

George, Ann Cathrice [Hrsg.]; Zintl, Ramona [Hrsg.]; Helm, Christoph [Hrsg.]
**10 Jahre Bildungsstandarterhebungen – Bedingungen (langfristiger)
Kompetenzentwicklungen auf Schul- und Systemebene**

Münster ; New York : Waxmann 2025, 328 S. - (Kompetenzmessungen im österreichischen Schulsystem: Analysen, Methoden & Perspektiven; 4)



Quellenangabe/ Reference:

George, Ann Cathrice [Hrsg.]; Zintl, Ramona [Hrsg.]; Helm, Christoph [Hrsg.]: 10 Jahre Bildungsstandarterhebungen – Bedingungen (langfristiger) Kompetenzentwicklungen auf Schul- und Systemebene. Münster ; New York : Waxmann 2025, 328 S. - (Kompetenzmessungen im österreichischen Schulsystem: Analysen, Methoden & Perspektiven; 4) - URN: urn:nbn:de:0111-pedocs-340341 - DOI: 10.25656/01:34034; 10.31244/9783818850593

<https://nbn-resolving.org/urn:nbn:de:0111-pedocs-340341>

<https://doi.org/10.25656/01:34034>

in Kooperation mit / in cooperation with:



WAXMANN
www.waxmann.com

<http://www.waxmann.com>

Nutzungsbedingungen

Dieses Dokument steht unter folgender Creative Commons-Lizenz: <http://creativecommons.org/licenses/by/4.0/deed.de> - Sie dürfen das Werk bzw. den Inhalt vervielfältigen, verbreiten und öffentlich zugänglich machen sowie Abwandlungen und Bearbeitungen des Werkes bzw. Inhaltes anfertigen, solange Sie den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

This document is published under following Creative Commons-License: <http://creativecommons.org/licenses/by/4.0/deed.en> - You may copy, distribute and render this document accessible, make adaptations of this work or its contents accessible to the public as long as you attribute the work in the manner specified by the author or licensor.

By using this particular document, you accept the above-stated conditions of use.



Kontakt / Contact:

peDOCS

DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation

Informationszentrum (IZ) Bildung

E-Mail: pedocs@dipf.de

Internet: www.pedocs.de

Mitglied der


Leibniz-Gemeinschaft



Ann Cathrice George, Ramona Zintl, Christoph Helm (Hrsg.)

10 Jahre Bildungsstandarderhebungen – Bedingungen (langfristiger) Kompetenz- entwicklungen auf Schul- und Systemebene



WAXMANN

 Institut des Bundes
für Qualitätssicherung im
österreichischen Schulwesen

Kompetenzmessungen im österreichischen Schulsystem: Analysen, Methoden & Perspektiven

herausgegeben vom
Institut des Bundes für Qualitätssicherung
im österreichischen Schulwesen (IQS)

Band 4

Ann Cathrice George, Ramona Zintl, Christoph Helm (Hrsg.)

10 Jahre Bildungsstandarderhebungen – Bedingungen (langfristiger) Kompetenz- entwicklungen auf Schul- und Systemebene



Waxmann 2025
Münster · New York

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter <http://dnb.dnb.de> abrufbar.

**Kompetenzmessungen im österreichischen Schulsystem:
Analysen, Methoden & Perspektiven, Bd. 4**

ISSN 2628-9598

Print-ISBN 978-3-8188-0059-8

E-Book-ISBN 987-3-8188-5059-3

DOI 10.31244/9783818850593

Waxmann Verlag GmbH, 2025
Steinfurter Straße 555, 48159 Münster
www.waxmann.com
info@waxmann.com

Umschlaggestaltung: Hannes Kaschnig-Löbel, IQS
& Anne Breitenbach, Münster
Coverfoto: stock.adobe.com – ABCDESIGN
Satz: Andreas Kamenik, IQS

Dieses E-Book steht unter der Lizenz CC BY 4.0 open access zur Verfügung.



Diese Lizenz gilt nur für das Originalmaterial. Alle gekennzeichneten Fremdinhalte (z.B. Abbildungen, Fotos, Zitate etc.) sind von der CC-Lizenz ausgenommen und für deren Wiederverwendung ist es ggf. erforderlich, weitere Nutzungsgenehmigungen beim jeweiligen Rechteinhaber einzuholen.

Inhalt

Vorwort.....	7
--------------	---

Christoph Helm, Ann Cathrice George, Herbert Altrichter

10 Jahre Bildungsstandarderhebungen – wissenschaftliche Einordnung, Bilanzen und Perspektiven.....	11
--	----

Systemebene

Christoph Weber, Johann Bacher & Michael Bruneforth

Änderungen in der sozialen und ethnischen Schulzusammensetzung im Zeitverlauf	31
---	----

Michael Bruneforth

Der Einfluss der Klassengröße auf Mathematikkompetenzen in der österreichischen Volksschule.....	69
--	----

Silvia Salchegger & Nele Kampa

Warum schneiden Mädchen bei der Mathematik-Zentralmatura schlechter ab als Burschen? Ein Erklärungsversuch zu Geschlechterungleichheiten im österreichischen Bildungssystem.....	107
--	-----

Lisa Wiesinger, Silvia Salchegger & Christina Wallner-Paschon

Das Zusammenspiel von Unterricht und Kompetenzen: Sind Volksschulkinder in Mathematik besser geworden und in Naturwissenschaft zurückgefallen, weil sich der Unterricht verändert hat?.....	141
---	-----

System- und Schulebene

Carina Gargitter & Burkhard Gniewosz

Einfluss von Schuleffektivitätsfaktoren über den fairen Vergleich hinaus.....	175
---	-----

Jana Groß Ophoff, David Kemethofer, Christoph Helm, Ramona Zintl,

Katharina Meusburger, Martina Ott

Maßnahmen der Qualitätssicherung und -entwicklung als „Motor“ schulischen Lernens? Befunde zur Leistungsentwicklung	203
---	-----

Schulebene

Christoph Helm, Ann Cathrice George, Ramona Zintl & Christoph Weber

Mechanismen der Klassenzusammensetzung und ihre Effekte auf die Schulleistung.....237

Klassenebene

Tanja Auer, Burkhard Gniewosz & Marion Reindl

Der Big-Fish-Little-Pond-Effekt im Zeitvergleich.....273

Juliane Schmich & Veronika Österbauer

Binnendifferenzierung und Individualisierung im Volksschulleseunterricht:

Österreich und ausgewählte EU-Vergleichsländer im 15-Jahres-Trend.....297

Autorinnen und Autoren.....329

Liebe Leserin, lieber Leser!

Mit dem vorliegenden Sammelband „10 Jahre Bildungsstandarderhebungen – Bedingungen (langfristiger) Kompetenzentwicklungen auf Schul- und Systemebene“ liegt der vierte Band der vom Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen (IQS) koordinierten Reihe „Kompetenzmessungen im österreichischen Schulsystem: Analysen, Methoden & Perspektiven“ vor Ihnen. Dies zeigt eindrucksvoll die zunehmende Etablierung der Reihe mit den Schwerpunkten Vertiefende Analysen zu (inter-)nationalen Kompetenzmessungen, Vorstellung, Erläuterung und Weiterentwicklung quantitativer Methoden im Bildungsbereich sowie wissenschaftlicher Diskurs von bildungspolitischen Entwicklungs- und Handlungsperspektiven.

Der Fokus des aktuellen Bands ist es, den Datenschatz, der in den vergangenen zehn Jahren durch die kontinuierlichen Erhebungen der österreichischen Bildungsstandards (Bildungsstandardüberprüfungen; BIST-Ü) generiert wurde, für die vertiefende Analyse längsschnittlicher Fragestellungen – die gleichermaßen für Bildungsforschung und Bildungspolitik relevant sind – verstärkt zu nutzen. Diese wurden im Zeitraum zwischen 2012 und 2019 zum Zweck der standardisierten Überprüfung von aus dem Lehrplan abgeleiteten Kompetenzen in Deutsch, Mathematik und Englisch aller Schülerinnen und Schüler der vierten Schulstufe (Volksschule) und der achten Schulstufe (Sekundarstufe I) erhoben. Ergänzt wurden die insgesamt acht Vollerhebungen (drei auf der Primarstufe und fünf in den Schulen der Sekundarstufe I) um zwei Ausgangerhebungen in den Jahren 2009 und 2010.

Die verbindenden Elemente dieser Studien sind die Volksschulen bzw. die Schulen der Sekundarstufe I, da auf diesen Ebenen die Kompetenzen wiederholt und systematisch erfasst wurden. Vor diesem Hintergrund eignen sich diese Daten zur Bearbeitung längsschnittlicher Fragestellungen; mit vorliegenden Analysen dieser Art soll ein bis dato mit österreichischen Daten wenig untersuchtes Feld erschlossen werden.

Kernziel der Analysen ist es, jene Faktoren zu identifizieren, die über die Zeit hinweg beobachtbar die Kompetenzen von österreichischen Schülerinnen und Schülern beeinflussen, dabei aber gleichzeitig stabil gegenüber punktuellen gesellschaftlichen Einflüssen wie beispielsweise verstärkter Zuwanderung oder Pandemien bleiben. Aus den vielfältigen Ergebnissen sollen Ableitungen für die Qualitätsentwicklung in Österreich vorgeschlagen werden, die Erkenntnisse als Basis für die Weiterentwicklung nationaler Kompetenzmessungen dienen sowie richtungsweisend für weitere vertiefende wissenschaftliche Fragestellungen sein.

Die einzelnen Kapitel des Bands fokussieren – mit ihren Fragestellungen – auf die unterschiedlichen Ebenen des österreichischen Bildungssystems, d. h. auf die Systemebene (Kapitel 2 bis 5), auf die System- und Schulebene (Kapitel 6 und 7), auf die Schulebene (Kapitel 8) und die Klassenebene (Kapitel 9 und 10), und sind inhaltlich in dieser Reihenfolge kaskadenartig sortiert. Jedes Kapitel endet zudem mit möglichen (bildungspolitischen) Handlungsableitungen. Der Band wird einleitend eingeführt mit der Einordnung sämtlicher Ergebnisse in den wissenschaftlichen Kontext und daraus resultierenden Richtungen für die Qualitätsentwicklung und Bildungsforschung (Kapitel 1).

Auch darf ich an dieser Stelle die Gelegenheit nutzen und darauf hinweisen, dass sämtliche Analysen in diesem Band mit Daten aus der Forschungsdatenbibliothek (FDB) des IQS umgesetzt wurden. Interessierte Wissenschaftlerinnen und Wissenschaftler können in diesen Daten aus (inter-)nationalen Kompetenzmessungen beantragen, um selbstständig Fragestellungen vertiefend zu beforschen (nähere Informationen liefert die Infobox auf der nächsten Seite).

Abschließend möchte sich das IQS – stellvertretend für die Herausgeberinnen und Herausgeber – bei all jenen bedanken, die zur erfolgreichen Umsetzung dieses Sammelbands einen wesentlichen Beitrag geleistet haben, und hoffen, dass dieses Werk einen wertvollen Einblick in die Kompetenzentwicklung bei österreichischen Schülerinnen und Schülern über den Zeitverlauf liefern wird.

A handwritten signature in black ink, appearing to read 'Robert Klinglmair', written in a cursive style.

Robert Klinglmair
Direktor IQS

Informationen zur Forschungsdatenbibliothek des IQS

Die Analysen im vorliegenden Band wurden an Daten durchgeführt, die in der Forschungsdatenbibliothek (FDB) des IQS dokumentiert, archiviert und Wissenschaftlerinnen und Wissenschaftlern nach positiv abgeschlossenen Antragsverfahren zugänglich gemacht werden.

Neben Informationen zum Antragsverfahren werden auf der Webseite der FDB (<https://www.iqs.gv.at/fdb>) die österreichischen Daten aus nationalen und internationalen Kompetenzerhebungen zur Vorbereitung von Forschungsanträgen transparent gemacht. Für jeden Forschungsdatensatz finden sich dort Kontextfragebögen, Codebücher und Almanache (univariate deskriptive Statistiken der verfügbaren Variablen).

Die Wahrung der Anonymität der Untersuchungseinheiten und der Grundsatz der Datenminimierung sind im IQS-Gesetz¹ bzw. in der DSGVO² einerseits gesetzlich verankert und tragen andererseits einen wesentlichen Anteil zur Akzeptanz der Erhebungen bei. Die Weiterverarbeitung der BIST-Daten im Rahmen wissenschaftlicher Fragestellungen ist unter Einhaltung dieser Bedingungen nicht nur gestattet,³ sondern explizit gewünscht.

Die FDB gewährleistet damit die Anonymität der Untersuchungseinheiten durch ihr Sicherheitskonzept, in dem klare Verantwortlichkeiten festgelegt werden, sowie durch die Bereitstellung von technisch anonymisierten Forschungsdaten.

Mit dem R-Paket BIFIEsurvey (BIFIE, 2024; siehe auch Bruneforth, Oberwimmer & Robitzsch, 2016) und den BIFIE-SPSS-Makros (BIFIE, 2020) stellt die FDB zudem Werkzeuge zur Verfügung, die speziell für die Analysen von FDB-Forschungsdaten entwickelt wurden.

Literatur

- BIFIE (2024). *BIFIEsurvey: Tools for survey statistics in educational assessment*. R package version 3.6-6 [Software]. Verfügbar unter <https://CRAN.R-project.org/package=BIFIEsurvey>
- BIFIE (2020). *SPSS-Makros*. Version v1.8 [Software]. Verfügbar unter <https://www.bifie.at/fdb>
- Bruneforth, M., Oberwimmer, K. & Robitzsch, A. (2016). Reporting und Analysen. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandardüberprüfungen* (S. 333–362). Wien: facultas.

1 BGBl. I Nr. 25/2008 i. d. F. I 32/2018, § 3 Abs. 1 Z. 5, § 6 Abs. 3 bzw. § 7b Abs. 1
 2 Art. 5 lit c VO (EU) 2016/679
 3 Art. 5 lit b und 89 VO (EU) 2016/679 sowie § 2d Abs 4 FOG

Christoph Helm, Ann Cathrice George, Herbert Altrichter

10 Jahre Bildungsstandarderhebungen – wissenschaftliche Einordnung, Bilanzen und Perspektiven

Das Ziel des vorliegenden Bands ist es, den Datenschatz, der sich in den vergangenen zehn Jahren durch die kontinuierlichen Erhebungen der österreichischen Bildungsstandards ergeben hat, für die Bearbeitung einiger längsschnittlicher Fragestellungen, die gleichermaßen für Bildungsforschung und Bildungspolitik relevant sind, zu nutzen.

2008 wurden die Bildungsstandards durch eine Novelle des Schulunterrichtsgesetzes (BGBl. I. 117/2008) rechtlich für das österreichische Bildungswesen gefasst. Ihre Verordnung im Jahr 2009 spezifizierte konkret formulierte Lernergebnisse für die Primarstufe in Mathematik und Deutsch sowie für die Sekundarstufe I in Mathematik, Deutsch und Englisch (die sogenannten „Bildungsstandards“; BIST) und zielte auf eine „Veränderung der Unterrichts- und Schulpraxis durch Kompetenzorientierung“ ab (Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen [IQS], 2025). Die Umsetzung dieser „Bildungsstandard-Politik“ umfasste eine Reihe von Begleitmaßnahmen, wie Unterrichtsmaterialien, diagnostische Tests, Beispielaufgaben, Fortbildungskurse und andere Publikationen, die die Implementation dieser Standards und – genereller – von kompetenzorientiertem Unterricht unterstützen sollten (Altrichter & Gamsjäger, 2017). Einen wesentlichen Stellenwert dabei nahmen die seit 2012 durchgeführten begleitenden Erhebungen auf der vierten und achten Schulstufe zu den Schülerinnen- und Schülerleistungen in den durch die Bildungsstandards definierten Kompetenzbereichen ein (bezeichnet als „Bildungsstandardüberprüfungen“; BIST-Ü). Insgesamt wurden im Zeitraum zwischen 2012 und 2019 acht Vollerhebungen durchgeführt (vgl. Abb. 1), drei in den Volksschulen und fünf in den Schulen der Sekundarstufe; die Datenlage umfasst zudem zwei stichprobenartig umgesetzte Ausgangserhebungen in den Jahren 2009 und 2010 („Baseline“-Erhebungen). Die Schulen waren verpflichtet, an den Kompetenzerhebungen teilzunehmen.

Die Ergebnisse der einzelnen BIST-Ü wurden allen Ebenen des Schulsystems in jeweils unterschiedlichen Formaten rückgemeldet: Im Sinne einer sog. „evidenzorientierten Bildungsentwicklung und -politik“ sollten diese Ergebnis- und Datenrückmeldungen sowohl als empirische Grundlage zur Systementwicklung auf der Ebene der Bildungspolitik wie auch als Baustein zur Schul- und Unterrichtsentwicklung in den Regionen und Standorten dienen. Als Serviceleistung erhielten auch Schülerinnen und Schüler Rückmeldung zu ihrem Kompetenzstand.

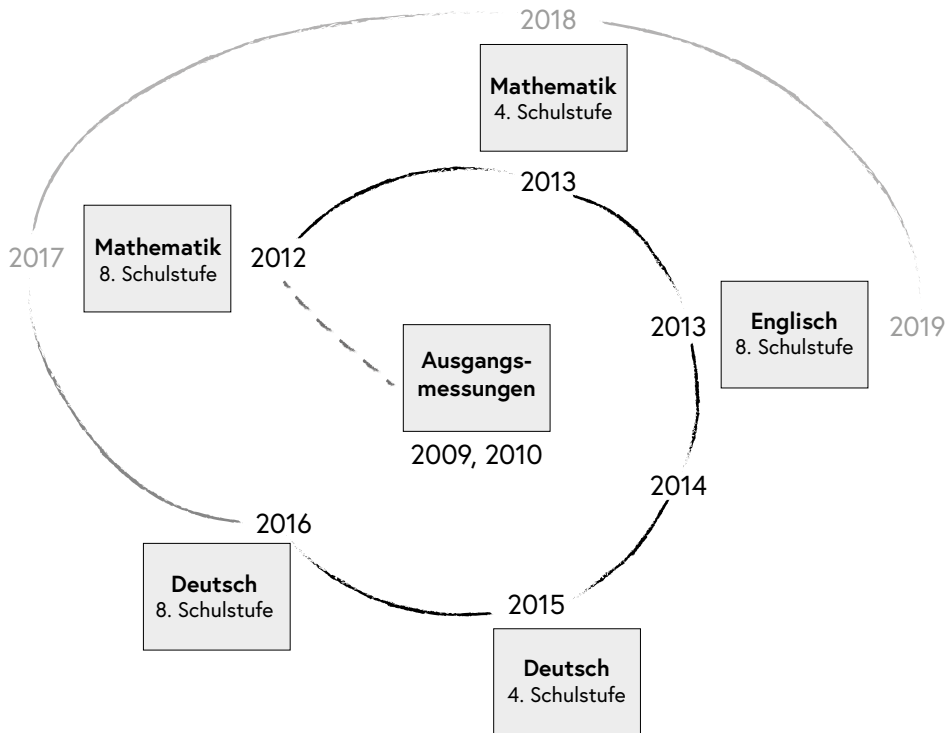


Abbildung 1: Zyklus der Standardüberprüfungen (in Anlehnung an Schreiner et al., 2018, S. 12)

Pro Fach und Jahrgangsstufe (Mathematik/4 – M4, Mathematik/8 – M8, Deutsch/4 – D4, Deutsch/8 – D8, Englisch/8 – E8) verfolgten die BIST-Ü das Ziel einer längsschnittlichen Leistungstrendmessung: Ausgehend von den in der Baseline-Studie in der Baseline verankerten 500 Punkten können nach jeder Wiederholung der Überprüfung die Leistungsänderungen des Gesamtsystems und der einzelnen Schulen nachvollzogen werden (Schreiner et al., 2018, S. 12). Der Längsschnitt der Leistungen über die Zeit ist ein wesentliches Element der BIST-Ü, welches an internationale Vorbilder im Large-Scale Assessment wie TIMSS/PIRLS oder PISA angelehnt wurde (z. B. Mullis & Martin, 2022). Auch in Deutschland gab es mit dem PISA-Schulpanel zeitgleich Bemühungen, auf nationaler Ebene Leistungstrends messen zu können (Bischof, Hochweber, Hartig & Klieme, 2013). Bei allen BIST-Ü-Wiederholungsmessungen (M4: 2010, 2013, 2018; M8: 2009, 2012, 2017; D4: 2010, 2015; D8: 2009, 2016; E8: 2009, 2013, 2019) konnten deutliche Leistungssteigerungen im System beobachtet werden. Die Leistungssteigerungen in Englisch zwischen 2010 und 2013 wurden im Zusammenhang mit den Maßnahmen der Bildungsstandard-Politik positiv evaluiert (Schmid, Trendtel, Bruneforth & Hartig, 2019).

Zusätzlich zu den Leistungswerten wurde für Schulen und Unterrichtsgruppen ein „Erwartungswert“ berichtet. Der berechnete Erwartungswert (George & Robitzsch, 2014) beschreibt die von einer Schule/Unterrichtsgruppe erwartete Leistung unter Berücksichtigung von wesentlichen Aspekten, die nicht von der Schule oder dem Unterricht be-

einflusst werden können (z. B. Urbanisierungsgrad am Schulstandort, Schulgröße, Schultart, sozioökonomische Herkunft der Schülerinnen und Schüler). Dies ermöglicht es, die Wirkung der am jeweiligen Standort geleisteten pädagogischen Arbeit besser zu erkennen (ebd.). Auch ist dadurch „fairerer Vergleich“ mit anderen Schulstandorten oder Unterrichtsrgruppen möglich.

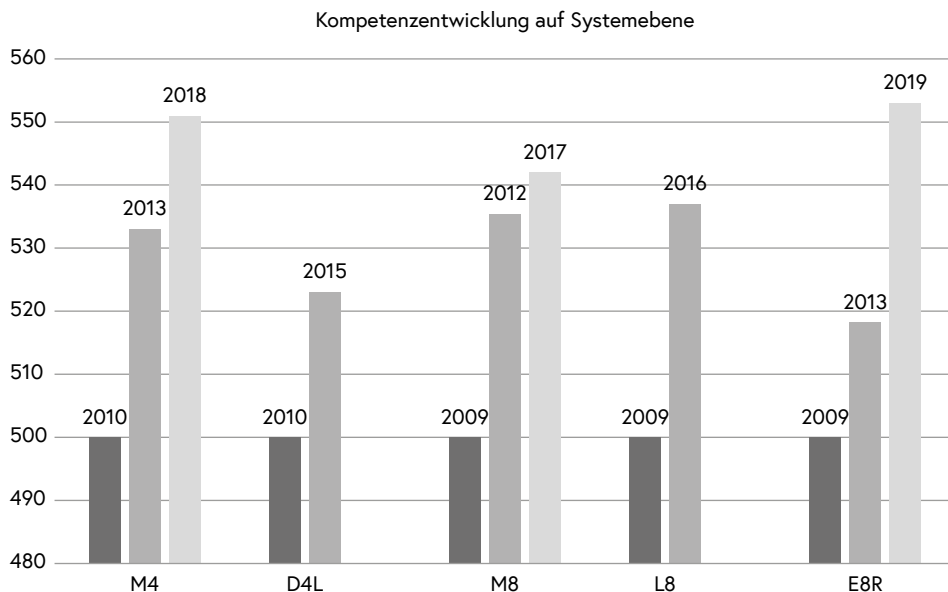


Abbildung 2: Zeitpunkt der BIST-Ü-Erhebungen mit jeweiligem Punktemittelwert. Erhebungen im gleichen Fach und der gleichen Jahrgangsstufe sind durch das Design der BIST-Üs verbunden, sodass Trends festgestellt werden können. Abgetragene Punkte für Mathematik 4 Schulstufe (M4), Mathematik 8 Schulstufe (M8), Deutsch Lesen 4 Schulstufe (D4L), Deutsch Lesen 8 Schulstufe (L8) und Englisch 8 Schulstufe (E8R).

Weitere Informationen zu den Grundlagen der Bildungsstandarterhebungen können George, Schreiner, Wiesner, Pointinger und Pacher (2019) entnommen werden. Im Jahr 2019 wurden die regelmäßigen BIST-Ü eingestellt; seit dem Schuljahr 2021/2022 wird an deren Stelle die verpflichtende „individuelle Kompetenzmessung PLUS“ eingesetzt (iKM^{PLUS}; Stauber, Stahl, Bruneforth & Illetschko, 2024).

1 Zielsetzung des vorliegenden Bands

Sowohl für die Systementwicklung als auch für die Einzelschulentwicklung stehen mehr und mehr Informationen aus unterschiedlichen Datenquellen und unterschiedlichen Zeitpunkten zur Verfügung, beispielsweise aus internationalen Studien (z. B. PISA, PIRLS, TIMSS, ICLIS), aus nationalen Erhebungen, die die Volksschule oder die Sekundarstufe betreffen (z. B. BIST-Ü), aus Ergebnissen der standardisierten kompetenzorientierten Diplom- oder Reifeprüfung (SRDP), Daten aus der amtlichen Statistik oder Studien bei Erwachsenen (z. B. PIAAC). Trotzdem werden Entscheidungen und Konsequenzen häufig aus einzelnen

Datenquellen abgeleitet. Im Kontext der Schulentwicklung rät Schildkamp (2019) dazu, von der Betrachtung dieser einzelnen „Datensilos“ abzuweichen und stattdessen Handlungsempfehlungen auf einer gemeinsamen (multiperspektivischen) Betrachtung (Kombination, Synthese) verschiedener Datenquellen zu basieren. Während diese ganzheitlichere Betrachtungsweise zwar für die Schulentwicklungspraxis empfohlen wird, scheint diese auf Ebene der Systementwicklung bisher kaum berücksichtigt zu werden. Häufig werden einzelne Large-Scale-Studien (bspw. PISA 2022) bereits zum Anlass für Entwicklungen im System genommen, ohne Verbindungen zu anderen Studien (bspw. den BIST-Ü) herzustellen. Auch im Rahmen von BIST-Ü wurden bisher weitestgehend die einzelnen Erhebungen isoliert voneinander betrachtet. So erfolgte beispielsweise die Aufbereitung und Auswertung der Bildungsstandarderhebungen auf Bundesebene jeweils nach jeder Erhebung in Form eines öffentlich verfügbaren Bundesergebnisberichts (<https://www.iqs.gv.at/downloads/archiv-des-bifie/bildungsstandardueberpruefungen/ergebnisberichte>), in dem der jeweilige Kompetenzstand der Schülerinnen und Schüler erläutert wurde. Kompetenzentwicklungen auf System- und Schulebene wurden nur im gleichen Fach auf der gleichen Stufe, also bezogen auf die jeweilige Ausgangsmessung und Vollerhebungen, die eventuell vorher stattgefunden haben, mitgeteilt. Auch die wissenschaftlichen Forschungsarbeiten, die aus den pseudonymisierten Datensätzen der BIST-Ü entstanden sind (vgl. dazu <https://www.iqs.gv.at/themen/bildungsforschung/forschungsdatenbibliothek/daten-der-bildungsstandardueberpruefungen>), fokussieren zuallermeist auf die Daten einer einzelnen Erhebung. Eine Ausnahme bildet der erste Band der hier vorliegenden Buchreihe „Fünf Jahre flächendeckende Bildungsstandardüberprüfungen in Österreich“, in welchem Forschungsfragen in der Zusammenschau über alle Bildungsstandarderhebungen zwischen 2012 und 2016 betrachtet wurden.

