehenden Herausforderungen im Hinblick auf die Qualität und Vergleichbarkeit des Abiturs rundet den Band ab. Damit bietet der Band sowohl eine Übersicht über das Forschungsfeld als auch eine Vertiefung in Spezialfragen und will so zur empirischen Fundierung, Differenzierung und Versachlichung der andauernden Diskussion über das Abitur beitragen.

ISBN: 978-3-7639-7056-8

wbv Publikation

wbv.de