Diese Limitation fordert Studien, die mehrere Datensätze synergetisch nutzen, denn auch im Schulsystem sollte nicht nach jeder Einzelerhebung eine „Aktion“, Intervention oder Reform erfolgen. Ein komplexes Gebilde wie das Schulsystem braucht gewisse Zeit, um Reformen in der Praxis zu verankern. Vielmehr sollte das erarbeitete Wissen in geordneter Weise gesichert, kombiniert und in Beziehung gesetzt werden, sodass Handlungsoptionen nach und nach deutlich werden (vgl. Bruneforth & Diedrich, 2024; im Kontext von Schulentwicklung).

Der Fokus des hier vorliegenden Bands liegt daher auf der Analyse mehrerer Datensätze, die im Zeitverlauf erhoben wurden. Eine Verbindung der Erhebungen (jeweils in der Volksschule und auf der Sekundarstufe I) ist über die Schulen hinweg möglich, welche an jeder Erhebung unabhängig vom Fach teilgenommen haben. Die Datenlage erlaubt, (trotz statistisch-methodischer Einschränkungen, siehe unten) u. a. folgende Fragen zu stellen.

1. Sind spezifische Befunde über mehrere Erhebungen hinweg (womöglich gar über den Erhebungszeitraum von 10 Jahren hinweg) stabil?
2. Können mögliche Veränderungen in diesem Zeitraum erklärt werden?
3. Oder legen die Daten nahe, dass gesellschaftliche Ereignisse, wie beispielsweise die Flüchtlingsbewegungen 2015, Einfluss auf die Befunde nehmen?

Bei den Fragestellungen handelt es sich um Veränderungen des (Bildungs-)Systems über die Zeit. So können beispielsweise Lehrmethoden im Zeitverlauf betrachtet werden oder Qualitätsmerkmale von Schulen. Die Verlinkung der Erhebungen über die Schulen hinweg unterscheidet sich von echten Längsschnitten, in denen die Entwicklung der Leistungen einzelner Schülerinnen und Schüler über einen Zeitraum hinweg beobachtet wird. Die Betrachtung kann mit dem Vorgehen in internationalen Large-Scale-Studien verglichen werden, in denen Schulsysteme von Ländern über die Zeit hinweg beobachtet werden.

Die BIST-Ü eignen sich für die Untersuchung von Fragen im Zeitverlauf besonders, da sowohl die Kompetenzen der Schülerinnen und Schüler wiederholt erfasst werden, ebenso wie auch eine Reihe interessierender Variablen (z. B. Unterrichtsqualitätsaspekte, Personal an Schulen). Die in diesem Band vorgestellten Studien nutzen dieses Potenzial und kombinieren jeweils mehrere Large-Scale-Datensätze (zumeist mehrere BIST-Ü).

2 Zum Potenzial der gemeinsamen Analyse verschiedener Bildungsstandarderhebungen

Auf *Ebene des Bildungssystems* können Erhebungen im gleichen Fach und der gleichen Jahrgangsstufe zu unterschiedlichen Messzeitpunkten direkt verbunden werden; dadurch können direkt Schlüsse über Kompetenzentwicklungen getroffen werden. Darüber hinaus können jene Informationen, die begleitend zu jeder Bildungsstandarderhebung durch die sog. „Kontextfragebögen“ erfasst wurden (IQS, o. J.), im Kohortenvergleich sowie in ihren Zusammenhängen zur Leistung betrachtet werden. Dadurch stehen Informationen zu Merkmalen der Schülerinnen und Schüler (soziodemografische, bildungsbiografische und persönliche Merkmale wie Motivation oder Selbstkonzept) den teilnehmenden Schulen (Schulgröße, soziale Zusammensetzung, Betreuungsangebote, Entfernung zur nächstgelegenen gymnasialen Schule) sowie zur Schul- und Unterrichtsgestaltung (Umgang mit Evaluationsergebnissen, Förderkonzepte, Abstimmung mit Lehrkräften, Zielvereinbarungen, Unterrichtsmethoden) zur Verfügung.

Während das Erhebungsdesign keine direkte Messung longitudinaler Kompetenzentwicklungen von der vierten auf die achte Schulstufe zulässt, bieten die Erhebungen auf der vierten Jahrgangsstufe in Mathematik 2013 und der achten Jahrgangsstufe in Mathematik 2017 einen Ausnahmefall. Durch den Abstand von vier Jahren zwischen beiden Studien handelt es sich um die identische Schülerkohorte in beiden Erhebungen. Zwar besteht keine Möglichkeit, die Schülerinnen und Schüler aus der vierten Stufe auf der achten Stufe direkt zu identifizieren und miteinander zu verbinden, doch erlauben statistische Methoden eine wahrscheinlichkeitsbasierte Zuordnung, wodurch bestimmte Entwicklungen sichtbar gemacht werden können (Süss-Stepancik, Ober, George & Varelija-Gerber, 2022).

Auf *Ebene der Schulen* können Erhebungsergebnisse schulbezogen verbunden werden, da ein Großteil der Standorte an allen BIST-Ü ab 2012 teilgenommen hat. Über gleiche Fächer und Kompetenzbereiche hinweg können Leistungsänderungen innerhalb der Schulen direkt

analysiert werden. Über die Erhebungen in den verschiedenen Fächern hinweg ist dies zwar nicht möglich, eine relative Platzierung der Schulen kann aber beurteilt werden: Ist eine Schule in allen Erhebungen über dem Kompetenzmittelwert platziert? Wie schneidet die Schule in den verschiedenen Erhebungen im Hinblick auf den Erwartungsbereich ab? Dadurch können Gruppierungen von Schulen mit verschiedenen im Kontextfragebogen erhobenen Schülerinnen- und Schüler- oder Schulmerkmalen in Beziehung gesetzt und diskutiert werden.

Auf *Ebene des Unterrichts* lassen sich keine echten Längsschnitte bilden, weil die Klassen nicht an mehreren Erhebungen teilgenommen haben. Allerdings sind Pseudo-Längsschnitte (Kohortenvergleiche) möglich. Beispielsweise kann der Frage nachgegangen werden, ob sich der Einsatz bestimmter Unterrichtsmethoden im Zeitverlauf für die gesamte Population verändert hat (siehe bspw. Schreiner & Helm, 2024). Zudem ist durch die Aggregation von Unterrichtsmerkmalen auf Schulebene ein Längsschnitt auf Schulebene über mehrere Erhebungen hinweg möglich. So können bspw. die an den Schulen von den Lehrenden im Durchschnitt eingesetzten Lehrmethoden über die Zeit hinweg bezüglich ihrer Entwicklung analysiert werden. Auch die mit den Veränderungen einhergehenden Auswirkungen auf die Schulleistung können so analysiert werden.

Die auf den Ebenen System, Schule und Unterrichtsklasse/-gruppe beispielhaft angesprochenen Fragestellungen zeigen einiges von dem Potenzial, das Bildungsstandarderhebungen wie auch andere Large-Scale-Studien für das Verständnis des Bildungssystems in einem Land, der Schulentwicklung am Standort und der Aus-, Fort- und Weiterbildung von Lehrpersonen haben können (Bos, Gebauer & Postlethwaite, 2018). So können empirisch gestützte Analysen als Hintergrundwissen zur (politischen) Bildungssystemplanung dienen, die die Erhöhung des Leistungs- und Bildungsniveaus der Schülerinnen und Schüler ebenso wie die Erhöhung der Chancengleichheit und Geschlechtergerechtigkeit im Bildungswesen als Zielperspektiven ausgegeben hat (Klinglmair & Thaller, 2025).

3 Diskussion der zentralen Ergebnisse des vorliegenden Bands

3.1 Diskussion der im Band präsentierten Studien

Die Befundlage der im vorliegenden Band präsentierten Studien wird nachfolgend entlang der Mehrebenenlogik des Schulsystems zusammenfassend diskutiert. Für alle Ebenen gilt: Ein zentraler Mehrwert der in diesem Band präsentierten Studien liegt in der Tatsache, dass ihre Analysen auf mehreren Erhebungen der BIST-Ü basieren. Dieser Ansatz ermöglicht eine Überprüfung der Stabilität und Generalisierbarkeit der gefundenen Zusammenhänge über verschiedene Erhebungszeitpunkte hinweg. Während frühere Studien meist auf einer einzelnen Erhebung basierten, werden dort identifizierte Effekte häufig als Nachweis für die Wirksamkeit einer bestimmten Größe interpretiert – ohne zu prüfen, ob diese Effekte auch in anderen Kohorten oder unter veränderten Rahmenbedingungen bestehen bleiben. Die hier vorgestellten Analysen tragen somit wesentlich zur Validierung bestehender Be-

funde bei, indem sie nicht nur mögliche einmalige oder kontextabhängige Zusammenhänge hinterfragen, sondern auch langfristige Muster und Replikationsmöglichkeiten aufzeigen. Damit wird jedoch eine Herausforderung sichtbar: Die Längsschnittperspektive liefert nicht zwangsläufig grundlegend neue Erkenntnisse, sondern kann bestehende Annahmen bestätigen oder differenzieren. Dies könnte den Eindruck erwecken, dass die Längsschnittanalyse lediglich bereits bekannte Zusammenhänge mit neuem Datenmaterial repliziert. Entscheidend ist daher, dass in zukünftigen Studien gezielt weiterführende Fragestellungen entwickelt und innovative methodische Ansätze genutzt werden, um das Potenzial der Längsschnittdaten voll auszuschöpfen und über die bloße Bestätigung früherer Ergebnisse hinaus neue Erkenntnisse zu generieren. Nur so kann die Längsschnittforschung nicht nur zur Stabilitätsprüfung, sondern auch zur Weiterentwicklung bildungstheoretischer Modelle und bildungspolitischer Maßnahmen beitragen.

3.1.1 Systemebene

Über alle Studien/Kapitel hinweg zeigt sich, dass für die untersuchten Maßnahmen, die im Einflussbereich der Bildungspolitik, Schulleitung und Lehrkräfte liegen – z. B. Maßnahmen der Schulentwicklung (Groß Ophoff et al., Kapitel 7), Selektionskriterien für die Klassenzusammensetzung (Helm et al., Kapitel 8), Maßnahmen der Klassenreduktion (Bruneforth, Kapitel 3), die Bereitstellung von zusätzlichen Ressourcen wie Unterstützungspersonal, Förderunterricht oder Ganztagsbetreuung, die Gestaltung des Unterrichts bspw. durch Individualisierung und Differenzierung (Gargitter & Gniewosz, Kapitel 6) etc. –, die erwarteten positiven Effekte auf die Schulleistung *nicht* beobachtbar sind. Stattdessen zeigt sich wiederholt, dass die soziale Zusammensetzung der Schülerinnen- und Schülerklientel am Schulstandort die Schulleistung deutlich vorhersagt (Weber, Bacher & Bruneforth, Kapitel 2). Dieser Effekt scheint nicht nur stabil über die Zeit zu sein, sondern sogar leicht zugenommen zu haben.

Dies bedeutet, dass für eine Vielzahl von Maßnahmen (siehe auch die Evaluation zur Einführung der Mittelschule; Eder, Altrichter, Hofmann & Weber, 2015), die im unmittelbaren Einflussbereich schulischer Akteure (insbesondere der Bildungspolitik) liegen, kaum positive Effekte auf die Schulleistung (und auch auf die Abweichung vom Erwartungswert) in den untersuchten Zeiträumen zu erwarten sind. Dieser Befund kann allerdings nicht so interpretiert werden, dass Schulen keinen Einfluss auf die Kompetenzen der Schülerinnen und Schüler hätten. Die Daten der BIST-Ü zeigen nämlich sehr deutlich, dass (1) auf Systemebene massive Verbesserungen in den Leistungen der Schülerinnen und Schüler über die Zeit hinweg zu beobachten sind (siehe Abbildung 1) und dass (2) eine Vielzahl von Schulen beobachtbar sind, die bessere Leistungen erzielen, als aufgrund ihres Erwartungswerts anzunehmen wäre. Vielmehr legen die Nullbefunde der berichteten Studien nahe, dass es mit den bisherigen Analysen nicht gelungen ist, herauszuarbeiten, was den Unterschied zwischen diesen Schulen mit Leistungen über dem Erwartungswert und Schulen mit Leistungen darunter ausmacht.

Der Einfluss der Schul- und Klassenzusammensetzung auf die individuellen Schülerinnen- und Schülerleistungen spiegelt die stark ausgeprägte Bildungsungleichheit im öster-

reichischen Schulsystem wider (Weber, Bacher & Bruneforth, Kapitel 2; Gargitter & Gniewosz, Kapitel 6; Salchegger & Kampa, Kapitel 4). Dies ist ein Merkmal, das schon länger bekannt ist und sich in einer beständigen Diskussion über Maßnahmen im Umgang mit Bildungsungleichheit niedergeschlagen hat (z. B. Bruneforth, 2025). Im NBB 2015 (Biedermann, Weber, Herzog-Punzenberger & Nagel, 2016) werden vier solcher Maßnahmen vorgestellt: Erstens eine sozialindexbasierte Mittelzuweisung zur Unterstützung von Schul- und Unterrichtsqualität, zweitens Profilangebote an benachteiligten Schulstandorten, drittens Strategien für eine ausgewogene Schülerzusammensetzung (kontrollierte Elternwahl der Schule, Busing, Veränderung Schulsprengel, Planung Schulstandorte, Elterninformation) und viertens Wohnbaupolitik.

Darüber hinaus legen der starke Einfluss des elterlichen Bildungshintergrunds und der des familiären Kontexts nahe (Weber, Bacher & Bruneforth, Kapitel 2; Salchegger & Kampa, Kapitel 4; Gargitter & Gniewosz, Kapitel 6), dass Investitionen in die Elementarpädagogik und intensiver Ausbau der Quantität und Qualität der Kinderbetreuung nötig sind. Diese Investitionen sollten von Anreizen für Eltern, diese Angebote zu nutzen, und der Aufklärung der Eltern über ihre Rolle im und ihren Einfluss auf den Lernprozess ihrer Kinder flankiert werden. Schließlich sind Förderprogramme für Risikoschülerinnen und -schüler (z. B. qualitativer Ausbau der Sommerschule) zur Absicherung von Grundkompetenzen und Vorläuferfähigkeiten sowie die Etablierung von Bildungslandschaften weitere überprüfenswerte Möglichkeiten der Reduktion von Bildungsungleichheit.

3.1.2 Schulebene

Mehrere Beiträge in diesem Band fokussieren auf Merkmale der Schule als Prädiktoren für die Entwicklung der Schulleistung. So untersuchen Helm et al. (Kapitel 8) den Einfluss des Selektierens von Schülerinnen und Schülern nach bestimmten Kriterien in Klassen auf die Abweichung vom Erwartungswert der Schule. Groß Ophoff et al. (Kapitel 7) prüfen, ob die Anzahl der Maßnahmen der Qualitätssicherung und -entwicklung Einfluss auf die Steigerung der Leistung in den darauffolgenden Schuljahren hat. Schließlich gehen Gargitter & Gniewosz (Kapitel 6) der Frage nach, ob zusätzliche Ressourcen am Schulstandort dazu beitragen, dass Schulen Leistungen über dem Erwartungswert erbringen.

In allen drei Studien erwiesen sich die untersuchten Prädiktoren als nicht bedeutsam für die Vorhersage (der Entwicklung) der Schulleistung. Die – bereits oben berichteten – Nullbefunde werfen die Frage auf, was gute Schulen ausmacht, also Schulen, die vor dem Hintergrund ihrer Schulzusammensetzung (Komposition) besser als erwartet abschneiden. In diesem Zusammenhang verweist die Literatur zu School Effectiveness Research (Ditton & Müller, 2011; Scheerens, 1990) auf eine relativ kohärente Anzahl von Qualitätsmerkmalen guter Schulen, die künftige Analysen auf Basis der Bildungsstandarddaten stärker berücksichtigen sollten.

Allerdings müssen auch hier die Erwartungen niedrig gehalten werden; aus mehreren Gründen: Zum einen, weil gute Schulen (und deren Unterricht) sehr unterschiedlich

charakterisierbar sind (Ditton & Müller, 2011; Helmke, 2009). Zum anderen zeigen Studien, dass die Schulleistungen stark mit der Schulzusammensetzung (z. B. Schult, Mahler, Fauth & Lindner, 2022) sowie mit früheren Leistungen (z. B. Ditton & Müller, 2011) korrelieren, sodass fraglich ist, inwiefern über die Schulzusammensetzung und bisherigen Leistungen (bzw. Leistungen aus Zubringerschulen oder Einrichtungen der Elementarpädagogik bzw. des Elternhauses) hinaus die Schulleistungen durch andere Merkmale beeinflussbar sind.

Diese Ergebnisse stehen in einer gewissen Tradition der Erkenntnisse des Coleman-Reports (Coleman et al., 1966). Der Coleman-Report, eine der einflussreichsten Studien zur Bildungsgerechtigkeit und Schulqualität, zeigte auf, dass die Unterschiede in der Schulleistung primär durch die soziale Herkunft und Zusammensetzung der Schülerschaft erklärbar sind, während schulische Merkmale, wie zusätzliche Ressourcen oder andere strukturelle Faktoren, nur einen marginalen Beitrag leisten. Dies brachte die Frage auf, ob und wie Schulen überhaupt in der Lage sind, unabhängig von der sozialen Zusammensetzung ihrer Schülerschaft Leistungsunterschiede zu beeinflussen. Da – wie die hier präsentierten Studien zeigen – es Schulen gibt, deren Leistungen vom Erwartungswert abweichen, muss die Antwort „Ja“ sein. Offenbar ist es aber bisher in einem nur unzureichenden Ausmaß gelungen, die für diese Abweichungen ursächlichen Merkmale von Schule mit quantitativ-empirischen Untersuchungen offenzulegen.

3.1.3 Unterrichtsebene

Auf Ebene der Lehrkräfte bzw. des Unterrichts sieht die Befundlage nicht viel anders aus. Auch die von Schmich und Österbauer (Kapitel 10) durchgeführte Studie zur Frage des Einflusses der „Sozialform“/Unterrichtsform (Individualisierung, Binnendifferenzierung, Unterricht mit ganzer Klasse) auf die Schulleistung konnte keine Effekte zutage fördern (wobei allerdings nicht untersucht wurde, ob der Zusammenhang in Abhängigkeit der sozialen Zusammensetzung der Klasse variiert).

Schülerzentriertem Unterricht werden in der Literatur positive Effekte auf die Lernmotivation und den Lernerfolg nachgesagt. Allerdings geben Large-Scale-Untersuchungen in Deutschland und Österreich (z. B. Gruehn, 2000; Helm, Kemethofer, Moosbrugger, Bröderbauer & Luthe, 2016; Klieme & Kuger, 2016; Schreiner et al., 2019) Hinweise darauf, dass Merkmale fördernd-unterstützenden, schülerzentrierten Unterrichts eher von Schülerinnen und Schülern in leistungsschwächeren Klassen wahrgenommen werden. Eine mögliche Erklärung ist, dass „derartige Unterrichtsformen unter Betonung ihrer remedialen Funktion speziell in problematischen, lernschwachen Klassen zum Einsatz kommen“ (Clausen, 2002, S. 52).

Die Befunde aus den vorliegenden Studien – sowohl auf System-, Schul- als auch Unterrichtsebene – bestätigen somit erneut die zentrale Rolle der Schulkomposition und unterstreichen, dass Reformen oder Interventionen, die die soziale Herkunft der Schülerinnen und Schüler stärker berücksichtigen, zielführender sein könnten als solche, die sich ausschließlich auf schulinterne Prozesse konzentrieren.

3.2 Rolle externer Ereignisse im Schulsystem

Der lange Erhebungszeitraum der BIST-Ü von 10 Jahren wirft die Frage auf, ob die Befunde aus den regelmäßig durchgeführten BIST-Ü-Analysen durch externe Einflüsse im System, wie beispielsweise die Einführung der Neuen Mittelschule – und die damit verbundene Einführung des Teamteachings – oder die Migrationswelle im Jahr 2015 – und den damit verbundenen Anstieg der Zahl der Kinder mit nichtdeutscher Muttersprache in den Schulen – beeinflusst wurde. Weitere zeitlich parallel verlaufende Entwicklungen sind bspw. in der Digitalisierung des Bildungssystems, der Zunahme sozialer Medien und der Lehrerinnen- und Lehrerbildungsreform sowie der demografischen Entwicklung der Lehrerinnen und Lehrer (Stichwort: Lehrpersonenmangel) zu sehen. Auch strukturbezogene Reformen – wie das Schulautonomiegesetz und das Dienstrecht – könnten Einfluss auf die Entwicklung der Schulleistungen im Beobachtungszeitraum genommen haben.

Umfangreiche und belastbare Studien zu derartigen Entwicklungen im Bildungssystem liegen nur wenige vor. Als Beispiel für Studien zu bildungspolitischen Reformen kann die Evaluation der Einführung der Neuen Mittelschule genannt werden. Sowohl in der Evaluationsstudie 2015 (Eder et al., 2015) als auch in jener aus 2024 (Jesacher-Rößler & Kemethofer, 2024) konnten keine oder nur schwache heterogene Effekte der Einführung der Konzeptmerkmale auf die Unterrichtsqualität, die Schülerinnen- und Schüler-/Schulleistung und Bildungsungleichheit nachgewiesen werden. Auch die in mehreren Kapiteln dieses Bands (Gargitter & Gniewosz, Kapitel 6; Groß Ophoff et al., Kapitel 7; Helm et al., Kapitel 8) berichteten Nullbefunde erweisen sich über die Zeit hinweg als stabil. D. h., die nicht signifikanten Effekte der untersuchten Variablen gelten sowohl für frühere als auch spätere Erhebungen der BIST-Ü. Dies stützt die These, dass sich Veränderungen im System kaum auf die schulischen Leistungen und ihre Zusammenhänge mit anderen Schulmerkmalen niederschlagen. Hinzu kommt, dass sich über die Zeit hinweg eine hohe Stabilität des Einflusses von schwer durch Schule veränderbaren Faktoren wie der Bildungsnähe der Eltern (z. B. Bildungsabschluss der Eltern, Anzahl der Bücher zuhause) und der zuhause gesprochenen Sprache auf die Schulleistung zeigt (z. B. Weber, Bacher & Bruneforth, Kapitel 2).

Zusammenfassend lässt sich feststellen, dass die in diesem Band berichteten „Nullbefunde“ sowohl für frühere als auch für spätere BIST-Erhebungen Bestand haben. Gleichzeitig legen manche Befunde (z. B. Weber, Bacher & Bruneforth, Kapitel 2) nahe, dass externe Einflüsse wie die zunehmende Migration bestehende soziale und ethnische Disparitäten eher verstärkten als abschwächten. Diese Entwicklungen sollten im Kontext bildungspolitischer Maßnahmen zur Erhöhung von Chancengerechtigkeit weiter untersucht werden. Darüber hinaus ist einerseits die Forschung zu Merkmalen von Schule, die über dem Erwartungswert performen, zu forcieren und andererseits das Bildungsmonitoring so weiterzuentwickeln, dass Schulen und Lehrkräfte in ihren pädagogischen und (fach-)didaktischen Bemühungen unterstützt werden (z. B. durch eine stärkere Verzahnung mit formativer Individualdiagnostik; Förster, Weber, Forthmann, Helm & Kemethofer, 2023; Tränkmann, 2025; siehe dazu auch weiter unten).

3.3 Wie können die ausbleibenden Befunde erklärt werden?

Obwohl die berichteten Studien umfassend in Theorie und bestehender Forschungslage eingebettet sind, auf hochwertigen Datensätzen (z. B. Vollerhebung, echter Längsschnitt auf Schulebene) basieren und fortgeschrittene Auswertungsverfahren zum Einsatz kommen, lassen sich in den meisten Fällen die vermuteten Effekte nicht bestätigen. Für diesen doch überraschenden Umstand werden in diesem Abschnitt mögliche theoretische und forschungsmethodische Erklärungen gesucht.

3.3.1 Theoretische Perspektiven

Ein möglicher Erklärungsansatz für die ausbleibenden Befunde liegt in der theoretischen Perspektive, die häufig auf einem Modell technischer Rationalität beruht. Diese Sichtweise geht davon aus, dass sich komplexe Lehr-Lern-Situationen durch eine Reduktion auf einzelne Disziplinen wie Fachwissenschaft, Fachdidaktik, Pädagogik, Psychologie oder Wissenschaftstheorie hinreichend erfassen lassen. Auf dieser Grundlage werden generalisierte, in Forschungskontexten erzeugte Erkenntnisse (z. B. über Schulklima, Individualisierung) weitergegeben, wobei jedoch stillschweigend angenommen wird, dass Ziele und Bedingungen – etwa die Zusammensetzung einer Klasse – weitgehend konstant sind. Diese Annahme vernachlässigt jedoch die Tatsache, dass sich Lehr-Lern-Prozesse in der Praxis häufig nicht als Routinen vollziehen, sondern sich in einer hochkomplexen Situation abspielen, in der Lehrkräfte zeitgleich auf eine Vielzahl von Herausforderungen und Zielen reagieren müssen (vgl. Doyle, 1977). Infolgedessen lässt sich die vermutete Wirkung einzelner Faktoren in empirischen Untersuchungen oftmals nicht klar nachweisen.

Ein weiterer theoretischer Aspekt betrifft die Ebene der Schülerinnen und Schüler, die in den meisten Untersuchungen maßgeblich zur Varianz der Lernergebnisse beiträgt, wie unter anderem an den Intraklassenkorrelationen (ICC) ablesbar ist (z. B. Wurster & Feldhoff, 2019). Da auch für Österreich zu vermuten ist, dass nach Berücksichtigung des Schultyps und des Schulstandorts der Großteil der Leistungsunterschiede auf der individuellen Ebene zu verorten ist, sollten spezifische kognitive, motivationale, volitionale und emotionale Faktoren (z. B. aus dem Modell der individuellen Voraussetzungen erfolgreicher Schülerinnen und Schüler [INVO], vgl. Ausubel, Novak & Hanesian, 1981; Hasselhorn, 2017) künftig stärker als mediierende und moderierende Variablen berücksichtigt werden. Dies lässt sich damit begründen, dass Vorwissen, allgemeine kognitive Fähigkeiten sowie Lernstrategien zentrale Prädiktoren für den Lernerfolg sind.

Darüber hinaus zeigt die aktuelle Unterrichtsforschung, dass gängige Modelle der Unterrichtsqualität zunehmend kritisch diskutiert werden (vgl. Begrich et al., 2023; Praetorius, Klieme, Herbert & Pinger, 2018). Ähnlich verhalten sich die Befunde im Bereich des Lehrkräftewissens oder des fachfremden Unterrichts: Mit Ausnahme einiger weniger Studien (z. B. COACTIV) ist es bislang kaum gelungen, einen deutlichen und konsistenten Einfluss dieser Faktoren auf die Unterrichtsqualität oder das Lernen der Schülerinnen und Schüler nachzuweisen. Aktuelle Entwicklungen wie die steigende Anzahl von Schülerinnen

und Schülern ohne Deutschkenntnisse, der Anstieg fehlender Vorläuferfähigkeiten der Schülerinnen und Schüler sowie der Einsatz digitaler Medien führen zu einer Rollenverschiebung der Lehrkraft. Die Vermittlungsfunktion tritt zurück, während der erzieherische und pädagogische Auftrag an Bedeutung gewinnt. Dadurch rückt die sozioemotionale Unterstützung stärker in den Fokus, wie in der Selbstbestimmungstheorie (Deci & Ryan, 1993) beschrieben. Dies erfordert eine Reflexion und mögliche Neuausrichtung bestehender Qualitätsmodelle hin zu mehr pädagogischem Handeln und Beziehungsgestaltung.

3.3.2 Forschungsmethodische Perspektiven

Neben theoretischen Aspekten können die ausbleibenden Effekte auch auf methodische Limitationen zurückzuführen sein.

Ein zentrales methodisches Problem besteht darin, dass viele Studien einen *distalen Zusammenhang* zwischen den erfassten unabhängigen Variablen (z. B. Schul- und Klassenzusammensetzung) und den abhängigen Variablen (z. B. fachliche Leistungen der Schülerinnen und Schüler) untersuchen. Die Wirkungskette erstreckt sich dabei über mehrere Glieder dieser Kette, angefangen von den Kontextmerkmalen (z. B. Zusammensetzung der Klasse, Region, Schultyp) über das Verhalten der Lehrpersonen bis hin zum Verhalten der Schülerinnen und Schüler und letztlich zur Schülerinnen- und Schülerleistung. Aufgrund dieser langen Wirkungskette besteht ein erhebliches Risiko, dass mediiierende Prozesse nicht hinreichend erfasst werden. Eine Folge dieser komplexen Zusammenhänge ist, dass einzelne Variablen (z. B. Schultyp, soziale Herkunft) potenziell durch diverse unbeobachtete mediiierende Prozesse (z. B. individuelle Lernaktivitäten, Klassenklima) unterschätzt werden. Um künftige Befunde zu präzisieren, ist daher eine stärkere Berücksichtigung mediiierender und moderierender Variablen notwendig, etwa hinsichtlich des Verhaltens der Lehrkräfte und des tatsächlichen Nutzungsverhaltens der Schülerinnen und Schüler im Unterricht. Darüber hinaus prüfen bisherige Studien kaum, inwieweit kumulative Effekte einer Vielzahl kleiner Einflussfaktoren die Leistungsentwicklung determinieren können. Hier könnten Profilanalysen Abhilfe schaffen.

Nicht nur die in diesem Band präsentierten Studien, sondern auch die Mehrheit der quantitativ-empirischen Schul- und Unterrichtsforschung stützt sich auf Fragebogenstudien, die sich zur Erfassung komplexer Konstrukte (z. B. Unterrichtsqualität) nur bedingt eignen. Insbesondere die Reliabilität und Validität einzelner Skalen oder gar Single-Item-Analysen (vgl. Gargitter & Gniewosz, Kapitel 6; Groß Ophoff et al., Kapitel 7) sind bei der Untersuchung vielschichtiger Konstrukte kritisch zu hinterfragen. Fragebögen erfassen zudem häufig nur subjektive Einschätzungen und sind anfällig für soziale Erwünschtheit, was zu Verzerrungen führen kann. Folglich bedarf es einer *methodischen Erweiterung* hin zu komplexeren *Mixed-Methods-Designs*, um die vielschichtigen Konstrukte angemessen zu untersuchen (Hagenauer, Gegenfurtner & Gläser-Zikuda, 2023). So könnten beispielsweise qualitative Interviews oder Beobachtungs- und Videostudien die Analyse tiefergehender Lehr-Lern-Prozesse ergänzen.

Methodisch bedeutsam ist außerdem das Ausmaß, in dem die Schülerinnen und Schüler bei Leistungserhebungen motiviert sind, ihre tatsächlichen Fähigkeiten zu zeigen. Ent-

sprechende Informationen zur Testmotivation wurden im Rahmen der BIST-Ü miterfasst und sollten künftig stärker in vertiefenden Analysen berücksichtigt werden.

Die hier diskutierten Studien unterstreichen die Notwendigkeit, sowohl die Komplexität schulischer und unterrichtlicher Prozesse zu berücksichtigen als auch differenzielle oder moderatorengeleitete Modelle zu nutzen. Künftige Forschung sollte neben klassischen querschnittlichen oder längsschnittlichen Fragebogenerhebungen verstärkt auf Methoden zurückgreifen, die die sozialen und kognitiven Prozesse im Unterricht unmittelbar erfassen (z. B. Videoanalysen, Lerntagebücher, Interventionsstudien). Dabei ist besonders zu beachten, dass das Einbeziehen von Mediatoren (z. B. Verhalten der Lehrpersonen) und Moderatoren (z. B. Schulart, Sprachkompetenz) die Effektstärken einzelner Prädiktoren erheblich verändern kann. Erst durch eine umfassende Betrachtung der Prozesshaftigkeit (z. B. Lehr-Lern-Interaktionen, Feedbackzyklen, Differenzierung im Unterricht) und der kontextuellen Rahmenbedingungen (z. B. Schulstandort, Lernvoraussetzungen der Schülerinnen und Schüler) lassen sich präzisere Aussagen über die Wirkmechanismen schulischer und unterrichtlicher Merkmale treffen. Die vorliegenden Befunde, wonach die meisten dieser Merkmale nur geringe Effekte auf das Lernen zeigen, dürfen also nicht vorschnell als Beleg für eine Nichtexistenz von Einflüssen interpretiert werden. Vielmehr verdeutlichen sie den Bedarf an methodischen Weiterentwicklungen, differenzierten Modellen und prozessorientierten Forschungsperspektiven.

Insgesamt machen diese Befunde deutlich, dass sowohl konzeptionell-theoretische als auch methodische Nuancierungen für zukünftige Untersuchungen unerlässlich sind. Die hier vorgestellten Ergebnisse dienen als Ausgangspunkt, um in Folgestudien die Wirksamkeit unterschiedlichster Schul- und Unterrichtsmerkmale im Zusammenspiel mit den komplexen Lernprozessen auf Ebene der Schülerinnen und Schüler weiter zu erforschen und damit zu einem tieferen Verständnis erfolgreicher Lernumgebungen beizutragen.

4 Ausblick

Was lässt sich aus den im vorangegangenen Kapitel beschriebenen Erkenntnissen für künftige Erhebungen ableiten? Zukünftige Erhebungen im Rahmen von iKM^{PLUS} sollten nicht nur Schulleistungen, sondern verstärkt die Abweichung vom Erwartungswert analysieren, da diese den schulischen Einfluss besser abbildet. Besonders relevant wäre die Untersuchung von Schulen mit überdurchschnittlicher Performance durch innovative Analyseansätze wie die Data Envelopment Analysis (DEA). Neben kognitiven sollten auch nonkognitive Zielgrößen, etwa Lernfreude oder lebenslanges Lernen, stärker berücksichtigt werden, da sie ebenfalls maßgeblich von schulischen Akteuren beeinflusst werden.

Hinsichtlich der Einflussgrößen sollten kontextbezogene Daten gezielter erhoben werden, um lernförderliche Aspekte im schulischen Einflussbereich besser zu identifizieren. Die bisherigen Instrumente zeigen geringe prädiktive Kraft für fachliches Lernen, sodass ihre Validität kritisch hinterfragt werden sollte. Zudem müssten zukünftige Studien unter-

suchen, welche Kontextvariablen – beispielsweise aus der School Effectiveness Research oder Lernmotivationstheorien – relevante Effekte auf schulische Leistung und nonkognitive Outcomes haben.

Ableitungen für künftige Erhebungen und Studien:

1. Fokus auf die Abweichung vom Erwartungswert anstelle reiner Schulleistungen.
2. Berücksichtigung nonkognitiver Zielgrößen (bspw. Motivation) als gleichwertige Bildungsziele.
3. Nutzung neuer Analyseverfahren zur Identifikation leistungsstarker Schulen.
4. Kritische Prüfung der Validität bestehender Kontextinstrumente.
5. Erhebung qualitativ aussagekräftiger Kontextvariablen mit erwartbaren Effekten.

Ein weiterer wichtiger Aspekt, der in zukünftigen Studien verstärkt berücksichtigt werden sollte, betrifft die forschungsmethodische Herangehensweise selbst. Es ist essenziell, nicht nur bestehende Analysen zu replizieren, sondern gezielt neue Fragestellungen zu entwickeln, die das Erkenntnispotenzial vorhandener Daten bestmöglich ausschöpfen. Die Vielzahl an Large-Scale Assessments, an denen Österreich teilnimmt, bietet eine wertvolle Grundlage für übergreifende Metaanalysen, die bisher zu wenig genutzt werden. Ein Beispiel hierfür ist die Untersuchung von Geschlechterunterschieden: Statt diese isoliert innerhalb einzelner Studien zu betrachten, wäre eine systematische Synthese der Befunde über verschiedene Erhebungen hinweg sinnvoll, um konsistente Muster und Kontextfaktoren zu identifizieren. Solche integrativen Analysen könnten dazu beitragen, zentrale Mechanismen der Bildungsungleichheit besser zu verstehen und evidenzbasierte Empfehlungen für die Bildungspolitik abzuleiten.

Trotz der beschriebenen Grenzen der vorliegenden Datensätze stellen die BIST-Ü in Österreich einen Meilenstein in der nationalen Bildungsforschung dar. Sie bieten durch ihre Voll-datensätze sowie die elaborierten Kompetenzmessungen eine Datenbasis, die von einzelnen Forscherteams an Universitäten und Hochschulen in diesem Umfang und dieser Qualität nicht leistbar wäre. Mit der Einführung von iKM^{PLUS} wird das Potenzial für die Bildungsforschung weiter gesteigert, da künftig auch längsschnittliche Erhebungen auf Schülerinnen- und Schülerebene sowie auf Unterrichtsebene möglich sind, was insbesondere für die Analyse von Entwicklungsprozessen und die Untersuchung von Unterrichtswirkungen neue Perspektiven eröffnet. Voraussetzung hierfür ist, dass die Vielfalt an Kontextvariablen sowie der Zugang für Bildungsforscherinnen und -forscher, unterstützt durch die professionelle Datenaufbereitung, -bereitstellung und -auswertung durch das IQS, weiter erhalten bleibt. Denn diese schaffen ideale Voraussetzungen für innovative und fundierte Forschung. Die im vorliegenden Band präsentierten Studien verdeutlichen dies eindrucksvoll.

Die Nutzung von Bildungsmonitoring-Daten birgt großes Potenzial, um die Qualität von Unterricht und schulischen Maßnahmen gezielt weiterzuentwickeln. Allerdings zeigt sich, dass das Bildungsmonitoring „den Blick nicht auf das einzelne Kind in seiner Individualität“ (Tränkmann, 2025, S. 403) richtet und „bei den Bildungsmonitoringdiskursen zu

den Lernstandserhebungen immer noch eher getrennt über Datenbereitstellung einerseits und Datenrezeption andererseits gesprochen wird, diese also nicht verzahnt entwickelt werden“ (ebd.). Genau hier setzt iKM^{PLUS} an, indem es Bildungsmonitoring stärker mit Förderdiagnostik zu verknüpfen und Lehrkräfte als zentrale Akteursgruppe intensiver einzubeziehen versucht. iKM^{PLUS} zielt darauf ab, Diagnostik und Monitoring kohärenter zusammenzuführen und schulische Entwicklungsprozesse auf datenbasierter Grundlage zu unterstützen. Die Zielerreichung wird davon abhängen, ob es gelingt, die Rückmeldung von Befunden in pädagogische und (fach-)didaktische Handlungszusammenhänge einzubetten und die relevanten Akteure unterschiedlicher Professionen und Expertise einzubeziehen. Dies vorausgesetzt, könnte Datennutzung nicht nur der Statusdiagnose dienen, sondern aktiv zur Verbesserung des Bildungssystems beitragen.

Literatur

- Altrichter, H. & Gamsjäger, M. (2017). A conceptual model for research in performance standard policies. *Nordic Journal of Studies in Education Policy*, 3 (1), 6–20. doi:10.1080/20020317.2017.1316180
- Ausubel, D. P., Novak, J. D. & Hanesian, H. (1981). *Psychologie des Unterrichts*. Beltz.
- Begrich, L., Praetorius, A.-K., Decristan, J., Fauth, B., Göllner, R., Herrmann, C., Klein-knecht, M., Taut, S. & Kunte, M. (2023). Was tun? Perspektiven für eine Unterrichtsqualitätsforschung der Zukunft. *Unterrichtswissenschaft*, 51, 63–97. doi:10.1007/s42010-023-00163-4
- Biedermann, H., Weber, C., Herzog-Punzenberger, B. & Nagel, A. (2016). Auf die Mitschüler/innen kommt es an? Schulische Segregation – Effekte der Schul- und Klassenzusammensetzung in der Primarstufe und der Sekundarstufe. In M. Bruneforth, F. Eder, K. Krainer, C. Schreiner, A. Seel & C. Spiel (Hrsg.), *Nationaler Bildungsbericht Österreich 2015, Band 2 – Fokussierte Analysen bildungspolitischer Schwerpunktthemen*. Graz: Leykam. doi:10.17888/nbb2015-2-4
- Bischof, L. M., Hochweber, J., Hartig, J. & Klieme, E. (2013). Schulentwicklung im Verlauf eines Jahrzehnts. Erste Ergebnisse des PISA-Schulpanels. In N. Jude & E. Klieme (Hrsg.), *PISA 2009 – Impulse für die Schul- und Unterrichtsforschung* (Zeitschrift für Pädagogik, 59. Beiheft, S. 172–199). Weinheim: Beltz Pädagogik.
- Bos, W., Gebauer, M. M. & Postlethwaite, T. N. (2018). Möglichkeiten, Grenzen und Perspektiven internationaler Schulleistungsforschung. In R. Tippelt & B. Schmidt-Hertha (Hrsg.), *Handbuch Bildungsforschung* (S. 325–348). Wiesbaden: Springer.
- Bruneforth, M. (2025, in Vorbereitung). Diversität und soziale Ungleichheit in den Grundkompetenzen. Ein empirischer Blick auf Vielfalt und Ungleichheiten im österreichischen Schulsystem. *SchulVerwaltung aktuell*, 13 (2).

- Bruneforth, M. & Diedrich, M. (2024). Was sind eigentlich Daten/Evidenzen im Schulsystem? *Erziehung und Unterricht*, 174 (5–6), 412–420.
- Clausen, M. (2002). *Unterrichtsqualität: eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstrukt- und Kriteriumsvalidität. Pädagogische Psychologie und Entwicklungspsychologie*. Waxmann.
- Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfield, F. D. & York, R. L. (1966). *Equality of educational opportunity*. US Government Printing Office.
- Deci, E. L. & Ryan, R. M. (1993). Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. *Zeitschrift für Pädagogik*, 39 (2), 223–238.
- Ditton, H. & Müller, A. (2011). Schulqualität. In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz (Hrsg.), *Lehrbuch. Empirische Bildungsforschung. Gegenstandsbereiche* (S. 99–111). VS Verl. für Sozialwiss. doi:10.1007/978-3-531-93021-3_9
- Doyle, W. (1977). Learning the Classroom Environment: An Ecological Analysis. *Journal of Teacher Education*, 28 (6), 51–55. doi:10.1177/002248717702800616
- Eder, F., Altrichter, H., Hofmann, F. & Weber, C. (Hrsg.). (2015). *Evaluation der Neuen Mittelschule (NMS). Befunde aus den Anfangskohorten*. Graz: Leykam.
- Förster, N., Weber, C., Forthmann, B., Helm, C. & Kemethofer, D. (2023). Kompensation von pandemiebedingten Lernlücken: Welche Rolle sollte die Diagnostik spielen? *Zeitschrift für Bildungsforschung*, 13 (2), 285–296. doi:10.1007/s35834-023-00395-5
- George, A. C. & Robitzsch, A. (2014). „Fairer Vergleich“ in der BIST-Ü-M4 2013. *Technische Dokumentation*. Salzburg.
- George, A. C., Schreiner, C., Wiesner, C., Pointinger, M. & Pacher, K. (Hrsg.). (2019). *Fünf Jahre flächendeckende Bildungsstandardüberprüfungen in Österreich. Vertiefende Analysen zum Zyklus 2012 bis 2016* (Kompetenzmessungen im österreichischen Schulsystem: Analysen, Methoden & Perspektiven). Münster, New York: Waxmann. doi:10.31244/9783830995586
- Gruehn, S. (2000). *Unterricht und schulisches Lernen. Schüler als Quellen der Unterrichtsbeschreibung*. Waxmann.
- Hagenauer, G., Gegenfurtner, A. & Gläser-Zikuda, M. (2023). *Grundlagen und Anwendung von Mixed Methods in der empirischen Bildungsforschung*. Wiesbaden: Springer VS.
- Hasselhorn, M. (2017). Was sind aus psychologischer Perspektive die individuellen Voraussetzungen gelingender Lern- und Bildungsprozesse? In N. McElvany, W. Bos, H. G. Holtappels, J. Hasselhorn & A. Ohle (Hrsg.), *Bedingungen gelingender Lern- und Bildungsprozesse: aktuelle Befunde und Perspektiven für die empirische Bildungsforschung* (S. 11–29). Waxmann.
- Helm, C., Kemethofer, D., Moosbrugger, R., Bröderbauer, S. & Luthe, S. (2016). Effekte der aus Schulleitersicht wahrgenommenen Belastungen des schulischen Lernens auf die Unterrichtsgestaltung und die Schülerkompetenzen in Mathematik sowie auf die Schulzufriedenheit. *Zeitschrift für Bildungsforschung*, 6 (3), 245–264. doi:10.1007/s35834-016-0165-0
- Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts*. Klett Kallmeyer.
- Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen [IQS] (2025). *Webpage Bildungsstandards und kompetenzorientierter Unterricht*. <https://www.iqs>.

- gv.at/themen/nationale-kompetenzerhebung/grundlagen-der-nationalen-kompetenzerhebung/bildungsstandards-und-kompetenzorientierter-unterricht [8.1.2025]
- Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen [IQS] (o. J.). *Erhebungsmaterialien und Fragebögen der Standardüberprüfungen*. <https://www.iqs.gv.at/downloads/archiv-des-bifie/bildungsstandardueberpruefungen/erhebungsmaterialien-und-frageboegen> [9.1.2025]
- Jesacher-Roessler, L. & Kemethofer, D. (Hrsg.). (2024). *10 Jahre Regelschule – die (Neue) Mittelschule*. Waxmann. doi:10.31244/9783830996392
- Klieme, E. & Kuger, S. (2016). *Umgang mit Heterogenität an Schulen des Sekundarbereichs im internationalen Vergleich*. GEBF, Berlin. http://www.gebf2016.de/aktuelles/Symposien_Abstracts_240216.pdf
- Klinglmair, R. & Thaller, A. (2025). (Gesetzliche) Rahmenbedingungen sowie ausgewählte Maßnahmen und Instrumente einer evidenzinformatierten Schul- und Unterrichtsentwicklung. In A. Fritz, M. Grillitsch, F. Kerschbaumer & L. Wölbitsch (Hrsg.), *Qualität in Schule und Unterricht Perspektiven – Spannungsfelder – Instrumente* (S. 79–88). Münster: Waxmann.
- Mullis, I. V. S. & Martin, M. O. (2022). IEA's TIMSS and PIRLS: Measuring Long-Term Trends in Student Achievement. In T. Nilsen, A. Stancel-Piątak & J.-E. Gustafsson (Hrsg.), *International Handbook of Comparative Large-Scale Studies in Education*. Springer International Handbooks of Education. Springer, Cham. doi:10.1007/978-3-030-88178-8_15
- Praetorius, A.-K., Klieme, E., Herbert, B. & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of Three Basic Dimensions. *ZDM*, 50 (3), 407–426. doi:10.1007/s11858-018-0918-4
- Scheerens, J. (1990). School Effectiveness Research and the Development of Process Indicators of School Functioning. *School Effectiveness and School Improvement*, 1 (1), 61–80. doi:10.1080/0924345900010106
- Schildkamp, K. (2019). Data-based decision-making for school improvement. Research insights and gaps. *Educational Research*, 61 (3), 257–273. doi:10.1080/00131881.2019.1625716
- Schmid, C., Trendtel, M., Bruneforth, M. & Hartig, J. (2019). Effectiveness of a governmental action to improve Austrian primary schools – results of multilevel analyses based on repeated cycles of educational standards assessments. *School Effectiveness and School Improvement*, 31 (2), 149–171. doi:10.1080/09243453.2019.1620294
- Schreiner, C., Breit, S., Pointinger, M., Pacher, K., Neubacher, M. & Wiesner, C. (Hrsg.). (2018). *Standardüberprüfung 2017 Mathematik, 8. Schulstufe – Bundesergebnisbericht*. Salzburg.
- Schreiner, C. & Helm, C. (2024). 10 Jahre NMS – Inwiefern hat sich der Unterricht aus Perspektive der Schüler*innen verändert? In L. Jesacher-Rößler & D. Kemethofer (Hrsg.), *10 Jahre Regelschule – die (Neue) Mittelschule* (S. 132–148). Waxmann. doi:10.31244/9783830996392
- Schreiner, C., Wiesner, C., Kemethofer, D., Helm, C., Freller-Töglhofer, M., Illetschko, M. & Kiefer, T. (2019). *Zyklusbericht zur Bildungsstandarderhebung 2017 – Charakterisierung des fachlichen Unterrichts*. Salzburg.

- Schult, J., Mahler, N., Fauth, B. & Lindner, M. A. (2022). Did students learn less during the COVID-19 pandemic? Reading and mathematics competencies before and after the first pandemic wave. *School Effectiveness and School Improvement*, 1–20. doi:10.1080/09243453.2022.2061014
- Stauber, A., Stahl, J., Bruneforth, M. & Illetschko, M. (2024). Die individuelle Kompetenzmessung PLUS (iKM^{PLUS}) – pädagogische Diagnostik im Spannungsfeld von Individualförderung und Systemmonitoring. In G. Kraler, G. Schauer & C. Schreiner (Hrsg.), *Pädagogische Diagnostik und Lehrer*innenbildung* (S. 157–172). Innsbruck: Klinkhardt.
- Süss-Stepancik, E., Ober, M., George, A. C. & Varelija-Gerber, A. (2022). Veränderungen im mathematischen Leistungsspektrum einer Kohorte von der Primar- zur Sekundarstufe. In A. C. George, S. Götz, M. Illetschko & E. Süss-Stepancik (Hrsg.), *Empirische Befunde zu Kompetenzen im Mathematikunterricht der Sekundarstufe 1 und Folgerungen für die Praxis* (S. 113–132). Münster: Waxmann. doi:10.31244/9783830995586
- Tränkmann, J. (2025). Bildungsmonitoring – ein Kommentar zum Sammelband. In A. Albers & N. Jude (Hrsg.), *Blickpunkt Bildungsmonitoring – Bilanzen und Perspektiven* (S. 397–407). Weinheim, Basel: Beltz Juventa.
- Wurster, S. & Feldhoff, T. (2019). Schul- und Unterrichtsqualität aus der Mehrebenenperspektive. Ist die Schule oder die Klasse die relevante pädagogische Gestaltungseinheit? *Zeitschrift für Pädagogik*, 65 (1), 24–39. doi:10.25656/01:23925

Systemebene

Änderungen in der sozialen und ethnischen Schulzusammensetzung im Zeitverlauf

1 Einleitung

In den letzten Jahrzehnten war die österreichische Population der Schülerinnen und Schüler einem stetigen Wandel unterworfen. Dieser Wandel macht sich auch in den Anteilen an Schülerinnen und Schülern unterschiedlicher sozioökonomischer und ethnischer Herkunft bemerkbar. Insbesondere zeigt sich, dass Ungleichheitsmerkmale, die in gut dokumentiertem Zusammenhang mit Schulleistung stehen und in der Vergangenheit zur Beschreibung der sozialen Benachteiligung von Schulen vorgeschlagen wurden (Bruneforth, Weber & Bacher, 2012), durch eine differenzielle Entwicklung gekennzeichnet sind. So zeigt sich auf Basis der Bildungsstandardüberprüfungen (Schreiner & Breit, 2012; Breit, Bruneforth & Schreiner, 2016, 2017; Schreiner, Breit, Pointinger, Pacher, Neubacher & Wiesner, 2018; BIFIE, 2019, 2020), dass in den 2010er-Jahren der Anteil der Schülerinnen und Schüler mit Migrationshintergrund – die im Vergleich zu Schülerinnen und Schülern ohne Migrationshintergrund im Schnitt geringere Leistungen erzielen – zugenommen hat. Konkret konnte zwischen den ersten Bildungsstandardüberprüfungen 2012 (8. Schulstufe) bzw. 2013 (4. Schulstufe) und den letzten Testungen 2018 (4. Schulstufe) bzw. 2019 (8. Schulstufe) ein Zuwachs von 4 Prozentpunkten (von 19 % auf 23 %) auf der 4. Schulstufe und von 5 Prozentpunkten (von 18 % auf 23 %) auf der 8. Schulstufe verzeichnet werden. Gleichzeitig nahm jedoch der Anteil von Kindern zu, deren Eltern eine Matura oder eine tertiäre Ausbildung absolviert haben (von 45 % auf 51 % auf der 4. Schulstufe und von 47 % auf 59 % auf der 8. Schulstufe) und die im Schnitt bessere Leistungen erzielen als Schülerinnen und Schüler mit formal geringer gebildeten Eltern.

Ungleichheitsmerkmale wie der Migrationshintergrund oder die Bildung der Eltern werden in der Forschung jedoch nicht nur als Individualmerkmal in den Fokus genommen, sondern auch im Hinblick auf die Zusammensetzung von Klassen oder Schulen (Bacher & Weber, 2018; Biedermann, Weber, Herzog-Punzenberger & Nagel, 2016; Dumont, Neumann, Maaz & Trautwein, 2013). Konkret steht in dieser Forschungstradition die Frage im Zentrum, ob Schülerinnen und Schüler eine unterschiedliche Leistungsentwicklung zeigen, je nachdem, mit „wem“ sie in die Klasse/Schule gehen. Angenommen wird, dass mit unterschiedlichen Zusammensetzungen der Schülerschaft differenzielle Lern- und Entwicklungsmilieus verbunden sind, die sich auf das Lernen und die Leistungsentwicklung der Schülerinnen

und Schüler niederschlagen (Baumert, Stanat & Watermann, 2006). Günstige Lern- und Entwicklungsmilieus wären etwa durch einen hohen Anteil an Akademikerkindern und durch einen geringen Anteil an Schülerinnen und Schülern mit Migrationshintergrund gekennzeichnet, wogegen eine Häufung von Schülerinnen und Schülern, deren Eltern formal geringere Schulabschlüsse aufweisen, bzw. von Schülerinnen und Schülern mit Migrationshintergrund als weniger lernförderlich erachtet wird.

Unterschiede in der sozioökonomischen und ethnischen Schul- und Klassenzusammensetzung und somit die Determinanten differenzieller Lern- und Entwicklungsniveaus sind eng mit schulischer Segregation – also der Ungleichverteilung von Schülerinnen und Schülern mit bestimmten Merkmalen auf Schulen/Klassen – verbunden (Weber, Moosbrugger, Hasengruber, Altrichter & Schrod, 2019).

Die damit verbundene Problematik hat sich in jüngster Zeit durch den Familiennachzug von Kindern mit Fluchthintergrund verschärft. Medial und politisch werden Befürchtungen geäußert, dass die damit verbundenen schulischen Herausforderungen (vor allem fehlende Räumlichkeiten und fehlende Lehrkräfte) nicht mehr bewältigt werden können. So z. B. betiteln Neuhold und Peterlik im Profil am 24.3.2024 ihren Beitrag mit „Jetzt kommen die syrischen Kinder nach: Schafft Wien das noch?“. Ähnlich lesen sich die Berichte anderer Medien. Vonseiten der Wiener Stadtpolitik werden daher u. a. mehr finanzielle Bundesmittel und eine gleichmäßigere Verteilung der Geflüchteten auf Österreich gefordert.¹ Damit wird Bezug auf mögliche Maßnahmen genommen, um den Effekt der sozialen Zusammensetzung von Schulen auf den Lern- und Schulerfolg der Kinder abzuschwächen, nämlich einer Desegregation und einer ausreichenden Ressourcenausstattung. Zur Erreichung einer Desegregation sind unterschiedliche Zugänge möglich, wie z. B. eingeschränkte Schulwahl, Quotenregelung, Busing, also die Re-Allokation von Schülerinnen und Schülern, Gutscheinsystem, Änderung der Schulsprengel und/oder Wohnpflicht/Residenzpflicht (siehe dazu u. a. Biedermann et al., 2016). Für eine ausreichende Ressourcenausstattung könnte eine sozial indizierte Mittelvergabe beitragen, deren Verwendung an Kriterien gebunden wird, die sich in der Literatur zur Schuleffektivitätsforschung als erfolgreich erwiesen haben (Bacher, 2015a, b).

Vor dem Hintergrund des eingangs beschriebenen Wandels der Population der Schülerinnen und Schüler geht der vorliegende Beitrag auf Basis der Stichprobe der Baseline-Erhebungen der Bildungsstandardüberprüfung (Primarstufe N = 267; Sekundarstufe N = 204 Schulen) unter Einbeziehung aller Bildungsstandardüberprüfungen (Primarstufe: Baseline 2010, Mathematik 2013, Deutsch 2015, Mathematik 2018. Sekundarstufe: Baseline 2009, Mathematik 2012, Englisch 2013, Deutsch 2016, Mathematik 2017, Englisch 2019) u. a. der Frage nach, ob und wie sich die Änderung der Population in der Änderung der sozialen und ethnischen Zusammensetzung der Schulen niederschlug und ob differenzielle Änderungen

1 Siehe dazu etwa folgende Medienbeiträge: <https://www.derstandard.at/story/3000000216540/wiens-stadtchef-ludwig-ich-erwarte-mir-vom-bund-solidaritaet>, <https://wien.orf.at/stories/3254370/> und <https://www.die-presse.com/18390928/lesen-sie-ruhig-weiter-hier-passiert-nichts>.

der Schulzusammensetzung mit Unterschieden in Änderungen der Leistungen auf Schulebene verbunden waren.

2 Soziale und ethnische Bildungsungleichheiten

Das nationale Bildungsmonitoring sowie internationale Large-Scale-Vergleichsstudien dokumentieren für Österreich wiederholt einen deutlichen Zusammenhang zwischen dem familiären Hintergrund und den Schulleistungen (siehe u. a. Bruneforth et al., 2012; Neubacher & Wimmer, 2021; OECD, 2012; Toferer, Lang & Salchegger, 2023). Bei der jüngsten PISA-Studie zeigte sich etwa, dass Schülerinnen und Schüler aus Familien mit geringem Sozialstatus² zwischen 109 Punkten (Mathematik) und 129 Punkten (Naturwissenschaften) hinter Gleichaltrigen mit hohem Sozialstatus zurückliegen. Bedient man sich der groben Daumenregel, wonach 40 Punkte auf der PISA-Skala etwa einem Schuljahr entsprechen (OECD, 2011), wird die Bedeutung dieser Unterschiede noch einmal verdeutlicht. Folglich lägen Schülerinnen und Schüler mit geringem Sozialstatus gegenüber den privilegiierteren Peers um rund 2,5 bzw. 3 Jahre in der Leistung zurück,³ wobei die Differenzen in Jahren möglicherweise insbesondere in den Naturwissenschaften noch größer ausfallen, wie eine Studie von Wendt, Kasper, Bos, Vennemann und Goy (2017) nahelegt. Ähnlich deutliche Unterschiede finden sich auch auf Basis der letzten Bildungsstandardüberprüfungen im Hinblick auf die elterliche Bildung. Die Unterschiede zwischen Schülerinnen bzw. Schülern, deren Eltern maximal die Pflichtschule abgeschlossen haben, und Schülerinnen bzw. Schülern mit tertiär gebildeten Eltern betragen in Mathematik 101 (8. Schulstufe; Schreiner et al., 2018) bzw. 119 Punkte (4. Schulstufe; BIFIE, 2019), in Lesen 113 (8. Schulstufe; Breit et al., 2017) bzw. 126 Punkte (4. Schulstufe; Breit et al., 2016) und in Englisch 79 Punkte (8. Schulstufe, BIFIE, 2020). Auch im Hinblick auf migrationsbezogene Merkmale werden deutliche Unterschiede sichtbar. PISA 2022 zeigt Unterschiede von 58 Punkten (Mathematik) bis 78 Punkte (Naturwissenschaften) zwischen 15-jährigen Schülerinnen und Schülern mit und ohne Migrationshintergrund (Toferer et al., 2023). Die Bildungsstandardüberprüfungen weisen ebenso auf deutliche Unterschiede zwischen Schülerinnen und Schülern mit und ohne Migrationshintergrund (Lesen 4. Schulstufe: 61 Punkte; Lesen 8. Schulstufe: 75 Punkte; Mathematik 4. Schulstufe: 61 Punkte; Mathematik 8. Schulstufe: 64 Punkte; Englisch Lesen 8. Schulstufe: 28 Punkte) und auch Schülerinnen und Schüler mit Deutsch als Erstsprache und Schülerinnen und Schüler mit einer anderen Erstsprache hin (Lesen 4. Schulstufe: 72 Punkte; Lesen 8. Schulstufe: 81 Punkte; Mathematik 4. Schulstufe: 64 Punkte; Mathematik 8. Schulstufe: 69 Punkte; Englisch Lesen 8. Schulstufe: 39 Punkte; Breit et al., 2016, 2017; Schreiner et al., 2018; BIFIE, 2019, 2020). Im Hinblick auf die berichteten Differenzen ist

- 2 In PISA2022 wird der soziale Status durch einen Index erfasst, in den als Indikatoren einfließen: die Bildung der Eltern, der Beruf der Eltern und die Besitztümer, die den materiellen Wohlstand und das kulturelle Kapital abbilden (Toferer et al., 2022, S. 72). Das unterste Quartil des Index wird als geringer bzw. unterdurchschnittlicher Sozialstatus und das oberste Quartil wird als hoher bzw. überdurchschnittlicher Sozialstatus bezeichnet.
- 3 In der jüngsten PISA-Publikation (Toferer et al., 2022) des IQS wird ein Schwellenwert von 26 Punkten berichtet, der auf empirischen Schätzungen basiert. Unter Verwendung dieses Schwellenwerts ergeben sich Leistungsrückstände von 4,2 bzw. 5,0 Jahren. Wir orientieren uns hier jedoch an dem konservativen Richtwert.

zu beachten, dass Ungleichheitsmerkmale häufig kumuliert auftreten, da etwa in Familien mit Migrationshintergrund häufiger Deutsch nicht die Erstsprache ist und diese Familien im Schnitt einen geringeren sozioökonomischen Status bzw. ein geringeres Bildungsniveau aufweisen (Oberwimmer, Zintl, Juen & Vogtenhuber, 2021). Auf Basis der M8-Bildungsstandardüberprüfungen aus 2017 berechneten in diesem Zusammenhang Rothe, Weber und Illetschko (2022) etwa eine Differenz von rund 150 Punkten zwischen Schülerinnen und Schülern mit geringem sozioökonomischem Status, Migrationshintergrund und nicht-deutscher Erstsprache und Schülerinnen und Schülern aus oberen sozialen Schichten, ohne Migrationshintergrund und Deutsch als Erstsprache.

Auf der einen Seite bestehen soziale und ethnische Bildungsungleichheiten bereits vor der Einschulung als Folge differenzieller Sozialisationserfahrungen (Bradley & Corwyn, 2002; Skopek & Passaretta, 2021; Schöfl, Weber, Steinmair, Zepnik & Grimm, im Review). Auf der anderen Seite wird in der Literatur davon ausgegangen, dass Bildungsungleichheiten im Laufe der Schulkarriere von Kindern entstehen bzw. verstärkt werden (Maaz, Baumert & Trautwein, 2010). In diesem Zusammenhang werden differenzielle Lern- und Entwicklungsmilieus, die aus Institutionsunterschieden (in Unterrichtskulturen, Lehrkraftkompetenzen ...) zwischen Schulen verschiedenen Typs und – darauf liegt der Fokus dieses Beitrags – aus Unterschieden in der Schulzusammensetzung resultieren können, als mögliche Erklärungsansätze in den Blick genommen (Baumert et al., 2006; Neumann, Becker & Maaz, 2014).

3 Effekte der ethnischen und sozialen Zusammensetzung der Schülerschaft auf die Lernergebnisse

Unterschiede in der Schul- und Klassenzusammensetzung können über sogenannte Kompositionseffekte zu einer Vergrößerung von Bildungsungleichheiten beitragen. Ein Kompositionseffekt liegt vor, wenn Schülerinnen und Schüler bei gleichen individuellen Voraussetzungen (gleiche Lernausgangslagen und gleicher familiärer Hintergrund) unterschiedliche Lernergebnisse erzielen, je nachdem, mit wem sie die Schule bzw. Klasse besuchen (Biedermann et al., 2016; Dumont, Neumann, Maaz & Trautwein, 2013). Die Grundannahme ist, dass eine Kumulation von Schülerinnen und Schülern mit Ungleichheitsmerkmalen (geringer Sozialstatus, Migrationshintergrund, nichtdeutsche Alltagssprache) – was medial häufig unter dem Begriff der Brennpunktschulen diskutiert wird – ein ungünstiges Lern- und Entwicklungsklima mit sich bringt und in der weiteren Folge die (Lern-)Entwicklungschancen der per se schon benachteiligten Schülerinnen und Schüler noch weiter reduziert. In der internationalen Forschung werden insbesondere Effekte der sozioökonomischen und ethnischen Zusammensetzung von Schulen bzw. Klassen untersucht.⁴ Insgesamt weist die Forschungslage auf die Bedeutung der sozioökonomischen Zusammensetzung der Schülerschaft hin (Holzberger, Reinhold, Lüdtke & Seidel, 2020; van

4 Eine weitere zentrale Forschungslinie ist die Zusammensetzung von Lerngruppen (siehe u. a. Becker, Kocaj, Jansen, Dumont & Lüdtke, 2022), auf die in diesem Beitrag nicht näher eingegangen werden soll.

Ewijk & Sleegers, 2010; für Österreich siehe auch Biedermann et al., 2016). Demnach zeigen Schülerinnen und Schüler unabhängig von ihren individuellen Merkmalen eine geringere Leistungsentwicklung, wenn sie eine Schule bzw. Klasse mit einem höheren Anteil an sozioökonomisch benachteiligten Schul- bzw. Klassenkameradinnen und -kameraden besuchen. Umgekehrt wird eine bessere Leistungsentwicklung ermittelt, wenn Schülerinnen und Schüler vermehrt sozioökonomisch privilegierte Mitschülerinnen und Mitschüler haben. Die ethnische Schul- und Klassenzusammensetzung wird häufig am Anteil an Schülerinnen und Schülern mit Migrationshintergrund und im deutschsprachigen Kontext auch am Anteil an Schülerinnen und Schülern mit nichtdeutscher Erstsprache festgemacht. Die Forschungslage zeigt sich diesbezüglich – insbesondere für den deutschsprachigen Raum – nicht eindeutig. Auf der einen Seite weisen manche Studien darauf hin, dass mit höheren Anteilen an Schülerinnen und Schülern mit Migrationshintergrund bzw. nichtdeutscher Alltagssprache unabhängig von individuellen Merkmalen schlechtere Leistungen zu erwarten sind (Bacher & Weber, 2018; Biedermann et al., 2016; Rjosk, Richter, Lüdtke & Eccles, 2017; Seuring, Rjosk & Stanat, 2021). Andererseits können andere, allerdings ältere Studien einen Effekt der ethnischen Schülerinnen- und Schülerzusammensetzung nicht bestätigen (Dumont, Neumann, Nagy, Becker & Rose, 2013; Stanat, Schwippert & Gröhlich, 2010).

Noch wenig Befunde liegen zur Frage vor, ob die Ungleichheitsmerkmale additiv oder interaktiv/multiplikativ wirken. Zumeist wird eine additive Wirkung angenommen, z. B. indem ein additiver Sozialindex berechnet wird (z. B. Bruneforth et al., 2012; Rothe et al., 2022). Auf eine multiplikative Wirkung wird im Rahmen der qualitativen Intersektionsforschung hingewiesen, quantitative Befunde fehlen aber weitgehend. Das gilt auch für die Frage, ob die soziale und ethnische Zusammensetzung linear, nichtlinear oder interaktiv mit anderen Faktoren auf die Kompetenzentwicklung und den Schulerfolg wirkt. Entsprechend dem Literaturüberblick von van Ackeren et al. (2021) zu Schulen in herausfordernder Lage gibt es Schulen, die diese Herausforderung positiv bewältigen, während andere scheitern, was durch unterschiedliche schulische Faktoren, wie z. B. das Schulleitungshandeln, die Unterrichtsgestaltung, die Nutzung von Daten, die Vernetzung mit Akteurinnen und Akteuren außerhalb und/oder die Ressourcen erklärt werden kann.

Das differenzielle Lern- und Entwicklungsklima, das aus Unterschieden der Schul- und Klassenzusammensetzung resultieren kann, wird in der Literatur durch unterschiedliche Mechanismen – u. a. Peer- und Unterrichtseffekte – zu erklären versucht (Dumont, Neumann, Maaz & Trautwein, 2013; Harker & Tymms, 2004). Ein Peereffekt würde etwa vorliegen, wenn sich an Schulen oder Klassen mit einem erhöhten Anteil an benachteiligten Schülerinnen und Schülern ein durch geringe Wertschätzung von Bildung und ungünstige Motivationslagen gekennzeichnetes Klima entwickelt, das auch alle anderen Schülerinnen und Schüler „ansteckt“ (Demanet & van Houtte, 2014; Harker & Tymms, 2004). Ein solches Klima kann weiters auch ein Nährboden für vermehrtes schulisches Problemverhalten sein, das in der weiteren Folge die effektive Lernzeit reduzieren kann (Demanet & van Houtte, 2011). Im Hinblick auf den Anteil an Schülerinnen und Schülern mit nichtdeutscher Erstsprache und/oder mit geringerem elterlichem Sozialstatus kann auch angenommen werden, dass in Schulen/Klassen mit hohen Anteilen die Möglichkeiten der gegenseitigen fachlichen

Hilfe und Unterstützung der Schülerinnen und Schüler untereinander eingeschränkt sind (Rjosk et al., 2017).

Unterrichtseffekte beziehen sich allgemein darauf, dass sich die ethnische und sozioökonomische Zusammensetzung von Klassen bzw. Schulen auf den Unterricht bzw. die Unterrichtsqualität niederschlägt. Hierbei können unterschiedliche Mechanismen eine Rolle spielen. Beispielsweise weisen Studien darauf hin, dass an Schulen mit einem erhöhten Anteil an benachteiligten Schülerinnen und Schülern vermehrt geringer qualifizierte und weniger erfahrene Lehrkräfte unterrichten (Allen & Sims, 2018; Weber et al., 2019). Darüber hinaus wird angenommen, dass Leistungserwartungen von Lehrkräften an die Schülerschaft von der Schul- bzw. Klassenzusammensetzung abhängen. Demnach haben Lehrkräfte in Schulen/Klassen mit einem hohen Anteil an benachteiligten Lernenden (geringer sozioökonomischer Status, Migrationshintergrund ...) geringere Erwartungen an die Leistungen der Schülerinnen und Schüler, was sich in der weiteren Folge in einem weniger leistungsförderlichen Unterricht niederschlagen kann (Ready & Wright, 2011; Rjosk et al., 2017). Eine belgische Mixed-Methods-Studie (Agirdag, van Avermaet & van Houtte, 2013) konnte in diesem Zusammenhang etwa zeigen, dass ein hoher Anteil an Schülerinnen und Schülern mit Migrationshintergrund bzw. Schülerinnen und Schülern mit geringerem sozioökonomischem Status damit einhergehend, dass Lehrkräfte die Schülerinnen und Schüler als weniger „unterrichtbar“ beschrieben, was sich in der weiteren Folge bei den Schülerinnen und Schülern in einem Gefühl der Sinnlosigkeit schulischer Anstrengung und schließlich in den Leistungen niederschlug. Interviews mit den Lehrkräften wiesen darauf hin, dass die geringe Unterrichtbarkeitseinschätzung vor allem durch vermutete sprachliche Defizite der Schülerinnen und Schüler begründet wird. In Bezug auf die Annahme, wonach sich die soziale und ethnische Komposition auf die Unterrichtsqualität auswirkt, ist die Studienlage jedoch allgemein als uneindeutig zu bezeichnen. So konnten zwar Rjosk et al. (2014) einen erwarteten Effekt der sozioökonomischen Klassenkomposition auf Aspekte des Unterrichts (Lehrkraft legt Wert auf sprachliche Kompetenzen) zeigen, gleichzeitig ergaben sich leichte Effekte des Anteils der nichtdeutschsprachigen Schülerinnen und Schüler – in nicht erwarteter Richtung – auf die Bedeutung sprachlicher Kompetenz, aber auch auf weitere Aspekte des Unterrichts (Schülerorientierung, Classroom Management). Mit steigendem Anteil an Schülerinnen und Schülern mit nichtdeutscher Erstsprache berichteten die Schülerinnen und Schüler von einer höheren Unterrichtsqualität. Eine jüngere Studie von Fauth, Atlay, Dumont und Decristan (2021) konnte weiters keine Effekte der sozioökonomischen und ethnischen Klassenzusammensetzung auf Aspekte der Unterrichtsqualität zeigen. Zusammenfassend ist festzuhalten, dass Studien zu Mediatoren von Kompositionseffekten – also zu jenen Mechanismen, die dazu führen, dass sich die Zusammensetzung der Schülerschaft auf die Leistungsentwicklung auswirkt – insgesamt rar sind und auch der diesbezügliche Forschungsstand als uneindeutig zu bezeichnen ist. Auch ist darauf hinzuweisen, dass die genannten Studien Kompositionseffekte und etwaige vermittelnde Prozesse aus einer *Durchschnittsperspektive* betrachten, wonach der Schul- bzw. Klassendurchschnitt auf Schülerinnen und Schüler oder Lehrkräfte wirkt. Theoretische Ansätze, die Kompositionseffekte aus einer netzwerkanalytischen Perspektive betrachten (u. a. Bolletta, 2021; Boucher,

Rendall, Ushchev & Zenou, 2024), können dazu beitragen, Kompositionseffekte besser zu verstehen, da sie ermöglichen, die relevante Referenzgruppe – z. B. bezogen auf Motivation als Mediator von Kompositionseffekten die Geringmotivierten einer Klasse, den Klassendurchschnitt und die Hochmotivierten einer Klasse – zu modellieren.

4 Schulische Segregation – Unterschiede in der ethnischen und sozialen Zusammensetzung der Schülerschaft

Eine nötige Voraussetzung für Kompositionseffekte sind bestehende Unterschiede in der Schul- und Klassenzusammensetzung. Entsprechende Unterschiede sind im Kontext schulischer Segregation zu betrachten, wobei nachfolgend der Fokus auf zwischen-schulische Segregation gelegt wird (Unterschiede zwischen Schulen). Für die Frage der innerschulischen Segregation (d. h. Ungleichverteilung zwischen Klassen innerhalb einer Schule) ist auf Weber et al. (2019) bzw. auf Helm, George, Zintl und Weber in diesem Band zu verweisen. Zwischenschulische ethnische Segregation bezieht sich in der Regel auf die ungleiche Verteilung von Schülerinnen und Schülern mit Migrationshintergrund bzw. nichtdeutscher Erstsprache zwischen Schulen. Soziale Segregation beschreibt die ungleiche Verteilung von Schülerinnen und Schülern mit unterschiedlicher sozioökonomischer Herkunft auf Schulen. Soziale und ethnische Segregation kann einerseits die Folge von Wohnsegregation sein (Karsten, 2010). In diesem Fall bildet die Zusammensetzung der Schulen die ethnische und soziale Zusammensetzung des Einzugsgebiets (Schulsprengel) ab. Bestehen etwa Unterschiede im Migrationsanteil zwischen zwei Stadtteilen, werden sich die Schulen der Stadtteile im Migrationsanteil unterscheiden (Altrichter, Bacher, Beham, Nagy & Wetzelhütter, 2011). Eine schwedische Studie legt nahe, dass ethnische und soziale schulische Segregation in erster Linie durch Wohnraumsegregation bestimmt wird (Böhlmark, Holmlund & Lindahl, 2016). Schulische Segregation kann jedoch auch Wohnraumsegregation verstärken, wenn etwa ein hoher Anteil an Schülerinnen und Schülern mit Migrationshintergrund an den Schulen dazu führt, dass Familien ohne Migrationshintergrund aus der Wohngegend wegziehen (Karsten, 2010). Auf der anderen Seite kann zwischenschulische Segregation durch die Schulwahlentscheidung der Schülerinnen und Schüler bzw. deren Eltern entstehen.

Im österreichischen Schulsystem kommt auch der intendierten Segregation nach Leistung am Übergang von der Primarstufe auf die Sekundarstufe (allgemeinbildende höhere Schule [AHS] vs. [Neue⁵] Mittelschule [(N)MS] bzw. Hauptschule [HS]) eine zentrale Bedeutung zu. Aufgrund der Zusammenhänge zwischen Leistungen und Ungleichheitsmerkmalen (siehe oben) unterscheiden sich Schulen unterschiedlichen Typs auf der Sekundarstufe 1 auch in ihrer Zusammensetzung. Sogenannte sekundäre Ungleichheitseffekte – sozial schwache Schülerinnen und Schüler besuchen bei gleichen Leistungen wie ihre privilegierten Mitschülerinnen und Mitschüler seltener eine AHS – können zusätzlich solche Unterschiede

5 Die Neue Mittelschule ersetze im Jahr 2012/13 die bisherige Hauptschule als Regelschule. Seit dem Schuljahr 2020/21 wird sie nur mehr als Mittelschule bezeichnet.

in der Schulzusammensetzung verstärken (Bruneforth et al., 2012). Für Österreich zeigen Analysen, dass auf der Sekundarstufe 1 die Trennung der Schülerschaft für einen Großteil der sozialen Segregation (59 %–77 %) verantwortlich ist, jedoch nur zu einem geringen Ausmaß die bestehende ethnische Segregation (6 %–28 %) erklären kann (Lassnigg et al., 2019). Eine Ursache für die geringe Erklärungskraft der frühen Selektion auf die ethnische Segregation ist die Tatsache, dass Schülerinnen und Schüler mit Migrationshintergrund häufiger in Städten mit einem dichteren Angebot an AHS-Unterstufen wohnen und dieses Angebot wegen höherer Bildungsaspirationen (z. B. Astleithner et al., 2021) auch nutzen. Unabhängig von der Zweigliedrigkeit des Schulsystems auf der Sekundarstufe 1 kann dort, wo Wahlmöglichkeiten vorhanden bzw. räumlich vergrößert sind – etwa keine Zuteilung zu einer Sprengelschule besteht –, durch ein selektives Wahlverhalten Segregation verstärkt werden. Für Österreich konnte ein entsprechender Effekt im Zuge der Auflösung der Schulsprengel in Linz jedoch nicht bestätigt werden (Altrichter et al., 2011), wenn Veränderungen der räumlichen Segregation mitberücksichtigt werden. Schwedische Studien hingegen legen nahe, dass durch die Ausweitung von Wahlfreiheit bei der Schulwahl ethnische Segregation verstärkt wurde (Böhlmark et al., 2016; Yang Hansen & Gustafsson, 2016). In diesem Kontext ist das Phänomen des *Native Flight* zu nennen, wonach mit steigendem Anteil an Schülerinnen und Schülern mit Migrationshintergrund an Schulstandorten autochthone Schülerinnen und Schüler und deren Eltern vermehrt versuchen, an Schulen mit einem geringeren Migrationsanteil zu gelangen (Gerdes, 2013; Rangvid, 2010).

5 Fragestellungen des Beitrags

Vor dem Hintergrund der eingangs beschriebenen Änderung der österreichischen Population der Schülerinnen und Schüler im Hinblick auf Ungleichheitsmerkmale und der Bedeutung der sozioökonomischen und ethnischen Zusammensetzung der Schülerschaft für die Leistungsentwicklung der Lernenden geht der Beitrag folgenden Fragen nach:

1. Wie änderte sich die Population in den 2010er-Jahren in Hinblick auf sozioökonomische (Bildung der Eltern, beruflicher Status der Eltern) und ethnische (Migrationshintergrund, nichtdeutsche Erstsprache) Ungleichheitsmerkmale?

Dabei soll ergänzend zu den eingangs beschriebenen Änderungen ein umfassenderer Blick auf vier zentrale Ungleichheitsindikatoren gelegt werden.

2. Wie änderte sich die soziale und ethnische schulische Segregation im Zeitverlauf?

Durch die Änderung der Population der Schülerinnen und Schüler stellt sich die Frage, ob das Ausmaß der Segregation auch einem Wandel unterworfen war. Denkbar wäre etwa, dass Änderungen in der Wohnraumsegregation oder Phänomene wie *Native Flight* zur Änderung der Segregation beigetragen haben.

3. Lassen sich Unterschiede in der Veränderung der sozialen und ethnischen Schulzusammensetzung beobachten und von welchen Faktoren hängen diese ab?

Hier steht die Frage im Zentrum, ob es zwischen Schulen Unterschiede in der Entwicklung der sozialen und ethnischen Zusammensetzung gab und ob solche Unterschiede vom Urbanisierungsgrad bzw. auf der Sekundarstufe auch dem Schultyp abhängig waren.

4. Wie hängt die Änderung der sozialen und ethnischen Schulzusammensetzung mit Entwicklung der Leistungen auf Schulebene zusammen?

Abschließend soll untersucht werden, ob Änderungen der Schulzusammensetzung auch mit der einer differenziellen Leistungsentwicklung an Schulen einhergehen; beispielsweise in der Form, dass die Leistungsentwicklung an Schulen vergleichsweise schlechter ausfällt, an denen es zu einem verstärkten Zuwachs an Schülerinnen und Schülern mit Migrationshintergrund kommt. Hierbei ist zu beachten, dass sich nach den Baseline-Erhebungen eine kontinuierliche Verbesserung der mittleren Kompetenzen in allen Bereichen für alle Schülergruppen gezeigt hat (Neubacher & Wimmer, 2021; Breit et al., 2016, 2017),

6 Methode

6.1 Stichprobe

Der Beitrag greift auf die Stichproben der Baseline-Erhebungen zurück. Für die 4. Schulstufe wurden im Frühjahr 2010 insgesamt 9.478 Schülerinnen und Schüler an 267 Schulen und für die 8. Schulstufe im Frühjahr 2009 10.749 Schülerinnen und Schüler an 204 Schulen getestet. Die Stichprobe der Sekundarstufe umfasst 7.346 Schülerinnen und Schüler (146 Schulen) der damaligen Hauptschulen (HS), die im letzten Jahrzehnt in (Neue) Mittelschulen überführt wurden, und 3.404 AHS-Schülerinnen und -Schüler (58 Schulen).⁶ In beiden Fällen wurde eine geschichtete Zufallsstichprobe für die Schulen gezogen, bei der an jeder ausgewählten Schule entweder alle Klassen bzw. Unterrichtsgruppen⁷ (wenn zwischen einer und drei gemeldet wurden) oder eine Stichprobe von zwei Klassen bzw. Unterrichtsgruppen (wenn vier oder mehr gemeldet wurden) getestet wurden. Darüber hinaus greift der Beitrag auf Daten aller Bildungsstandardüberprüfungen (BIST-Ü) für diese Schulen zurück: konkret

6 Eine AHS führt auch Schülerinnen und Schüler in Mittelschulklassen. Diese Schule wird für die gesamten Analysen als AHS behandelt.

7 Bei den Erhebungen wurden Unterrichtsgruppen gezogen und gemeinsam getestet, die Klassenstruktur wurde allerdings auch erhoben. In der Volksschule setzten sich Unterrichtsgruppen und Klassen in der Regel identisch zusammen, auf der Sekundarstufe ist das hingegen nicht immer der Fall, insbesondere an den HS, wo durch die Bildung von Leistungsgruppen in den Fächern eine von den Klassen unterschiedliche Gruppierung unterrichtet wurde.

die BIST-Ü der 4. Schulstufe in Mathematik (2013 und 2018) und Deutsch (2015) sowie die BIST-Ü der 8. Schulstufe in Mathematik (2012 und 2017), in Englisch (2013 und 2019) und Deutsch (2016). Da die BIST-Ü jeweils Vollerhebungen waren, konnten deren Daten den Schulen der Baseline-Stichprobe zugeordnet werden. Somit konnte für Schulen der Baseline-Erhebung ein Längsschnitt auf Schulebene konstruiert werden, bei dem für jede Schule durch Schülerfragebögen⁸ wiederholte Informationen zur sozialen und ethnischen Schulzusammensetzung bzw. die BIST-Ü-Testergebnisse der Schülerinnen und Schüler verfügbar sind. Details zur Baseline-Erhebung sind in der IQS-Forschungsdatenbibliothek (2024a, 2024b) dargestellt, die Methodik zur Stichprobenziehung beschreiben George, Itzlinger-Bruneforth und Trendtel (2015).

6.2 Messungen

Ethnischer Hintergrund. Als Indikatoren des ethnischen Hintergrunds wurden der Migrationshintergrund (0 = nein, 1 = ja) entsprechend der BIFIE-Definition (u. a. Breit et al., 2017) und die Erstsprache (0 = nur deutsch, 1 = nicht nur deutsch) verwendet.

Anzumerken ist, dass in der Baseline-Erhebung der 8. Schulstufe Schülerinnen und Schüler sich für eine Erstsprache entscheiden mussten, während in den folgenden Erhebungen mehrere Erstsprachen angegeben werden konnten. Daher ist bei der Baseline-Erhebung der 8. Schulstufe der Anteil der Schülerinnen und Schüler mit nichtdeutscher Erstsprache relativ zu den anderen Erhebungen leicht unterschätzt. Folglich wurde diese Messung in den Analysen nicht verwendet.

Sozioökonomischer Hintergrund. Zur Erfassung des sozioökonomischen Hintergrunds wurde erstens der höchste berufliche Status der Eltern (höchster ISEI-Wert [HISEI], *International Socio-Economic Index of Occupational Status*; Ganzeboom, 2010) verwendet. Der ISEI wird auf Basis der IISCO-(International Standard Classification of Occupations-) Berufsklassifikation gebildet und basiert auf Durchschnittsgehältern und dem nötigen Ausbildungsniveau für Berufe. Anzumerken ist, dass für die Baseline-Erhebungen der HISEI noch auf Basis des ISCO-88 gebildet wurde, bei den BIST-Ü jedoch die ISCO-08-Version die Grundlage für die HISEI-Berechnung darstellt. Nach Ganzeboom (2013) bestehen drei Möglichkeiten im Umgang mit den unterschiedlichen ISCO-Versionen, die sich seiner Einschätzung nach nur geringfügig in unterschiedlichen Ergebnissen niederschlagen werden: Transformation von ISCO-08 in ISCO-88 und Verwendung von ISCO-88, Transformation von ISCO-88 in ISCO-08 und Verwendung von ISCO-08 und direkte Verwendung des HISEI auf Basis von ISCO-88 bzw. ISCO-08. Da in den Daten nicht die nötigen ISCO-Informationen (4-stelliger Code) für eine Neuberechnung der HISEI-Werte verfügbar sind, wird im vorliegenden Beitrag die dritte Variante gewählt. Darüber hinaus wird die höchste

8 Für die Primarstufe sind auch Daten auf Basis von Elternfragebögen vorhanden. Für den Beitrag wurden jedoch nur Daten der Fragebögen für die Schülerinnen und Schüler verwendet.

Bildung der Eltern auf einer 4-stufigen Skala (1 = max. Pflichtschule; 2 = Berufsausbildung; 3 = Matura; 4 = tertiärer Abschluss) erfasst.

Die verwendeten Daten zum ethnischen und sozioökonomischen Hintergrund der Schülerinnen und Schüler werden von den Schülerinnen und Schülern erhoben. Für die 4. Stufe liegen auch Angaben der Eltern vor, diese werden aufgrund höherer Anteile fehlender Daten und mit dem Ziel der größeren Konsistenz zwischen 4. und 8. Schulstufe nicht verwendet.

Als Leistungsmessungen werden die Bildungsstandardüberprüfungen in Mathematik (Primarstufe: Baseline 2010; BIST-Ü 2013 und BIST-Ü 2018; Sekundarstufe: Baseline 2009, BIST-Ü 2012, BIST-2017), Deutsch Lesen (Baseline 2010 und BIST-Ü 2015; Sekundarstufe: Baseline 2009 und BIST-Ü 2016) und Englisch (Sekundarstufe: Baseline 2009, BIST-Ü 2013, BIST-2019) herangezogen. Für alle gemessenen Fächer und Kompetenzbereiche wird für die Baseline ein Mittelwert von 500 Punkten mit einer Standardabweichung von 100 festgelegt, alle späteren Messungen werden auf diese Skala hin verlinkt.

Für Erläuterungen zur Testentwicklung siehe Itzlinger-Brune forth, Kuhn und Kiefer (2016) und Trendtel, Schwabe und Fellingner (2016). Details zur Skalierung der Erhebungen sowie zu deren Verlinkung – wodurch eine Vergleichbarkeit der erhobenen Kompetenzen im Zeitverlauf gewährleistet ist – finden sich in Trendtel, Pham und Yanagida (2016). Die Beschreibungen der erhobenen Kompetenzen finden sich in den jährlichen Studienberichten (Breit et al., 2016, 2017; Schreiner et al., 2019, 2020).

Darüber hinaus werden die Variablen Schultyp (0 = Hauptschule bzw. [Neue] Mittelschule [HS/(N)MS], 1 = allgemeinbildende höhere Schule [AHS]) und Urbanisierungsgrad des Schulstandorts (1 = dicht besiedelt/urban, 2 = mittel besiedelt, 3 = dünn besiedelt/ländlich) in den Analysen als erklärende Variablen verwendet. Die Variable Urbanität basiert auf der Klassifikation des Grads der Urbanisierung der Europäischen Kommission (Statistik Austria, 2012) mit Stand 2015 und kategorisiert die Schulstandorte gemäß der Besiedlungsdichte als überwiegend städtische Gemeinden (dicht besiedelt), mittel besiedelte Gemeinden und überwiegend ländliche Gemeinden (dünn besiedelt).

6.3 Analyse

Alle Analysen wurden mit Mplus 8.2 (Muthén & Muthén, 1998–2017) durchgeführt. Für alle Analysen wurden die Schülergewichte der Baseline-Erhebung verwendet, für spätere Zeitpunkte wurde das Baseline-Gewicht mit dem Gewichtungsfaktor des jeweiligen BIST-Durchgangs (Gewicht der Vollerhebungen, durch das Non-Response-Fälle berücksichtigt werden) multipliziert. Alle Ergebnisse basieren auf vom IQS zur Verfügung gestellten imputierten Daten (Robitzsch, Pham & Yanagida, 2016).

Zur Beantwortung der ersten Forschungsfrage (Änderung der Population der Schülerinnen und Schüler) wurden deskriptive Kennwerte (Mittelwerte und Standardabweichung für HISEI und Bildung der Eltern bzw. Anteilswerte für Migrationshintergrund und Erstsprache) der Ungleichheitsmerkmale für alle Testjahre berechnet. Zusätzlich wurde ein linearer Trend geschätzt (durchschnittliche jährliche Änderung über den Untersuchungszeitraum), der Auskunft darüber gibt, ob es im Untersuchungszeitraum zu einer von 0 verschiedenen Änderung in der Population kam. Um das komplexe Stichprobendesign zu berücksichtigen, wurde die Standardfehler mittels $Type = Complex$ korrigiert.

Zur Beantwortung der zweiten Forschungsfrage (Änderung der Segregation) wurde für die Ungleichheitsmerkmale die Intraklassenkorrelation (ICC) auf Basis eines Zweiebenenmodells (siehe u. a. Raudenbush & Bryk, 2002) für alle Testjahre berechnet. Die ICC ist zwischen 0 und 1 normiert und misst die Ähnlichkeit innerhalb von Schulen hinsichtlich der Ungleichheitsmerkmale bzw. die Unterschiedlichkeit zwischen Schulen.⁹ Ein Wert von 1 würde eine vollständige Segregation bedeuten, d. h., dass alle Schülerinnen und Schüler innerhalb von Schulen sich hinsichtlich der Ungleichheitsmerkmale nicht unterscheiden (z. B. alle Schülerinnen und Schüler mit Migrationshintergrund gehen gemeinsam in Schulen und alle Schülerinnen und Schüler ohne Migrationshintergrund gehen gemeinsam in Schulen). Ein ICC von 0 würde hingegen bedeuten, dass sich Schulen etwa in Hinblick auf den Migrationshintergrund nicht unterscheiden, also alle Schulen den gleichen Anteil von Schülerinnen und Schülern mit Migrationshintergrund haben. Ebenfalls wurde ein linearer Trend berechnet, der die durchschnittliche jährliche Änderung im ICC wiedergibt.

Zur Beantwortung der dritten Forschungsfrage (Unterschiede in den Änderungen der Schulzusammensetzung) wurde ein Strukturgleichungsmodellansatz zur Analyse der Entwicklung auf Institutionsebene auf Basis von Querschnittsdaten verwendet (Hochweber & Hartig, 2017). Dieser Mehrebenenansatz erlaubt die Spezifikation eines – im vorliegenden Fall linearen – latenten Wachstumsmodells auf Institutionsebene (hier die Schulebene) auf Basis einer nichtverbundenen Stichprobe auf Individualebene. Durch ein lineares Wachstumsmodell kann die Entwicklung der Zusammensetzung der Schulen durch einen Ausgangswert (Intercept) und einen linearen Wachstumsparameter (Slope) beschrieben werden. Für beide Parameter kann die Varianz geschätzt werden, die Schulunterschiede im Intercept und dem Slope abbilden. Neben einem un konditionalen Modell wurden in einem weiteren Schritt auch Prädiktoren (Urbanisierungsgrad und für die Sekundarstufe auch der Schultyp) für den Intercept und den Slope berücksichtigt.

Schlussendlich wurden zur Beantwortung der vierten Frage (Zusammenhang mit Leistungen) parallele Wachstumsmodelle spezifiziert. Neben der Entwicklung der Schul-

9 Es wurde für alle Merkmale der ICC für metrische Merkmale berechnet. Der ICC wurde auch unter der Annahme dichotomer Merkmale berechnet. Im Wesentlichen zeigen sich ähnliche Muster (zeitliche Trends). Da der ICC für dichotome Merkmale jedoch in der Regel höher ausfällt als der ICC bei metrischen Variablen, berichten wir im Sinne einer leichteren Vergleichbarkeit für alle Merkmale den ICC unter der Annahme metrischen Messniveaus.

zusammensetzung wurde hier auch für die Leistung ein Wachstumsmodell geschätzt.¹⁰ Die Korrelationen der Wachstumsparameter geben dann Auskunft über Entwicklungszusammenhänge. Ergibt sich etwa eine negative Korrelation zwischen den Slopes des Migrationsanteils und dem Leistungsslope, so würde das bedeuten, dass eine stärkere Zunahme des Migrationsanteils mit einer schwächeren Zunahme (oder auch Abnahme) der Leistung einhergeht.

Zur Einordnung der Stärke von zeitlichen Änderungen wurden für die Analysen zu den Forschungsfragen 1, 2 und 3 Effektstärkekoeffizienten berechnet, die einheitlich in einer Korrelationsmetrik berichtet werden.¹¹ Zur Einordnung folgen wir dem Vorschlag von Cohen (1992), wonach $|r| \geq 0,1$ einen kleinen, $|r| \geq 0,3$ einen mittleren und $|r| \geq 0,5$ einen großen Effekt darstellt.

7 Ergebnisse

7.1 Änderung der Population der Schülerinnen und Schüler in Bezug auf Ungleichheitsmerkmale (Forschungsfrage 1)

Die Ergebnisse zur Änderung der Population der Schülerinnen und Schüler sind in Tabelle 1 (Primarstufe) und Tabelle 2 (Sekundarstufe) dargestellt. Für die Primarstufe (Tabelle 1) zeigt sich erwartungsgemäß, dass einerseits der Anteil an Schülerinnen und Schülern mit Migrationshintergrund (Trend = 0,903, $p < 0,001$) und nichtdeutscher Erstsprache (Trend = 0,945, $p < 0,001$) im Schnitt um ein schwaches Prozent pro Jahr zugenommen hat. Auf der anderen Seite hat jedoch der durchschnittliche berufliche Status der Eltern (Trend = 0,367, $p < 0,001$) und auch deren Bildungsniveau (Trend = 0,020, $p < 0,001$) zugelegt. Diese Änderungen sind aber numerisch gering. Für den beruflichen Status und die Bildung der Eltern kann auch eine Zunahme der Streuung verzeichnet werden, wobei es sich beim beruflichen Status möglicherweise um ein Artefakt handelt, da 2010 noch eine andere Berechnung verwendet wurde (siehe Messungen). Wird der Trend ohne 2010 geschätzt, kann keine Zunahme der Streuung verzeichnet werden. Der signifikante Trend im Mittelwert bleibt jedoch bestehen.

Auch auf der Sekundarstufe zeigt sich das gleiche Muster. Einer Zunahme des Migrationsanteils und des Anteils der Schülerinnen und Schüler mit nichtdeutscher Erstsprache steht eine Zunahme des Bildungsniveaus und des beruflichen Status der Eltern gegenüber. Die Änderungen fallen aber schwach aus. Darüber hinaus zeigt sich für den beruflichen Status, dass die Zunahme vermutlich im Kontext der unterschiedlichen Berechnung des HISEI

10 Anzumerken ist, dass wir die Leistungsentwicklung in einem linearen Wachstum für Mathematik verankert haben und die Wachstumsform für die anderen Domänen frei geschätzt wurde, da sonst Punktunterschiede zwischen den Domänen die Schätzung der Entwicklung verzerren würden.

11 Koeffizienten in einer anderen Metrik (Cohen's d für Mittelwertunterschiede, Cohen's h für Unterschiede in Anteilswerten) wurden in eine Korrelationsmetrik entsprechend Lenhard und Lenhard (2022) umgerechnet.

Tabelle 1: Änderung der Schülerpopulation auf der Primarstufe

		2010	2013	2015	2018	Trend	
						M (SE)	ES
Beruflicher Status (HISEI)	M (SE)	45,836 (0,451)	46,572 (0,601)	47,904 (0,583)	48,615 (0,602)	0,367*** (0,050) 0,395*** (0,074)	0,087 0,076
	SD (SE)	16,774 (0,212)	20,856 (0,183)	20,791 (0,165)	21,055 (0,172)	0,545*** (0,025) 0,043 (0,035)	0,076
Höchste Bildung der Eltern	M (SE)	2,580 (0,024)	2,602 (0,029)	2,695 (0,026)	2,726 (0,027)	0,020*** (0,002)	0,144
	SD (SE)	0,928 (0,011)	0,944 (0,010)	0,942 (0,008)	0,975 (0,009)	0,005*** (0,001)	
Migrationshintergrund	% (SE)	16,996 (1,190)	20,139 (1,525)	21,476 (1,641)	24,341 (1,731)	0,903*** (0,120)	0,088
Nichtdeutsche Erstsprache	% (SE)	24,969 (1,520)	27,726 (1,769)	29,936 (1,859)	32,453 (1,892)	0,945*** (0,132)	0,084

Anmerkungen: M = Mittelwert, SD = Standardabweichung, SE = Standardfehler, ES = Effektstärke in Korrelationsmetrik über die gesamten acht Jahre. Die zweiten Werte in der Spalte Trend für den beruflichen Status geben den Trend ohne das Jahr 2010 wieder. Für Mittelwertänderungen wurde der Trend durch die Standardabweichung von 2010 bzw. 2013 dividiert. Für Änderungen in Anteilswerten wurden zur Berechnung der Effektstärken entsprechend Cohen's h die Anteilswerte Arkussinus-transformiert und mit den transformierten Werten der Trend berechnet. *** p < 0,001.

Tabelle 2: Änderung der Population der Schülerinnen und Schüler auf der Sekundarstufe

		2009	2012	2013	2016	2017	2019	Trend	
								M (SE)	ES
Beruflicher Status (HISEI)	M (SE)	46,575 (0,510)	50,551 (0,802)	50,741 (0,773)	51,284 (0,773)	51,260 (0,796)	50,688 (0,772)	0,325*** (0,038) 0,050 (0,044)	0,101 0,012
	SD (SE)	16,024 (0,185)	20,862 (0,128)	20,895 (0,132)	20,739 (0,138)	20,857 (0,132)	20,887 (0,128)	0,355*** (0,019) 0,002 (0,015)	
Höchste Bildung der Eltern	M (SE)	2,466 (0,027)	2,640 (0,038)	2,683 (0,039)	2,806 (0,037)	2,821 (0,039)	2,886 (0,034)	0,041*** (0,002)	0,206
	SD (SE)	0,972 (0,010)	0,976 (0,008)	0,976 (0,008)	0,976 (0,007)	0,967 (0,007)	0,966 (0,008)	-0,001 (0,001)	
Migrationshintergrund	% (SE)	17,278 (0,993)	18,766 (1,284)	19,428 (1,338)	20,875 (1,402)	21,822 (1,418)	24,314 (1,559)	0,672*** (0,095)	0,104
Nichtdeutsche Erstsprache	% (SE)		25,422 (1,447)	26,713 (1,534)	29,919 (1,648)	30,677 (1,645)	34,419 (1,787)	1,214*** (0,115)	0,134

Anmerkungen: M = Mittelwert, SD = Standardabweichung, SE = Standardfehler, ES = Effektstärke in Korrelationsmetrik über die gesamten zehn Jahre. Die zweiten Werte in der Spalte Trend für den beruflichen Status geben den Trend ohne das Jahr 2009 wieder. Für Mittelwertänderungen wurde der Trend durch die Standardabweichung von 2009 bzw. 2012 dividiert. Für Änderungen in Anteilswerten wurden zur Berechnung der Effektstärken entsprechend Cohen's h die Anteilswerte Arkussinus-transformiert und mit den transformierten Werten der Trend berechnet. *** p < 0,001.

im Jahr 2009 zu sehen ist, denn ohne Berücksichtigung des Jahrs 2009 ist der Trend nicht signifikant. Eine Änderung der Streuung der elterlichen Bildung und des beruflichen Status kann (nach Ausschluss der Daten aus 2009) auch nicht festgestellt werden.

7.2 Änderung der sozialen und ethnischen Segregation (Forschungsfrage 2)

Die Tabellen 3 und 4 stellen die Intraklassenkorrelationen (ICC) als Segregationsmaßzahl für die Ungleichheitsmerkmale im Zeitverlauf dar. Je höher die ICC-Werte ausfallen, desto stärker unterscheiden sich Schulen hinsichtlich der Ungleichheitsmerkmale und umso homogener (in Bezug auf die Ungleichheitsmerkmale) sind die Schülerinnen und Schüler innerhalb der Schulen. Ein höherer Wert ist also ein Hinweis auf eine stärkere Segregation. Insgesamt zeigt sich, dass sowohl auf der Primar- als auch Sekundarstufe die ethnische Segregation stärker als die soziale Segregation ausfällt. Im Hinblick auf die Entwicklung der Segregation zeigt sich auf der Primarstufe (Tabelle 3) für den Migrationshintergrund ein signifikanter positiver Trend (d. h. Zunahme der Segregation; $\Delta ICC = 0,004$, $p < 0,05$). Dieses Ergebnis legt also nahe, dass die Segregation von Schülerinnen und Schülern mit Migrationshintergrund zwischen Schulen im Zeitverlauf zugenommen hat, wobei die Zunahme als gering zu beurteilen ist. Für die anderen Merkmale kann keine signifikante (lineare) Änderung der Segregation festgestellt werden.

Auf der Sekundarstufe kann im Hinblick auf die elterliche Bildung ($\Delta ICC = 0,005$, $p < 0,01$), den Migrationshintergrund ($\Delta ICC = 0,011$, $p < 0,001$) und die nichtdeutsche Erstsprache ($\Delta ICC = 0,005$, $p < 0,05$) eine schwache, jedoch statistisch signifikante Zunahme der Segregation verzeichnet werden. Auffällig dabei ist, dass für Bildung der Eltern und Migrationshintergrund insbesondere zwischen 2009 und 2012 ein deutlicher Anstieg zu verzeichnen ist (Bildung von 0,122 auf 0,202; Migrationshintergrund von 0,161 auf 0,255) und dass – trotz positiven Trends über die gesamte untersuchte Zeitspanne – für Bildung zwischen 2017 und 2019 wieder ein Rückgang der Segregation sichtbar wurde. Da insbesondere die soziale Segregation auf der Sekundarstufe 1 im Zusammenhang mit dem zweigliedrigen Schulsystem zu sehen ist, werden in Tabelle 4 getrennte Analysen für AHS und (N)MS/HS berichtet. Erwartungsgemäß zeigt sich, dass die soziale Segregation innerhalb der Schultypen deutlich geringer ausfällt. Für die HS/(N)MS zeigt sich – wie in der gesamten Sekundarstufe – eine schwache, jedoch statistisch signifikante Zunahme der ethnischen Segregation (Migrationshintergrund: $\Delta ICC = 0,010$, $p < 0,001$; nichtdeutsche Erstsprache: $\Delta ICC = 0,007$, $p < 0,05$) und eine Zunahme der Segregation in Bezug auf den elterlichen Bildungshintergrund ($\Delta ICC = 0,003$, $p < 0,01$). Für die AHS zeigt sich zwar ein ähnliches Muster, jedoch fällt hier nur die Zunahme der Segregation im Hinblick auf den Migrationshintergrund signifikant aus ($\Delta ICC = 0,015$, $p < 0,001$).

Tabelle 4: Änderung der Segregation (Intraklassenkorrelation) auf der Sekundarstufe

	2009	2012	2013	2016	2017	2019	Trend	
	ICC (SE)	ICC (SE)	ICC (SE)	ICC (SE)	ICC (SE)	ICC (SE)	Δ ICC pro Jahr (SE)	ES
Beruflicher Status (HISEI)	0,184 (0,014)	0,209 (0,014)	0,201 (0,015)	0,206 (0,015)	0,209 (0,015)	0,203 (0,014)	0,002 (0,001) 0,000 (0,001)	0,002 0,000
Höchste Bildung der Eltern	0,122 (0,011)	0,202 (0,015)	0,206 (0,015)	0,200 (0,015)	0,211 (0,015)	0,167 (0,014)	0,005** (0,002)	0,050
Migrationshintergrund	0,161 (0,018)	0,268 (0,024)	0,255 (0,022)	0,270 (0,023)	0,278 (0,020)	0,279 (0,021)	0,011*** (0,002)	0,110
Nichtdeutsche Erstsprache		0,276 (0,024)	0,274 (0,023)	0,299 (0,024)	0,303 (0,023)	0,298 (0,022)	0,005* (0,002)	0,050
AHS								
Beruflicher Status (HISEI)	0,047 (0,011)	0,059 (0,014)	0,065 (0,015)	0,058 (0,014)	0,057 (0,013)	0,081 (0,016)	0,002 (0,001) 0,002 (0,001)	0,020 0,020
Höchste Bildung der Eltern	0,029 (0,009)	0,060 (0,012)	0,062 (0,013)	0,059 (0,011)	0,047 (0,010)	0,059 (0,011)	0,002 (0,001)	0,020
Migrationshintergrund	0,061 (0,018)	0,167 (0,041)	0,157 (0,039)	0,182 (0,043)	0,176 (0,037)	0,196 (0,044)	0,015*** (0,004)	0,150
Nichtdeutsche Erstsprache		0,153 (0,035)	0,142 (0,033)	0,185 (0,039)	0,160 (0,031)	0,176 (0,035)	0,005 (0,003)	0,050
NMS								
Beruflicher Status (HISEI)	0,057 (0,011)	0,074 (0,016)	0,075 (0,014)	0,082 (0,012)	0,090 (0,015)	0,087 (0,014)	0,003** (0,001) 0,002 (0,002)	0,030 0,020
Höchste Bildung der Eltern	0,030 (0,009)	0,054 (0,013)	0,065 (0,012)	0,068 (0,011)	0,067 (0,011)	0,057 (0,011)	0,003** (0,001)	0,030
Migrationshintergrund	0,189 (0,020)	0,285 (0,026)	0,273 (0,024)	0,289 (0,025)	0,299 (0,022)	0,290 (0,024)	0,010*** (0,002)	0,100
Nichtdeutsche Erstsprache		0,422 (0,069)	0,424 (0,064)	0,433 (0,051)	0,435 (0,045)	0,435 (0,046)	0,007* (0,003)	0,070

Anmerkungen: ICC = Intraklassenkorrelation, SE = Standardfehler, ES = Effektstärke in Korrelationsmetrik über die gesamten zehn Jahre. Für die Berechnung von Δ ICC pro Jahr wurden die ICCs vorab einer Fisher-z-Transformation unterzogen und anschließend wurde der Trend wieder auf die ICC-Metrik rücktransformiert. Für die Berechnung der Effektstärken wurde entsprechend Cohen's q der Trend auf Basis der Fisher-z-transformierten Werte verwendet. * p < 0,05; ** p < 0,01; *** p < 0,001.

7.3 Änderung der Schulzusammensetzung (Forschungsfrage 3)

In Tabelle 5 sind die Ergebnisse der Wachstumsmodelle zur Beschreibung der Änderung der sozioökonomischen und ethnischen Schulzusammensetzung dargestellt. Betrachtet man die durchschnittliche Entwicklung (d.h. Mittelwerte von Intercept und Slope), so zeigt sich das gleiche Bild wie in der gesamten untersuchten Population, d.h., die Anteile an Schülerinnen und Schülern mit Migrationshintergrund und nichtdeutscher Erstsprache nehmen auf Primar- und Sekundarstufe zu, ebenso kommt es zu einem Anstieg des durchschnittlichen elterlichen Bildungsniveaus. Auf der Primarstufe zeigt sich auch eine Zunahme des elterlichen beruflichen Status, wobei auf der Sekundarstufe nach Ausschluss des Jahres 2009 die Zunahme des beruflichen Status nicht signifikant ausfällt. Im Detail zeigt sich etwa für die Primarstufe, dass die durchschnittliche Schule im Jahr 2010 einen Anteil von 24,1 % (Mittelwert Intercept = 24,131) an Schülerinnen und Schülern mit nichtdeutscher Alltagssprache aufweist. In den Folgejahren steigt dieser Anteil im Durchschnitt um rund 0,6 Prozentpunkte (Mittelwert Slope = 0,572; $p < 0,001$) pro Jahr an, was einem kleinen Effekt ($ES = 0,104$) entspricht.¹²

Von zentralem Interesse ist bei diesen Analysen jedoch vor allem die Varianz des Slopes, da diese Entwicklungsunterschiede zwischen Schulen widerspiegelt, wodurch weitere Aussagen über die Entwicklungsunterschiede getroffen werden. Die Slope-Varianzen fallen für alle Merkmale sowohl auf der Primar- als auch Sekundarstufe statistisch signifikant aus.

Führt man das obige Beispiel zum Anteil der Schülerinnen und Schüler mit nichtdeutscher Erstsprache auf der Primarstufe fort, so kann eine Slope-Varianz von 0,388 ($p < 0,001$) festgestellt werden, was einer Standardabweichung von 0,623 entspricht. Unter der Annahme eines normalverteilten Slopes kann nun abgeleitet werden, dass zwar im Schnitt eine Zunahme des Anteils an Schülerinnen und Schülern mit nichtdeutscher Erstsprache erfolgt (Mittelwert Slope = 0,572), für manche Schulen jedoch keine Änderung oder sogar eine Abnahme vorliegt. Konkret ist für 13,5 % der Schulen ein Rückgang zwischen $-0,051$ Prozentpunkten (Mittelwert – Standardabweichung = $0,572 - 0,623 = -0,051$) und $-0,674$ Prozentpunkten (Mittelwert – $2 \times$ Standardabweichung) zu erwarten und für 2,4 % fielen die Abnahmen sogar noch stärker aus. Demgegenüber stehen Schulen mit einer deutlich stärkeren Zunahme (für 13,5 % der Schulen etwa zwischen 1,195 Prozentpunkten und 1,818 Prozentpunkten pro Jahr). Diese Interpretation, wonach es an manchen Schulen eine deutlich über dem durchschnittlichen Trend liegende Zunahme im Hinblick auf die untersuchten Merkmale gibt und an manchen Schulen auch eine entsprechende Abnahme zu verzeichnen war, kann für alle Merkmale übertragen werden. Eine Ausnahme stellt

12 Der Unterschied im Vergleich zu den Analysen zu Forschungsfrage 1 ist dadurch zu erklären, dass für Forschungsfrage 1 der Mittelwert der Gesamtstichprobe in den Blick genommen wurde, für Forschungsfrage 3 jedoch der Mittelwert von Schulmittelwerten analysiert wird. Das heißt, bei der Beantwortung der Forschungsfrage 3 wurde angenommen, dass alle Schulen dieselbe Schülerinnen- und Schülerzahl haben. Unterschiede in den Effektstärken ergeben sich des Weiteren dadurch, dass für Forschungsfrage 1 die Änderungen in Beziehung zur Gesamtstreuung gesetzt wurden und für Forschungsfrage 3 – entsprechend dem Fokus auf die Schulebene – Änderungen in Beziehung zur Streuung auf Schulebene gesetzt wurden.

der berufliche Status in der Sekundarstufe dar, da hier – nach Ausschluss der Daten aus 2009 – keine durchschnittliche Zunahme zu verzeichnen ist (für nachfolgende Analysen wurden die Daten aus 2009 für den beruflichen Status ausgeschlossen). Folglich impliziert die signifikante Slope-Varianz, dass ausgehend von keiner durchschnittlichen Änderung bei manchen Schulen eine Zunahme, bei anderen wiederum eine Abnahme des durchschnittlichen beruflichen Status der Eltern zu verzeichnen war.

Weitere zentrale Koeffizienten sind die Intercept-Slope-Kovarianzen, die Auskunft über den Zusammenhang zwischen Ausgangslage und Entwicklung geben. Hier zeigen sich sowohl auf der Primarstufe als auch Sekundarstufe signifikante positive Zusammenhänge zwischen dem Ausgangswert und der weiteren Entwicklung der ethnischen Schulzusammensetzung (Anteil der Schülerinnen und Schüler mit Migrationshintergrund bzw. Schülerinnen und Schüler mit nichtdeutscher Erstsprache). Diese Zusammenhänge implizieren, dass an Schulen mit einem höheren Anteil an Schülerinnen und Schülern mit nichtdeutscher Erstsprache und einem höheren Anteil an Schülerinnen und Schülern mit Migrationshintergrund zu Beginn der 2010er-Jahre auch die Zunahmen der jeweiligen Schülergruppen stärker ausfielen. Auf der Primarstufe zeigt sich ein analoger Zusammenhang für den beruflichen Status der Eltern. Diese Ergebnisse implizieren, dass die Unterschiede in der Schulzusammensetzung über die Zeit größer wurden, was sich zum Teil mit den Befunden zu Forschungsfrage 2 deckt. Diskrepanzen zwischen den Analysen zu Forschungsfrage 2 und den Ergebnissen in Tabelle 5 sind dadurch zu erklären, dass bei Forschungsfrage 2 (Tabelle 4) die Unterschiede zwischen Schulen immer in Relation zur gesamten Streuung der Ungleichheitsmerkmale betrachtet wurden und hier nur die Schulunterschiede betrachtet werden.¹³

In einem weiteren Schritt wurden Prädiktoren (Urbanisierungsgrad und für die Sekundarstufe auch der Schultyp) in die Wachstumsmodelle aufgenommen, um in den Blick zu nehmen, ob die Entwicklung der ethnischen und sozioökonomischen Schulzusammensetzung unterschiedlich ausfällt, je nachdem, in welcher Region die Schule sich befindet bzw. um welchen Schultyp es sich handelt. Die Ergebnisse sind in Tabelle 6 dargestellt.

Für die Primarstufe zeigt sich, dass zu Beginn der 2010er-Jahre Schulen in dicht besiedelten Gegenden sowohl durch eine günstigere sozioökonomische Zusammensetzung (höheres Bildungsniveau und beruflicher Status der Eltern) als auch durch einen höheren Anteil an Schülerinnen und Schülern mit Migrationshintergrund und nichtdeutscher Erstsprache gekennzeichnet waren als Schulen in dünn besiedelten Gegenden (siehe dazu die signifikanten Wald- χ^2 -Tests auf Unterschiede in den Ausgangswerten und die entsprechenden Effekte in Referenz zu einem mittleren Urbanisierungsgrad). Im Hinblick auf den beruflichen Status und die ethnische Schulzusammensetzung ergeben sich auch differenzielle Entwicklungen in Abhängigkeit vom Urbanisierungsgrad (siehe signifikanten Wald- χ^2 -Tests auf Unterschiede in der Entwicklung). So zeigt sich, dass bei Schulen in urbanen Gegenden eine stärkere

13 Würde etwa bei einem Merkmal die gesamte Streuung zunehmen (wie z. B. bei der Bildung auf der Primarstufe) und würde sich das gleichermaßen in den Schulen niederschlagen, so würden zwar Unterschiede auf Schulebene größer, jedoch würde sich das Verhältnis der Schulunterschiede zur Gesamtstreuung nicht ändern.

Tabelle 6: Latente Wachstumsmodelle zur Beschreibung der Entwicklung der Schulzusammensetzung unter Berücksichtigung von Prädiktoren

Primarstufe			Sekundarstufe				
HISEI	Bildung ^a	Sprache	Migration	HISEI	Bildung ^a	Sprache	Migration
Effekte auf Intercept							
Konstante	46,768 (0,761)	25,974 (0,418)	28,774 (1,954)	19,578 (1,619)	43,410 (0,683)	22,137 (0,253)	26,355 (2,254)
Urbanisierungsgrad (Referenz = mittel besiedelt)							
Dicht besiedelt	1,417 (1,391)	1,562* (0,744)	18,710*** (3,751)	14,819*** (3,168)	-0,944 (1,051)	0,309 (0,416)	24,984*** (3,304)
Dünn besiedelt	-2,818*** (0,861)	-1,338** (0,474)	-18,639*** (2,137)	-14,058*** (1,748)	-0,848 (0,827)	-0,277 (0,295)	-16,094*** (2,405)
Test auf Unterschied der Ausgangswerte:							
Wald- χ^2 , df = 2	19,528***	24,029***	185,493***	158,666***	1,381	2,355	162,428***
Schultyp (Referenz = NMS/HS)							
AHS					17,361*** (0,855)	7,660*** (0,375)	-15,832*** (2,635)
Effekte auf Slope							
Konstante	0,257 (0,084)	0,193 (0,038)	0,717 (0,148)	0,547 (0,135)	0,076 (0,070)	0,306 (0,030)	1,117 (0,196)
Urbanisierungsgrad (Referenz = mittel besiedelt)							
Dicht besiedelt	-0,044 (0,115)	-0,079 (0,054)	0,199 (0,220)	0,645** (0,212)	-0,237** (0,087)	-0,074* (0,033)	0,779** (0,252)
Dünn besiedelt	0,244* (0,103)	0,029 (0,047)	-0,426* (0,195)	-0,220 (0,173)	0,040 (0,090)	0,039 (0,035)	-0,437* (0,197)
Test auf Unterschiede in der Entwicklung:							
Wald- χ^2 , df = 2	10,465**	5,203	10,677**	20,064***	11,059**	14,024***	29,972***
							67,876***

Zunahme der Anteile der Schülerinnen und Schüler mit Migrationshintergrund (1,193 Prozentpunkte pro Jahr; $p < 0,001$; siehe Zeile Änderung für dicht besiedelt) bzw. nichtdeutscher Erstsprache (0,916 Prozentpunkte pro Jahr; $p < 0,001$) zu verzeichnen war, während an Schulen in dünn besiedelten Gegenden die Zunahme geringer, aber dennoch statistisch signifikant ausfiel (Erstsprache: 0,291 Prozentpunkte pro Jahr, $p < 0,05$; Migrationsanteil: 0,327 Prozentpunkte pro Jahr, $p < 0,001$). Für den beruflichen Status zeigt sich, dass die Zunahme an Schulen in dünn besiedelten Gegenden stärker ausfiel (durchschnittliche Zunahme des HISEI pro Jahr = 0,501; $p < 0,001$) als an Schulen im urbanen Bereich (durchschnittliche Zunahme des HISEI pro Jahr = 0,213; $p < 0,01$)¹⁴.

Für die Sekundarstufe ist analog zur Primarstufe an Schulen in urbanen Gegenden ein höherer Anteil an Schülerinnen und Schülern mit Migrationshintergrund bzw. nichtdeutscher Erstsprache zu verzeichnen. Unterschiede in der sozioökonomischen Zusammensetzung in Abhängigkeit des Urbanisierungsgrads bei Kontrolle des Schultyps können nicht festgestellt werden. Jedoch zeigt sich – wie zu erwarten – dass AHS durch eine günstigere sozioökonomische Zusammensetzung und einen geringeren Anteil an Schülerinnen und Schülern mit Migrationshintergrund bzw. nichtdeutscher Erstsprache gekennzeichnet sind.

Das Muster der Entwicklung der ethnischen Zusammensetzung in Abhängigkeit vom Urbanisierungsgrad fällt analog zur Primarstufe aus, d. h., es gibt höhere Zuwächse im Anteil an Schülerinnen und Schülern mit Migrationshintergrund bzw. nichtdeutscher Erstsprache an Schulen in urbanen Gegenden. Für den beruflichen Status ist interessant, dass an urbanen HS/(N)MS sogar ein signifikanter, von der Stärke her jedoch geringer Rückgang des durchschnittlichen HISEI (durchschnittliche Abnahme des HISEI pro Jahr = -0,162; $p < 0,05$) sichtbar wird, während an ruralen HS/(N)MS keine signifikante durchschnittliche Änderung festzustellen ist (durchschnittliche Zunahme des HISEI pro Jahr = 0,115; $p > 0,05$). Ein ähnlicher Effekt zeigt sich für die elterliche Bildung. An Schulen in urbanen Gegenden fällt die Zunahme des durchschnittlichen Bildungsniveaus geringer aus als an ruralen Schulen. Im Hinblick auf Effekte des Schultyps zeigt sich, dass an AHS (in mittel besiedelten Gegenden) keine Zunahme des Migrationsanteils zu verzeichnen ist (0,018 Prozentpunkte pro Jahr; $p > 0,05$), während an HS/(N)MS mittel besiedelter Gegenden eine signifikante Zunahme zu verzeichnen ist (0,532 Prozentpunkte pro Jahr; $p < 0,001$).

7.4 Änderung der Schulzusammensetzung und Leistungsänderung auf Schulebene (Forschungsfrage 4)

Abschließend wurde der Frage nachgegangen, ob die Entwicklung der Schulzusammensetzung mit der Entwicklung der durchschnittlichen Leistung in Zusammenhang steht. Dazu wurde neben der Entwicklung der Schulzusammensetzung die Leistungsentwicklung

14 Ein ähnliches Muster zeigt sich für die Bildung, wobei die Entwicklungsunterschiede marginal signifikant ($p < 0,10$) ausfallen.

modelliert.¹⁵ Für die Sekundarstufe wurden sowohl ein Gesamtmodell als auch getrennte Modelle für AHS und HS/(N)MS geschätzt. In Bezug auf die Leistungsentwicklung zeigt sich allgemein, dass im Schnitt die BIST-Ergebnisse der Schulen besser wurden (Primarstufe: Mittelwert Slope = 5,997, $p < 0,001$; Sekundarstufe: 4,311, $p < 0,001$; HS/(N)MS: Mittelwert Slope = 4,221, $p < 0,001$; AHS: Mittelwert Slope = 4,614, $p < 0,001$), jedoch signifikante Entwicklungsunterschiede bestehen (Primarstufe: Varianz Slope = 9,603, $p < 0,001$; Sekundarstufe: Varianz Slope = 1,304, $p < 0,01$; HS/(N)MS: Varianz Slope = 1,013, $p < 0,10$; AHS: Varianz Slope = 1,432, $p < 0,05$).

Die Ergebnisse im Hinblick auf die Zusammenhänge bei der Entwicklung sind in Tabelle 7 dargestellt. Als erwartbares Ergebnis zeigt sich, dass die Intercepts der Leistungsentwicklung mit den Intercepts der sozioökonomischen und ethnischen Schulzusammensetzung korrelieren. Kurz gesagt, an Schulen mit Schülerinnen und Schülern, deren Eltern einen höheren beruflichen Status bzw. eine höhere Bildung aufweisen, sowie an Schulen mit einem geringeren Anteil an Schülerinnen und Schülern mit Migrationshintergrund bzw. nichtdeutscher Erstsprache fallen die Leistungen besser aus als an Schulen mit einer ungünstigeren sozialen bzw. ethnischen Zusammensetzung. Darüber hinaus ergeben sich fallweise signifikante Korrelationen zwischen dem Intercept der Leistung und dem Slope der Schulzusammensetzung. Konkret zeigt sich etwa, dass an Schulen mit einer höheren Ausgangsleistung der durchschnittliche berufliche Status (Primarstufe: $r = 0,751$; $p < 0,001$; HS/(N)MS: $r = 0,578$; $p < 0,001$) und das elterliche Bildungsniveau vermehrt zunehmen (Sekundarstufe: $r = 0,341$, $p < 0,05$; HS/(N)MS: $r = 0,690$, $p < 0,001$) bzw. der Anteil an Schülerinnen und Schülern mit Migrationshintergrund (Primarstufe: $r = -0,510$; $p < 0,001$; Sekundarstufe gesamt: $r = -0,369$, $p < 0,001$; HS/(N)MS: $r = -0,610$; $p < 0,001$; AHS: $r = -0,378$, $p < 0,05$) und nichtdeutscher Erstsprache (Sekundarstufe: $r = -0,218$, $p < 0,05$; HS/(N)MS: $r = -0,398$, $p < 0,001$) weniger stark zunimmt als an Schulen mit geringeren Ausgangsleistungen. Einige dieser Zusammenhänge sind auch Abbildung 1 zur Illustration dargestellt. Es wird sichtbar, dass an Schulen mit einer anfänglich überdurchschnittlichen Leistung und einer damit einhergehenden benachteiligten sozialen bzw. ethnischen Zusammensetzung eine günstigere Entwicklung der Schulkomposition (hier dargestellt: geringere Zunahme an Schülerinnen und Schülern mit Migrationshintergrund bzw. stärkere Zunahme des durchschnittlichen beruflichen Status) erfolgt. Demgegenüber zeigt sich für Schulen mit einer unterdurchschnittlichen Ausgangsleistung und einer damit einhergehenden benachteiligten Zusammensetzung eine ungünstigere Entwicklung der Schulkomposition,¹⁶ wodurch sich die Schere in der Zusammensetzung zwischen Schulen mit anfänglich über- bzw. unterdurchschnittlichen Leistungen im Zeitverlauf öffnet. Schlussendlich zeigen sich – vor allem auf der Primarstufe – echte Entwicklungszusammenhänge in der Form, dass eine verstärkte

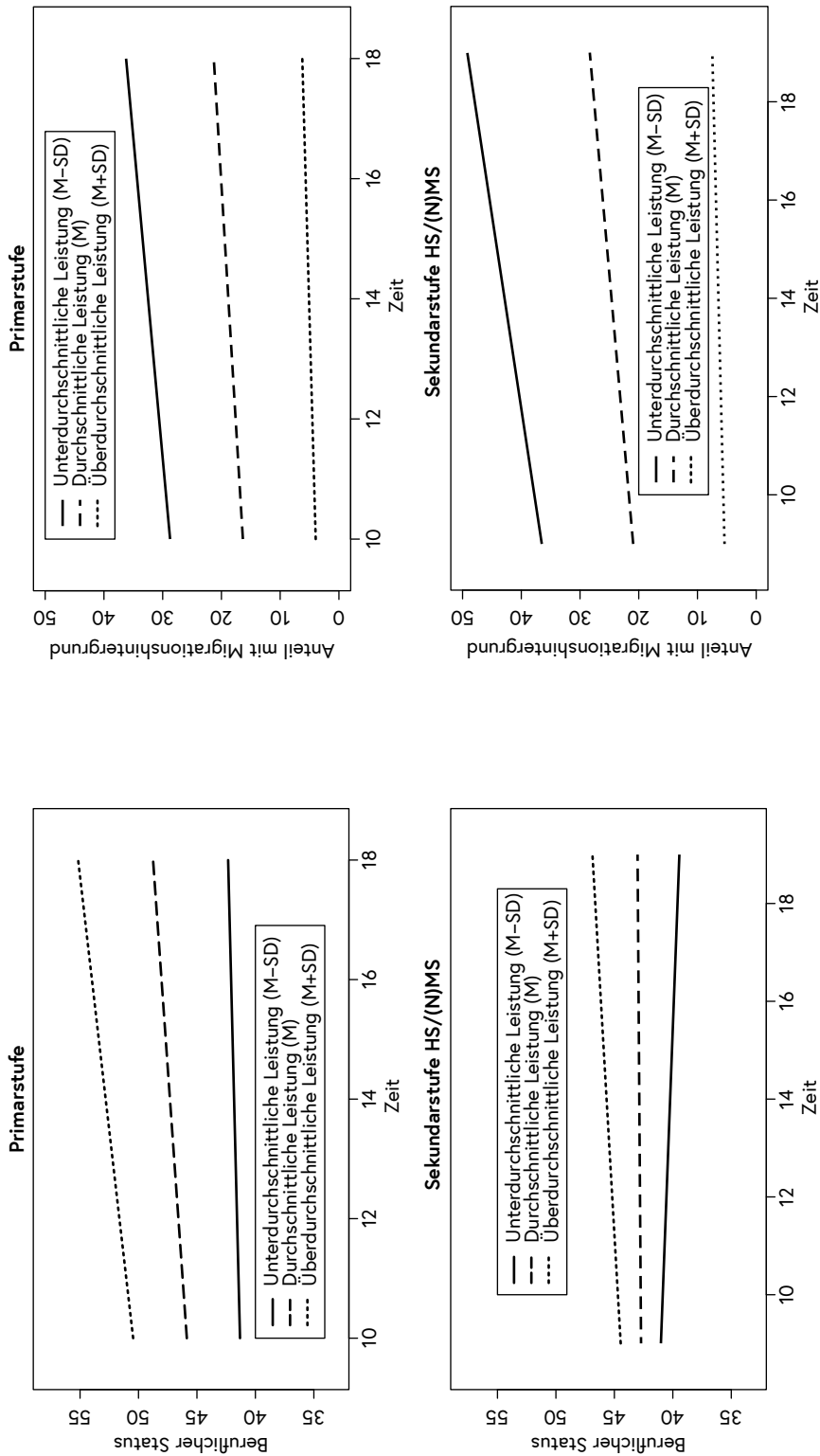
15 Bei den Analysen zur Leistungsentwicklung wurde ein lineares Wachstum für Mathematik (jeweils drei Messzeitpunkte) angenommen. Die Entwicklung der anderen Domänen wurde frei geschätzt, da sonst Domänenunterschiede die Leistungsentwicklung überlagern und verzerren würden. Durch diese Modellierung fließen Informationen der anderen Domänen in die Schätzung ein, Intercept und Slope sind aber in Referenz zur Mathematikentwicklung zu interpretieren.

16 Für HS/(N)MS ergibt sich bei Schulen mit einer unterdurchschnittlichen Eingangsleistung sogar ein Rückgang im durchschnittlichen beruflichen Status der Schülerschaft.

Zunahme der Leistungen mit einer Zunahme des beruflichen Status (Primarstufe: $r = 0,493$; $p < 0,05$; AHS: $r = 0,647$; $p < 0,05$), des Bildungsniveaus (Primarstufe: $r = 0,492$, $p < 0,01$) und einer geringeren Zunahme des Anteils an Schülerinnen und Schülern mit nichtdeutscher Erstsprache (Primarstufe: $r = -0,511$, $p < 0,01$) einhergeht. Diese Zusammenhänge sind zur Illustration in Abbildung 2 dargestellt.

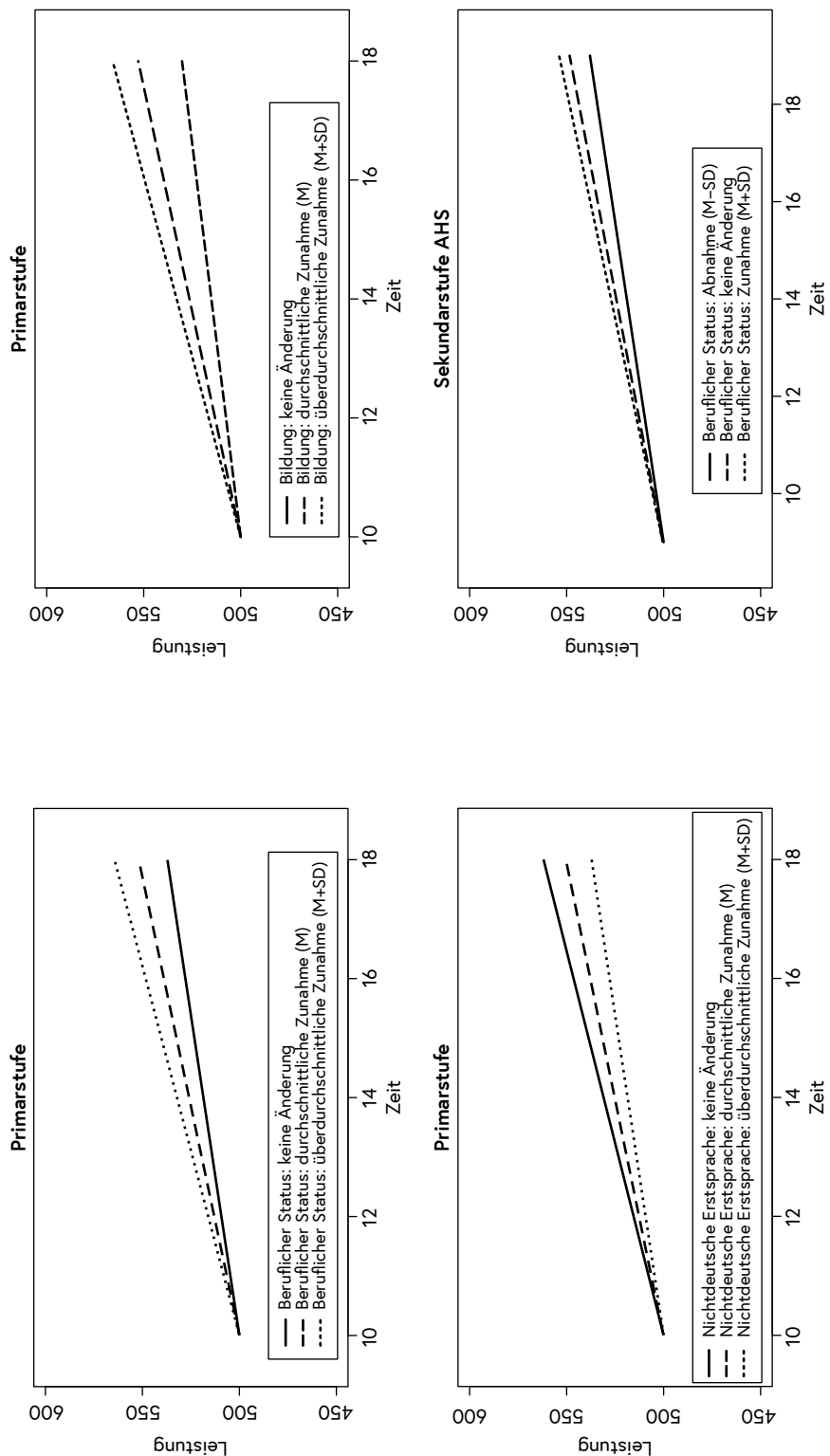
Tabelle 7: Parallele latente Wachstumsmodelle zum Zusammenhang der Entwicklung der Schulzusammensetzung und der durchschnittlichen Leistung der Schulen

		Leistungsentwicklung	
		Korrelationen der Wachstumsparameter (SE)	
		Intercept (I)	Slope (S)
		r (SE)	r (SE)
Primarstufe			
	I	0,746*** (0,039)	-0,097 (0,112)
Beruflicher Status	S	0,751*** (0,186)	0,493* (0,210)
Bildung	I	0,644*** (0,045)	-0,077 (0,112)
	S	0,203 (0,191)	0,592** (0,222)
Nichtdeutsche Erstsprache	I	-0,672*** (0,049)	-0,042 (0,118)
	S	-0,274 (0,163)	-0,511** (0,190)
Migrationshintergrund	I	-0,690*** (0,046)	-0,066 (0,108)
	S	-0,510** (0,155)	-0,287 (0,164)
Sekundarstufe			
Beruflicher Status	I	0,845*** (0,030)	-0,187 (0,138)
	S	0,107 (0,113)	0,500 (0,271)
Bildung	I	0,855*** (0,023)	-0,112 (0,128)
	S	0,341* (0,141)	0,172 (0,278)
Nichtdeutsche Erstsprache	I	-0,550*** (0,051)	0,201 (0,199)
	S	-0,218* (0,085)	-0,223 (0,230)
Migrationshintergrund	I	-0,586*** (0,044)	0,159 (0,189)
	S	-0,369*** (0,082)	-0,124 (0,236)
NMS			
Beruflicher Status	I	0,371*** (0,096)	-0,432 (0,264)
	S	0,578*** (0,145)	0,503 (0,410)
Bildung	I	0,412*** (0,095)	-0,403 (0,241)
	S	0,690*** (0,200)	0,193 (0,397)
Nichtdeutsche Erstsprache	I	-0,794 (0,040)	0,003 (0,340)
	S	-0,398*** (0,104)	-0,480 (0,402)
Migrationshintergrund	I	-0,795*** (0,038)	-0,010 (0,307)
	S	-0,610*** (0,088)	-0,313 (0,371)
AHS			
Beruflicher Status	I	0,765*** (0,087)	-0,388 (0,237)
	S	-0,107 (0,299)	0,647* (0,325)
Bildung	I	0,662*** (0,092)	-0,237 (0,241)
	S	-0,609 (0,469)	0,611 (0,538)
Nichtdeutsche Erstsprache	I	-0,619*** (0,123)	0,192 (0,279)
	S	-0,086 (0,233)	-0,318 (0,288)
Migrationshintergrund	I	-0,655*** (0,101)	0,189 (0,269)
	S	-0,378* (0,162)	-0,129 (0,267)
Anmerkungen: r = Korrelation, SE = Standardfehler. * $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$.			



Anmerkung: Für die grafische Darstellung wurde die Änderung der Schulzusammensetzung (Slope) auf die Änderung der Leistung bzw. der Ausgangswert der Leistung (Intercept) auf die Entwicklung der Schulzusammensetzung regressiert.

Abbildung 1: Ausgewählte Entwicklungszusammenhänge von Schulzusammensetzung und durchschnittlicher Leistung der Schulen, Abschnitt A.



Anmerkung: Für die grafische Darstellung wurde die Änderung der Schulzusammensetzung (Slope) auf die Änderung der Leistung bzw. der Ausgangswert der Leistung (Intercept) auf die Entwicklung der Schulzusammensetzung regressiert.

Abbildung 2: Ausgewählte Entwicklungszusammenhänge von Schulzusammensetzung und durchschnittlicher Leistung der Schulen, Abschnitt B.

8 Diskussion

Wie Daten aus den Bildungsstandardüberprüfungen des letzten Jahrzehnts erkennen lassen, kommt es in Österreich in den 2010er-Jahren zu einem Wandel der Population der Schülerinnen und Schüler (siehe auch Statistik Austria, 2024): Sowohl der Anteil der Schülerinnen und Schüler mit Migrationshintergrund als auch das Bildungsniveau der Eltern steigt in dieser Zeit an. Vor diesem Hintergrund verfolgt der vorliegende Beitrag das Ziel, die Entwicklung der Schülerpopulation differenzierter zu beschreiben. Darüber hinaus wird untersucht, wie sich die soziale und ethnische Schulzusammensetzung im Zeitverlauf entwickelt und ob ein Zusammenhang der Entwicklung der Schulzusammensetzung mit der Entwicklung der durchschnittlichen Leistung der Schulen besteht.

Im Hinblick auf die erste Forschungsfrage (Änderung der Population) liefern die Analysen Ergänzungen zu bisher schon Bekanntem (Zunahme Schülerinnen und Schüler mit Migrationshintergrund und nichtdeutscher Erstsprache, Zunahme des Bildungsniveaus der Eltern): Auf der Primarstufe steigt der berufliche Status der Eltern im Zeitverlauf, wobei dieser Effekt auf der Sekundarstufe nicht gezeigt werden kann.¹⁷ Der Effekt ist aber numerisch klein und seine inhaltliche Relevanz ist daher gering. Das gilt zum Teil auch für die Änderung im durchschnittlichen Bildungsniveau.

Die Analysen zur zweiten Forschungsfrage (Änderung der Segregation) zeigen, dass in den 2010er-Jahren zumindest die ethnische Segregation sowohl auf der Primarstufe als auch Sekundarstufe zunimmt, wobei auf der Sekundarstufe auch eine Zunahme der sozialen Segregation sichtbar wird. Diese Befunde legen nahe, dass Unterschiede in der ethnischen und sozioökonomischen Zusammensetzung zwischen Schulen – zumindest zu einem geringen Ausmaß – zunehmen.

Dieser Befund bekommt im Kern von den Analysen zur dritten Forschungsfrage (Änderung der Schulzusammensetzung) Bestätigung. Hier zeigt sich, dass sich Schulen in der Entwicklung der sozialen und ethnischen Zusammensetzung unterscheiden, d. h., auch wenn im Schnitt Zunahmen der untersuchten Merkmale zu verzeichnen sind, bei manchen Schulen keine Änderungen oder sogar Abnahmen vorliegen bzw. die Zunahmen deutlich über dem Durchschnitt liegen. Insbesondere zeigt sich in Bezug auf die Zunahme der Segregation, dass Schulen, die bereits zu Beginn der 2010er-Jahre durch einen höheren

17 Warum die Zunahme des HISEI nicht auch – zumindest zeitversetzt – in der Sekundarstufe sichtbar wird, kann an dieser Stelle nicht beantwortet werden. Anzumerken ist, dass ein ähnliches Muster auch in der Vollerhebungen der BIST-Ü auftritt. Hier zeigt sich zwischen 2013 und 2018 für die Primarstufe eine Zunahme im HISEI von $M = 47,33$ (Kinderangaben 2013) bzw. $M = 50,31$ (Elternangaben 2013) auf $M = 49,08$ (Kinderangaben 2018) bzw. $M = 52,37$ (Elternangaben 2018). Für die Sekundarstufe ergibt sich zwischen 2012 ($M = 49,02$) und 2019 ($M = 49,99$) ein geringerer Anstieg. Denkbar ist, dass Änderungen in der Population der Schülerinnen und Schüler u. a. durch Migration dazu geführt haben, dass ein gewisser Anteil an Schüler/innen erst in der Sekundarstufe in das österreichische Schulsystem eingetreten ist. Gestützt wird diese Annahme durch Daten der Schulstatistik (Statistik Austria, 2024). So gab es 2013 österreichweit 81.458 Schülerinnen und Schüler der 4. Schulstufe (Anteil nichtdeutsche Umgangssprache 25,3 %). Vier Jahre (2017) später umfasste die 8. Schulstufe 86.445 (Anteil nichtdeutsche Umgangssprache 27,5 %).

Anteil an Schülerinnen und Schülern mit Migrationshintergrund bzw. nichtdeutscher Erstsprache gekennzeichnet sind, in den Folgejahren einen größeren Zuwachs der jeweiligen Schülergruppe zu verzeichnen haben; also die Schere in der ethnischen Zusammensetzung zwischen Schulen weiter aufgeht. Dieses Ergebnis ist im Kontext regionaler Unterschiede im Anteil an Schülerinnen und Schülern mit Migrationshintergrund bzw. nichtdeutscher Erstsprache zu sehen (Biedermann et al., 2016), so zeigen die weiterführenden Analysen, dass an Schulen im urbanen Raum ein signifikant stärkerer Zuwachs dieser Schülergruppen zu verzeichnen ist als im ländlichen Raum. Da die Zunahme der Bevölkerung (Kinder) mit Migrationshintergrund bzw. nichtdeutscher Erstsprache vermehrt in urbanen Gegenden stattfindet (Statistik Austria, 2024), schlägt sich das in der weiteren Folge auch an den dortigen Schulen nieder. Verstärkend kommt vermutlich elterliches Wahlverhalten hinzu.

Schließlich zeigen sich Entwicklungszusammenhänge zwischen der Änderung der Schulsammensetzung und der Entwicklung der Schulleistungen. Hier ist einerseits interessant, dass an jenen Schulen, die zu Beginn der 2010er-Jahre gute Leistungen erzielen, die Zunahme an Schülerinnen und Schülern mit Migrationshintergrund bzw. nichtdeutscher Alltagssprache weniger stark ausfällt – was auch hier wieder mit regionalen Unterschieden erklärt werden kann. Auf der Sekundarstufe I lässt sich dieser Effekt nur bei der Hauptschule/Mittelschule beobachten. Eine Rolle spielen kann – vor allem in städtischen Gebieten – auch elterliches Wahlverhalten. So können die Ergebnisse ein Hinweis darauf sein, dass Eltern mit höherer Bildung und beruflicher Position sowie mit deutscher Erstsprache und ohne Migrationshintergrund Schulen wählen, wenn sie gute Leistungen erzielen und Eltern die Arbeit der Schulen als von hoher Qualität wahrnehmen.

Auf der Primarstufe zeigen sich Entwicklungszusammenhänge in der Form, dass Schulen mit einer günstigeren Entwicklung der Zusammensetzung (d. h. Zunahme der Bildung und des beruflichen Status der Eltern bzw. Abnahme des Anteils an Schülerinnen und Schülern mit nichtdeutscher Erstsprache) eine stärkere Zunahme der Leistungen aufweisen, was durch die im Theorieteil genannten Effekte erklärt werden kann, wie dadurch, dass Lehrkräfte ihre Leistungserwartungen an die soziale und ethnische Zusammensetzung oder die Leistungserwartungen der Eltern anpassen.

Die Studie weist bestimmte Limitationen auf. Sie vermittelt ein Gesamtbild für die Primarstufe und die Sekundarstufe I, aber kein regionales Detailbild etwa hinsichtlich der Segregation innerhalb der Städte. Diese differenzierten Analysen sind einerseits aufgrund der verwendeten Stichprobe (Baseline-Erhebung), die nicht alle Schulen beinhaltet und andererseits aus Platzgründen – sofern die Analysen auf die BIST-Ü-Daten beschränkt würden – nicht möglich. Für Detailanalysen stehen zudem nur wenige Kontextmerkmale zur Verfügung. Auch ist anzumerken, dass eine Prüfung der im Theorieteil angeführten Erklärungen unter Verwendung der BIST-Ü-Daten nicht bzw. nur bedingt möglich ist. Kompositionseffekte können etwa nur bedingt geprüft werden (siehe u. a. Biedermann et al., 2016) und werden in diesem Beitrag aus Platzgründen nicht untersucht. Abstand genommen wird aus diesem Grund auch von der Modellierung nichtlinearer Effekte, die z. B. bei der Bildung oder dem sozioökonomischen Status auftreten können. Des Weiteren

werden Entwicklungszusammenhänge der einzelnen Aspekte der ethnischen und sozialen Schulzusammensetzung sowie Interaktionseffekte der Ungleichheitsmerkmale im Hinblick auf die Leistungsentwicklung nicht untersucht. Mögliche Mediatoren von Kompositionseffekten können aufgrund fehlender Informationen z. B. zu Lehrkrafterwartungen nicht geprüft werden. Eine weitere Restriktion besteht darin, dass die Angaben zu den Ungleichheitsmerkmalen auf Angaben der Schülerinnen und Schüler basieren. Hier lässt sich vor allem hinsichtlich der beruflichen Position vermuten, dass die Messfehler stärker ausfallen als bei den anderen Merkmalen.

Trotz dieser Limitationen lassen sich aus den Befunden Implikationen für unterschiedliche Ebenen ableiten. Auf der Systemebene stellt sich die Frage, wie im Allgemeinen auf die Änderung der Population der Schülerinnen und Schüler und im Speziellen auf eine zunehmende ethnische Segregation zu reagieren ist. Bezogen auf die Zunahme der Segregation können einerseits Maßnahmen zur Reduktion in den Blick genommen werden und auf der anderen Seite Maßnahmen zur Kompensation der Segregation und deren Folgen. Da die Zunahme der Segregation vor allem im Zusammenhang mit einer gestiegenen großräumigen ethnischen Wohnraumsegregation zu betrachten ist (größerer Zuwachs in Städten als in ruralen Gegenden), ist hier vor allem die städtische Wohnpolitik gefordert, deren Ziel es sein sollte, eine zu starke Segregation zu vermeiden. Derzeit werden – wie einleitend erwähnt – Desegregationsmaßnahmen im Zusammenhang mit der in Wien aufgrund der Familienzusammenführung stark gestiegenen Anzahl von geflüchteten Kindern diskutiert. Diese zielen auf eine gleichmäßigere Verteilung von Geflüchteten auf alle Bundesländer ab und damit eine Reduktion des Anteils von Geflüchteten in Wien. Vorgeschlagen werden dafür u. a. eine Residenzpflicht bzw. eine Koppelung der Mindestsicherung an einen bestimmten Wohnort.

Weniger diskutiert werden Maßnahmen einer Desegregation, die direkt die Schulen betreffen. Als solche käme z. B. eine Quotenregelung¹⁸ in Frage, die Schulen verpflichtet, einen bestimmten Anteil an sozial benachteiligten Schülerinnen und Schülern aufzunehmen, eine Änderung der Schulsprengel, sodass das Einzugsgebiet eine größere Heterogenität in der Schülerschaft aufweist, ein Gutscheinsystem, das Eltern bzw. Schülerinnen und Schüler Schulen als zusätzliche Ressourcen anbieten können, und/oder eine begrenzte Schulwahl, bei der z. B. Eltern Wunschschulen nennen können, die Letztzuteilung aber durch die Schulbehörde erfolgt, (siehe dazu auch Biedermann et al., 2016). Folgt man der Studie von Gutiérrez, Jerrim und Torres (2020), die die Entwicklung der sozialen Segregation auf Basis von PISA-Daten zwischen 2000 und 2015 untersuchten, sind bisherige Ansätze noch wenig erfolgreich bei der Reduktion von Segregation. So können die Autoren keine wesentlichen Änderungen der sozialen Segregation in den OECD-Ländern feststellen, woraus sie schließen, dass stabile strukturelle Faktoren (z. B. Wohnraumsegregation) die zentralen Mechanismen hinter der schulischen Segregation sind, sich bildungspolitische Maßnahmen und Reformen kaum niederschlagen und folglich grundlegendere Maßnahmen zum Abbau von Segregation anzudenken sind.

18 Das aus den USA bekannte System des Busing sieht eine Zuweisung von benachteiligten Schülerinnen und Schülern vor. Ihm liegt somit eine Art Quotenregelung zugrunde.

Die Befunde von Gutiérrez et al. (2020) legen auch nahe, dass es immer ein bestimmtes Ausmaß an schulischer Segregation geben wird und daher kompensatorische Maßnahmen erforderlich sind. Eine diesbezügliche Maßnahme¹⁹ ist eine indexbasierte Mittelvergabe (Bacher, 2015a, b; Bruneforth et al., 2012; Groot-Wilken et al., 2016; Kuschaj & Schönpflug, 2013 sowie OECD, 2017), der zufolge Schulen mit einem höheren Anteil an sozial und ethnisch benachteiligten Schülerinnen und Schülern auf der Grundlage einer Formel mehr finanzielle Ressourcen erhalten, die sie entsprechend den konkreten Bedürfnissen auf der Grundlage eines Plans unter Nutzung der Ergebnisse der Schuleffektivitätsforschung einsetzen (Bacher, 2015a, b).

Richtet man nun den Blick von Maßnahmen der Systemsteuerung auf Implikationen für den Unterricht, so sind jene Mechanismen in den Blick zu nehmen, die auch für Effekte der Schul- und Klassenzusammensetzung auf die Leistungsentwicklung verantwortlich sein können (siehe dazu auch Helm et al., in diesem Band). Die Zunahme des Anteils an Schülerinnen und Schülern mit Migrationshintergrund und nichtdeutscher Alltagssprache ist die Realität des schulischen Alltags, auch wenn die Entwicklungen sich zwischen Schulen unterscheiden. Breitet sich dadurch eine Haltung aus, dass insbesondere an Schulen mit hohen Anteilen die Schülerinnen und Schüler aufgrund von „Sprachproblemen wenig unterrichtbar“ sind und Lehrkräfte folglich ihren Unterricht weniger förderlich und fordernd gestalten oder sich auch ein Klima der „Sinnlosigkeit schulischen Lernens“ unter den Schülerinnen und Schülern etabliert, dann kann sich das negativ in der Lernentwicklung der Schülerinnen und Schüler niederschlagen (Agirdag et al., 2013). Ähnlich wirken können gebiaste Leistungserwartungen – wonach bei gleichem Leistungsniveau Lehrkräfte für Schülerinnen und Schüler mit Migrationshintergrund oder nichtdeutscher Erstsprache eine geringere Leistungsentwicklung erwarten als für ihre Mitschülerinnen und Mitschüler (Lorenz, Gentrup, Kristen, Stanat & Kogan, 2016). Lehrkräfte bzw. angehende Lehrkräfte sollten im Rahmen ihrer Professionalisierung mit diesen Mechanismen konfrontiert und diesbezüglich sensibilisiert werden. Hier sind einerseits Hochschulen bei der Erstausbildung sowie der Fortbildung gefordert, diese Themen aufzugreifen und Angebote zum Aufbau nötiger Kompetenzen zu entwickeln (u. a. Ehmke, Hammer, Köker, Ohm & Koch-Priewe, 2018).

Diese Themen sollten auch im Rahmen von Schulentwicklungsprozessen durch Schulleitungen – insbesondere an Standorten mit hohen Anteilen an Schülerinnen und Schülern mit Migrationshintergrund bzw. nichtdeutscher Alltagssprache – adressiert werden,

19 In Österreich wurde im Bildungsreformgesetz 2017 in § 5 Bildungsdirektionen-Einrichtungsgesetz (BD-EG i. d. g. F.) die Grundlage einer indexbasierten Finanzierung als Kann-Bestimmung gesetzlich verankert. Demnach hat sich die Bewirtschaftung der Lehrpersonalressourcen jedenfalls an der Zahl der Schülerinnen und Schüler, am Bildungsangebot, am sozioökonomischen Hintergrund, am Förderbedarf der Schülerinnen und Schüler sowie an deren im Alltag gebrauchter Sprache und an den regionalen Bedürfnissen zu orientieren. Das zuständige Mitglied der Bundesregierung kann zur Berücksichtigung des sozioökonomischen Hintergrunds der Schülerinnen und Schüler durch Verordnung entsprechende Kriterien festlegen. Aktuell läuft ein Forschungs- und Entwicklungsprojekt mit dem Ziel, die Umsetzbarkeit einer indexbasierten Finanzierung empirisch zu prüfen (https://www.bmbwf.gv.at/Themen/schule/zrp/100schulen_1000chancen.html).

wobei hier auf aktuelle Forschung zu hinderlichen und förderlichen Faktoren für kulturell responsives Schulleitungshandeln zurückgegriffen werden kann (Brown et al., 2022).

Abschließend ist aber festzuhalten, dass die Schule diese Probleme allein nicht lösen kann, sondern dass alle gesellschaftlichen Bereiche und Politikfelder sowie -ebenen gefordert sind.

Literatur

- Agirdag, O., van Avermaet, P. & van Houtte, M. (2013). School segregation and math achievement: A mixed-method study on the role of self-fulfilling prophecies. *Teacher College Record*, 115, 1–50.
- Allen, R. & Sims, S. (2018). Do pupils from low-income families get low-quality teachers? Indirect evidence from English schools. *Oxford Review of Education*, 44 (4), 441–458. doi: 10.1080/03054985.2017.1421152
- Altrichter, H., Bacher, J., Beham, M., Nagy, G. & Wetzelschütter, D. (2011). The effects of a free school choice policy on parents' school choice behaviour. *Studies in Educational Evaluation*, 37 (4), 230–238. doi:10.1016/j.stueduc.2011.12.003
- Astleithner, F., Vogl, S. & Parzer, M. (2021). Zwischen Wunsch und Wirklichkeit: Zum Zusammenhang von sozialer Herkunft, Migration und Bildungsaspirationen. *Österreichische Zeitschrift für Soziologie*, 46 (3), 233–256. doi:10.1007/s11614-021-00442-3
- Bacher J. (2015a). Indexbasierte Finanzierung des österreichischen Schulsystems. *SchulVerwaltung*, 3 (4), 102–105.
- Bacher, J. (2015b). Mehr Bildungsgerechtigkeit durch sozialindexbasierte Finanzierung des österreichischen Bildungssystems? In H. Seckauer, C. Stelzer-Orthofer & B. Keplinger (Hrsg.), *Das Vorgefundene und das Mögliche* (S. 142–156). Wien: Mandelbaum.
- Bacher, J. & Weber, C. (2018). „Der hohe Anteil an Zuwandererkindern in den Schulen reduziert die Qualität der Ausbildung.“ In M. Haller (Hrsg.), *Migration & Integration – Fakten und Mythen. Siebzehn Schlagwörter auf dem Prüfstand*. Wien: Verlag der Österr. Akademie der Wissenschaften.
- Baumert, J., Stanat, P. & Watermann, R. (Hrsg.). (2006). *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit*. Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/978-3-531-90082-7
- Becker, M., Kocaj, A., Jansen, M., Dumont, H., & Lüdtke, O. (2022). Class-average achievement and individual achievement development: Testing achievement composition and peer spillover effects using five German longitudinal studies. *Journal of Educational Psychology*, 114 (1), 177.
- Biedermann, H., Weber, C., Herzog-Punzenberger, B. & Nagel, A. (2016). Auf die Mitschüler/innen kommt es an? Schulische Segregation – Effekte der Schul- und Klassenzusammensetzung in der Primarstufe und der Sekundarstufe I. In M. Bruneforth, F. Eder,

- K. Krainer, C. Schreiner, A. Seel & C. Spiel (Hrsg.), *Nationaler Bildungsbericht Österreich 2015, Band 2. Fokussierte Analysen bildungspolitischer Schwerpunktthemen* (S. 133–174). Graz: Leykam. doi:10.17888/nbb2015-2
- Böhlmark, A., Holmlund, H. & Lindahl, M. (2016). Parental choice, neighbourhood segregation or cream skimming? An analysis of school segregation after a generalized choice reform. *Journal of Population Economics*, 29 (4), 1155–1190. doi:10.1007/s00148-016-0595-y
- Bolletta, U. (2021). A model of peer effects in school. *Mathematical Social Sciences*, 114, 1–10.
- Boucher, V., Rendall, M., Ushchev, P. & Zenou, Y. (2024). Toward a general theory of peer effects. *Econometrica*, 92 (2), 543–565. doi:10.3982/ECTA21048
- Bradley, R. H. & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology*, 53, 371–399.
- Breit, S., Bruneforth, M. & Schreiner, C. (Hrsg.). (2016). *Standardüberprüfung 2015 Deutsch, 4. Schulstufe. Bundesergebnisbericht*. Salzburg: BIFIE.
- Breit, S., Bruneforth, M. & Schreiner, C. (Hrsg.). (2017). *Standardüberprüfung 2016 Deutsch, 8. Schulstufe. Bundesergebnisbericht*. Salzburg: BIFIE.
- Brown, M., Altrichter, H., Shiyan, I., Rodríguez Conde, M. J., McNamara, G., Herzog-Punzenberger, B. et al. (2022). Challenges and opportunities for culturally responsive leadership in schools: Evidence from four European countries. *Policy Futures in Education*, 20 (5), 580–607. doi:10.1177/14782103211040909
- Bruneforth, M., Weber, C. & Bacher, J. (2012). Chancengleichheit und garantiertes Bildungsminimum in Österreich. In B. Herzog-Punzenberger (Hrsg.), *Nationaler Bildungsbericht Österreich 2012, Band 2. Fokussierte Analysen bildungspolitischer Schwerpunkte* (S. 189–227). Graz: Leykam. doi:10.17888/nbb2012-2-5
- Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (BIFIE) (2019). *Standardüberprüfung 2018. Mathematik, 4. Schulstufe. Bundesergebnisbericht*. Salzburg: BIFIE.
- Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (BIFIE) (Hrsg.). (2020). *Standardüberprüfung 2019. Englisch, 8. Schulstufe. Bundesergebnisbericht*. Salzburg: BIFIE.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112 (1), 155–159.
- Demanet, J. & van Houtte, M. (2011). Social-ethnic school composition and school misconduct: Does sense of futility clarify the picture? *Sociological Spectrum*, 31 (2), 224–256. doi:10.1080/02732173.2011.541343
- Demanet, J. & van Houtte, M. (2014). Social-ethnic school composition and disengagement: An inquiry into the perceived control explanation. *The Social Science Journal*, 51 (4), 659–675. doi:10.1016/j.soscij.2014.09.001
- Dumont, H., Neumann, M., Maaz, K. & Trautwein, U. (2013). Die Zusammensetzung der Schülerschaft als Einflussfaktor für Schulleistungen. *Psychologie in Erziehung und Unterricht*, 60 (3), 163–183. doi:10.2378/peu2013.art14d
- Dumont, H., Neumann, M., Nagy, G., Becker, M. & Rose, N. (2013). Einfluss der Klassenkomposition auf die Leistungsentwicklung in Haupt- und Realschulen in Baden-Württemberg. *Psychologie in Erziehung und Unterricht*, 60 (3), 198–213. doi:10.2378/peu2013.art16d

- Ehmke, T., Hammer, S., Köker, A., Ohm, U. & Koch-Priewe, B. (Hrsg.). (2018). *Professionelle Kompetenzen angehender Lehrkräfte im Bereich Deutsch als Zweitsprache*. Münster: Waxmann.
- Fauth, B., Atlay, C., Dumont, H. & Decristan, J. (2021). Does what you get depend on who you are with? Effects of student composition on teaching quality. *Learning and Instruction*, 71, 101355. doi:10.1016/j.learninstruc.2020.101355
- Ganzeboom, H. B. G. (2010). *A new international socio-economic index (ISEI) of occupational status for the international standard classification of occupation 2008 (ISCO-08) constructed with data from the ISSP 2002–2007*. Paper presented at the Annual Conference of International Social Survey Programme, Lisbon.
- Ganzeboom, H. B. G. (2013). *Questions and Answers about ISCO-08*. Verfügbar unter: <http://www.harryganzeboom.nl/isco08/qa-isco-08.htm>
- George, A. C., Itzlinger-Bruneorth, U. & Trendtel, M. (2015). *Stichprobenziehung in Bildungsstandardüberprüfungen und Pilotierungen Technische Dokumentation – BIST-Ü*. Salzburg: BIFIE. Verfügbar unter: <https://www.iqs.gv.at/downloads/archiv-des-bifie/bildungsstandardueberpruefungen/technische-dokumentation>
- Gerdes, C. (2013). Does immigration induce “native flight” from public schools? *The Annals of Regional Science*, 50 (2), 645–666. doi:10.1007/s00168-012-0514-4
- Groot-Wilken, B., Isaac, K. & Schräpler, J. P. (Hrsg.). (2016). *Sozialindices für Schulen. Hintergründe, Methoden und Anwendung*. Münster: Waxmann.
- Gutiérrez, G., Jerrim, J. & Torres, R. (2020). School segregation across the world: Has any progress been made in reducing the separation of the rich from the poor? *The Journal of Economic Inequality*, 18 (2), 157–179. doi:10.1007/s10888-019-09437-3
- Harker, R. & Tymms, P. (2004). The effects of student composition on school outcomes. *School Effectiveness and School Improvement*, 15 (2), 177–199. doi:10.1076/sesi.15.2.177.30432
- Hochweber, J. & Hartig, J. (2017). Analyzing organizational growth in repeated cross-sectional designs using multilevel structural equation modeling. *Methodology*, 13 (3), 83–97. doi:10.25656/01:15867
- Holzberger, D., Reinhold, S., Lüdtke, O. & Seidel, T. (2020). A meta-analysis on the relationship between school characteristics and student outcomes in science and maths – evidence from large-scale studies. *Studies in Science Education*, 56 (1), 1–34. doi:10.1080/03057267.2020.1735758
- IQS-Forschungsdatenbibliothek (2024a). *Ausgangsmessung, 8. Schulstufe (2009)*. Verfügbar unter: <https://iqs.gv.at/themen/bildungsforschung/forschungsdatenbibliothek/daten-der-bildungsstandardueberpruefungen/fdb-ausgangsmessung-8-schulstufe-2009>
- IQS-Forschungsdatenbibliothek (2024b). *Ausgangsmessung, 4. Schulstufe (2010)*. Verfügbar unter: <https://iqs.gv.at/themen/bildungsforschung/forschungsdatenbibliothek/daten-der-bildungsstandardueberpruefungen/fdb-ausgangsmessung-4-schulstufe-2010>
- Itzlinger-Bruneorth, U., Kuhn, T. & Kiefer, T. (2016). Testkonstruktion. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandard-Überprüfung* (S. 21–50). Wien: facultas.
- Karsten, S. (2010). School segregation. In OECD (Hrsg.), *Equal opportunities? The labour market integration of the children of immigrants* (S. 193–209). Paris: OECD Publishing. doi:10.1787/9789264086395-8-en

- Kuschej, H. & Schönpflug, K. (2013). *Indikatoren bedarfsorientierter Mittelverteilung im österreichischen Pflichtschulwesen*. Wien: IHS.
- Lassnigg, L., Mayrhofer, L., Baumegger, D., Vogtenhuber, S., Weber, C., Aspetsberger, R. et al. (2019). Indikatoren B: Input – Personelle und finanzielle Ressourcen. In K. Oberwimmer, S. Vogtenhuber, L. Lassnigg & C. Schreiner (Hrsg.), *Nationaler Bildungsbericht Österreich 2018, Band 1. Das Schulsystem im Spiegel von Daten und Indikatoren* (S. 49–121). doi:10.17888/nbb2018-1-B.2
- Lenhard, W. & Lenhard, A. (2022). *Berechnung von Effektstärken*. Verfügbar unter: <https://www.psychometrica.de/effektstaerke.html>
- Lorenz, G., Gentrup, S., Kristen, C., Stanat, P. & Kogan, I. (2016). Stereotype bei Lehrkräften? Eine Untersuchung systematisch verzerrter Lehrererwartungen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 68 (1), 89–111. doi:10.1007/s11577-015-0352-3
- Maaz, K., Baumert, J. & Trautwein, U. (2010). Genese sozialer Ungleichheit im institutionellen Kontext der Schule: Wo entsteht und vergrößert sich soziale Ungleichheit? In J. Baumert, K. Maaz & U. Trautwein (Hrsg.), *Bildungsentscheidungen* (S. 11–46). Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/978-3-531-92216-4_2
- Muthén, L. K. & Muthén, B. O. (1998–2017). *Mplus User's Guide*. Eighth Edition. Los Angeles, CA: Muthén & Muthén.
- Neubacher, M. & Wimmer, C. (2021). Indikatoren D: Output – Ergebnisse des Schulsystems. In BMBWF (Hrsg.), *Nationaler Bildungsbericht Österreich 2021, Teil 2. Bildungsindikatoren* (S. 286–328). Wien. doi:10.17888/nbb2021-2-D
- Neumann, M., Becker, M. & Maaz, K. (2014). Soziale Ungleichheiten in der Kompetenzentwicklung in der Grundschule und der Sekundarstufe I. *Zeitschrift für Erziehungswissenschaft*, 17 (Suppl. 24), 167–203.
- Oberwimmer, K., Zintl, R., Juen, I. & Vogtenhuber, S. (2021). Indikatoren A: Kontext des Schul- und Bildungswesens. In BMBWF (Hrsg.), *Nationaler Bildungsbericht Österreich 2021, Teil 2. Bildungsindikatoren* (S. 158–185). Wien: BMBWF. doi:10.17888/nbb2021-2-A.
- OECD (2011). *PISA 2009 results: Overcoming social background – equity in learning opportunities and outcomes* (Volume II). Paris: OECD Publishing.
- OECD (2012). *Equity and quality in education. Supporting disadvantaged students and schools*. Paris: OECD Publishing. doi:10.1787/9789264130852-en
- OECD (2017). *The funding of school education. Connecting resources and learning*. Paris: OECD Publishing.
- Rangvid, B. S. (2010). School choice, universal vouchers and native flight from local schools. *European Sociological Review*, 26 (3), 319–335. doi:10.1093/esr/jcp024
- Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical Linear Models. Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.
- Ready, D. D. & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities. *American Educational Research Journal*, 48 (2), 335–360. doi:10.3102/0002831210374874
- Rjosk, C., Richter, D., Hochweber, J., Lüdtke, O., Klieme, E. & Stanat, P. (2014). Socioeconomic and language minority classroom composition and individual reading achievement: The

- mediating role of instructional quality. *Learning and Instruction*, 32, 63–72. doi:10.1016/j.learninstruc.2014.01.007
- Rjosk, C., Richter, D., Lüdtke, O. & Eccles, J. S. (2017). Ethnic composition and heterogeneity in the classroom: Their measurement and relationship with student outcomes. *Journal of Educational Psychology*, 109 (8), 1188–1204. doi:10.1037/edu0000185
- Robitzsch, A., Pham, G. & Yanagida, T. (2016). Fehlende Daten und Plausible Values. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandard-Überprüfung* (S. 293–259). Wien: facultas.
- Rothe, E., Weber, C. & Illetschko, M. (2022). Ethnische und soziale Bildungsungleichheiten in Mathematik – Anregungen für den Mathematikunterricht. In A. C. George, S. Götz, M. Illetschko & E. Süss-Stepancik (Hrsg.), *Empirische Befunde zu Kompetenzen im Mathematikunterricht der Sekundarstufe 1 und Folgerungen für die Praxis: Ergänzende Analysen zu den Bildungsstandardüberprüfungen* (S. 187–209). Münster: Waxmann.
- Schöfl, M., Weber, C., Steinmair, G., Zepnik, S. & Grimm, A. (im Review). The mediating role of language abilities and phonological processing in the association of bilingualism, socioeconomic status and spelling. *Frontiers in Education*.
- Schreiner, C. & Breit, S. (Hrsg.). (2012). *Standardüberprüfung 2012 Mathematik, 8. Schulstufe. Bundesergebnisbericht*. Salzburg: BIFIE.
- Schreiner, C., Breit, S., Pointinger, M., Pacher, K., Neubacher, M. & Wiesner, C. (Hrsg.). (2018). *Standardüberprüfung 2017 Mathematik, 8. Schulstufe. Bundesergebnisbericht*. Salzburg: BIFIE.
- Seuring, J., Rjosk, C. & Stanat, P. (2021). Ethnic classroom composition and minority language use among classmates: Do peers matter for students' language achievement? *European Sociological Review*, 36 (6), 920–936. doi:10.1093/esr/jcaa022
- Skopek, J. & Passaretta, G. (2021). Socioeconomic inequality in children's achievement from infancy to adolescence: The case of Germany. *Social Forces*, 100 (1), 86–112. doi:10.1093/sf/soaa093
- Stanat, P., Schwippert, K. & Gröhlich, C. (2010). Der Einfluss des Migrantenanteils in Schulklassen auf den Kompetenzerwerb. Längsschnittliche Überprüfung eines umstrittenen Effekts. In C. Allemann-Ghionda, P. Stanat, K. Göbel & C. Röhner (Hrsg.), *Migration, Identität, Sprache und Bildungserfolg*. 55. Beiheft der Zeitschrift für Pädagogik (S. 147–164). Weinheim: Beltz.
- Statistik Austria (Hrsg.). (2012). *Kurzbeschreibung internationaler Verfahren zur Klassifikation von Stadt und Land*. Wien: Statistik Austria.
- Statistik Austria (2024). *Schulstatistik. Schüler:innen ab 2006. STATcube – Statistische Datenbank von STATISTIK AUSTRIA*. Verfügbar unter: <https://statcube.at/>
- Toferer, B., Lang, B. & Salchegger, S. (Hrsg.). (2023). *PISA 2022. Kompetenzen in Mathematik, Lesen und Naturwissenschaft am Ende der Pflichtschulzeit im internationalen Vergleich*. Salzburg: IQS. doi:10.17888/pisa2022-eb
- Trendtel, M., Pham, G. & Yanagida, T. (2016). Skalierung und Linking. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandard-Überprüfung* (S. 185–224). Wien: facultas.
- Trendtel, M., Schwabe, F. & Fellingner, R. (2016). Differenzielles Itemfunktionieren in Subgruppen. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R. Methodische*

- Grundlagen der österreichischen Bildungsstandard-Überprüfung* (S. 111–148). Wien: facultas.
- Van Ackeren, I., Holtappels, H. G., Bremm, N., Hillebrand-Petri, A. & Kamski, I. (2021). Zur Einführung: Schulen in herausfordernden Lagen als Forschungs- und Entwicklungsfeld. In I. van Ackeren, H. G. Holtappels, N. Bremm & A. Hillebrand-Petri (Hrsg.), *Schulen in herausfordernden Lagen – Forschungsbefunde und Schulentwicklung in der Region Ruhr: Das Projekt „Potenziale entwickeln – Schulen stärken“* (S. 14–37). Weinheim: Beltz Juventa.
- Van Ewijk, R. & Sleegers, P. (2010). The effect of peer socioeconomic status on student achievement. A meta-analysis. *Educational Research Review*, 5 (2), 134–150. doi:10.1016/j.edurev.2010.02.001
- Weber, C., Moosbrugger, R., Hasengruber, K., Altrichter, H. & Schrodtt, H. (2019). Wer unterrichtet wen? Die Zusammensetzung von Klassen und Schulen und die Zuteilung von Lehrkräften. In S. Breit, F. Eder, K. Krainer, C. Schreiner, A. Seel & C. Spiel (Hrsg.), *Nationaler Bildungsbericht Österreich 2018, Band 2. Fokussierte Analysen und Zukunftsperspektiven für das Bildungswesen* (S. 143–182). Graz: Leykam. doi:10.17888/nbb2018-2-4
- Wendt, H., Kasper, D., Bos, W., Vennemann, M. & Goy, M. (2017). Wie viele Punkte auf der TIMSS-Metrik entsprechen einem Lernjahr? In T. Eckert & B. Gniewosz (Hrsg.), *Bildungsgerechtigkeit* (S. 121–152). Wiesbaden: Springer Fachmedien Wiesbaden. doi:10.1007/978-3-658-15003-7_8
- Yang Hansen, K. & Gustafsson, J.-E. (2016). Causes of educational segregation in Sweden – school choice or residential segregation. *Educational Research and Evaluation*, 22 (1–2), 23–44. doi:10.1080/13803611.2016.1178589

Michael Bruneforth

Der Einfluss der Klassengröße auf Mathematikkompetenzen in der österreichischen Volksschule

1 Einleitung

Der Wunsch nach kleineren Klassen, um bessere Unterrichtsbedingungen für Schülerinnen und Schüler sowie Lehrerinnen und Lehrer zu schaffen, ist zugleich alt wie auch aktuell vernehmbar. Bereits für 1842 dokumentiert Saldern (1993) die Beschwerden über große Klassen, die sich damals auf eine Klasse von 80 Knaben bezogen. Aber allen Veränderungen im Schulsystem zum Trotz, wird dieser Wunsch auch aktuell vielfach geäußert. So auch in Österreich: Laut „Der Standard“ (Arztmann, 29.8.2023) ist der Wunsch nach kleineren Klassen die dritthäufigste Nennung auf die Frage: „Was sollte verändert werden, um den Alltag von Lehrerinnen und Lehrern zu erleichtern?“ Zwar steht dieser Wunsch mit einer Nennung bei 13 % der Lehrpersonen deutlich hinter dem Wunsch nach weniger Bürokratie (51 %), aber doch nahezu gleichauf mit dem Wunsch nach mehr Unterstützungspersonal (18 %) und besseren Räumlichkeiten (12 %) und vor dem nach mehr Lehrpersonal (10 %) und den anderen fünf Vorschlägen, z. B. nach mehr Teamteaching. Auch in Interviews mit Lehrerinnen und Lehrern zur Zukunft von Schule (George & Klinglmair, 2023) zeigt sich der Wunsch nach kleineren Klassen.

Der Wunsch nach kleineren Klassen hat in Österreich im Schuljahr 2007/2008 zur Senkung der Klassenschülerhöchstzahl geführt (Rechnungshof Österreich, 2011) und auch in anderen Systemen folgten entsprechende Maßnahmen, so in Ontario, Kanada (Levin, 2008) oder in Kalifornien (Jepsen, 2015).

Gleichzeitig ist unumstritten, dass die Reduktion der Klassengröße zu den kostenintensivsten Maßnahmen zur Verbesserung des Unterrichts gehört. Hattie (2005) beurteilt aufgrund zu geringer Effekte kleinere Klassen nicht als sinnvolle Maßnahme und beschreibt die Forderungen danach als eine Flut des Lobbyismus („a tide of parent and teacher lobbying for smaller class-sizes“).

Der österreichische Rechnungshof (2011) schätzt die zusätzlichen Personalausgaben für die Maßnahme „Klassenschülerhöchstzahl 25“ allein im Schuljahr 2008/2009 auf rund

146 Millionen Euro. Die für die Maßnahme zusätzlich notwendigen 1.900 Klassen und rund 2.800 Lehrpersonen-Planstellen wurden allerdings nicht neu finanziert, sondern konnten aufgrund der damals günstigen demografischen Situation kostenneutral bereitgestellt werden, weil die Zahl der Schülerinnen und Schüler gleichzeitig sank (Rechnungshof Österreich, 2011). Angesichts der im internationalen Vergleich bereits sehr niedrigen mittleren Klassengrößen identifiziert die OECD die Erhöhung der mittleren Klassengröße für Österreich als einen Weg, die Effizienz des Schulsystems zu steigern (Nusche, Radinger, Theisens & Busemeyer, 2016). Auch für Deutschland stellt die Ständige Wissenschaftliche Kommission der KMK eine Erhöhung der mittleren Klassengröße als Option in den Raum, mit dem Hinweis darauf, dass es nur eine heterogene Studienlage zu den Effekten der Klassengröße auf Unterrichtsqualität und Lernleistungen gebe (Radisch & Vatterrott, 2024).

Stünde eine weitere teure Reduktion der Klassengröße als qualitätssteigernde Maßnahme zur Diskussion, sollte sich die Frage anschließen, ob diese Maßnahme für alle Schulen oder nur für ausgewählte, besonders förderbedürftige Gruppen Anwendung finden sollte. Kleinere Klassenschülerzahlen werden laut Altrichter, Trautmann, Wischer, Sommerauer und Doppler (2009) häufig als wesentliche Bedingung für produktiven Umgang mit heterogenen Klassen genannt. In Hamburg und Bremen erlaubt es die Mittelzuteilung über den Sozialindex in Schulen in sozial herausfordernder Lage, kleinere Klassen zu bilden (Bremische Bürgerschaft, 2017; Bürgerschaft der freien und Hansestadt Hamburg, 2012). In Frankreich werden zusätzliche Mittel in den benachteiligten Schulen der „Priority Education Zones“ hauptsächlich für kleinere Klassen eingesetzt (Bénabou, Kramarz & Prost, 2009). In Korea sind kleinere Klassen Teil eines Anreizsystems, das das Unterrichten in benachteiligten Schulen für die Lehrerinnen und Lehrer attraktiver machen soll (OECD, 2012). Insgesamt beurteilt die von der europäischen Union bestellte *expert group on quality investment in education and training* (European Commission, 2022, dort findet sich weitere Literatur) eine Reduktion der Klassengröße für benachteiligte Schülerinnen und Schüler zwar als teure, aber effiziente Maßnahme zur Steigerung der Schulqualität. Auch in Österreich wäre es vorstellbar, in diese Richtung zu denken, da laut § 5 des BDEG bei der Personalzuteilung auch sozioökonomische Rahmenbedingungen berücksichtigt werden sollten.

Da den hohen Erwartungen die hohen Kosten, die mit kleineren Klassen verbunden sind, gegenüberstehen, hat die grundlegende Frage große Bedeutung, ob und wie stark die Klassengröße mit den Unterrichtsergebnissen – in diesem Beitrag exemplarisch am Kompetenzerwerb – zusammenhängt.

Dieser Beitrag untersucht daher für Volksschulen in Österreich auf Basis eines natürlichen Experiments, ob und in welcher Stärke die Klassengröße mit den Mathematikkompetenzen am Ende der 4. Schulstufe zusammenhängt. Hierbei wird im Querschnitt und auch für ein Schulpanel von 1.716 Volksschulen mit ca. 6.000 Klassen im Längsschnitt untersucht, wie sich eine Änderung der mittleren Klassengröße an einer bestimmten Schule zu der Änderung der Mathematikkompetenzen im gleichen Zeitraum verhält. Aufgrund der großen Anzahl an Schulen in den Vollerhebungen der Überprüfung der Bildungsstandards ergeben sich zwischen den Messzeitpunkten in vielen Schulen Veränderungen

in der Klassengröße, die als natürliches Experiment interpretiert werden können. Die vorliegende Untersuchung bezieht sich alleinig auf Klassengröße und nicht generell auf Personalressourcen, d. h., sie erlaubt keine Aussagen über die Wirkung von Teamteaching und temporären Klassenteilungen z. B. im Rahmen des Fachunterrichts.

2 Forschungslage

Idealerweise würde eine Untersuchung zur Klassengröße sich nicht darauf beschränken, verallgemeinernd zu untersuchen, ob kleinere Klassen mit besseren und größere Klassen mit schlechteren Ergebnissen verbunden sind. Vielmehr sollte versucht werden, die eigentliche Kernfrage für die Bildungssteuerung anzugehen, welche Klassengröße pädagogisch ideal ist und in welchem Bereich die Größe vielleicht keinen Unterschied macht. Saldern (1993) kam aufgrund einer Befragung von Lehrpersonen vor drei Dekaden zu dem Schluss, dass aus Sicht von Lehrerinnen und Lehrern im Mittel eine Klasse von 20 Schülerinnen und Schülern (mit einer Spannweite von 10 bis 26 Kindern) ideal wäre, eine Zahl, die aus heutiger österreichischer Sicht überrascht. Mazrekaj, Witte und Triebs (2022) finden nahezu 30 Jahre später für Belgien auf Basis einer Stochastic Frontier Analysis den idealen Wert in der Volksschule auch bei 20 Kindern.

Als „ideal“ könnte, unabhängig von den Kosten, die Größe bezeichnet werden, mit der die größten Lernzuwächse zu erzielen wären. Hier ist anzunehmen, dass der mögliche Gewinn durch kleinere Klassen sich bei der Verkleinerung größerer Klassen eines Systems stärker zeigen würde als bei Änderungen für schon kleinere Klassen. Es könnte sogar angenommen werden, dass unter einer gewissen Größe sich die Effekte umdrehen, da zu kleine Klassen eventuell nachteilige Effekte haben (Saldern, 1993).

Aus Sicht der Steuerung steht dabei die Frage im Raum, welche zusätzlichen Ressourcen bzw. Kosten durch den erwartbaren Nutzen in verbesserten schulischen Ergebnissen bzw. anderen Aspekten der Schulqualität gerechtfertigt werden könnten. Dies ist zuerst eine vermutlich willkürlich zu beantwortende pädagogische Frage. Eine Steigerung von Schulqualität könnte sich hierbei auch in anderen Ergebnissen als im Kompetenzerwerb zeigen, z. B. (Klassenklima, soziale Kompetenzen, Lernfreude). Immer würde sich dabei aber die Frage der Kosten-Nutzen-Abwägung stellen. In Bezug auf Schulleistungen könnte/müsste man thematisieren: Welchen Preis darf eine Leistungssteigerung von x Punkten in den jeweiligen Kompetenztests haben? Aber damit verbunden ist die Frage, ob sich mit den benötigten Ressourcen evtl. durch andere Maßnahmen mehr erreichen ließe. Alternativ könnte man versuchen, den ökonomischen Nutzen der durch eine Klassengrößenreduktion ermöglichten besseren Bildung abzuschätzen (Jepsen, 2015). Krüger (2002) zeigt für das STAR-Programm, dass hier der langfristige ökonomische Nutzen die Kosten vermutlich übertrifft.

Der vorliegende Beitrag muss allerdings davon absehen, diese zentralen Fragen konkret zu beantworten. Vielmehr wird für die Situation in Österreich tatsächlich die Forschungsfrage

darauf reduziert, wie stark der Zusammenhang zwischen Klassengröße und Mathematikkompetenz ist. Allerdings soll versucht werden, diesen Zusammenhang gesondert für Schulen zu untersuchen, die zu wenigstens einem Messzeitpunkt große Klassen haben, die damit vermutlich über einer unbekannten idealen Größe liegen. Ferner sollen die Effekte für Schulen mit unterschiedlichen sozioökonomischen Rahmenbedingungen getrennt betrachtet werden.

2.1 Effekte der Klassengröße

Die Forschungslage zu den Effekten der Klassengröße wird auch international häufig als unbefriedigend und unklar eingeschätzt (z. B. in ihrer Übersichtsarbeit: Filges, Sonne-Schmidt & Nielsen, 2018), auch wenn deutlich mehr Arbeiten verfügbar sind, als hier berücksichtigt werden können. Insgesamt finden sich mehr Untersuchungen aus dem angelsächsischen Raum als aus Europa.

Vielfach werden negative Zusammenhänge zwischen Klassengröße und schulischen Ergebnissen (d. h., kleinere Klassen sind besser) gefunden, diese aber in der Größenordnung als zu gering angesehen, um der Klassengröße einen relevanten Einfluss zuzuschreiben. Dies schließen Jepsen (2015) und Filges et al. (2018) in ihren Forschungsübersichten, Letztere allerdings nur in Bezug auf Lesen, aber nicht für Mathematik. Auch Hattie (2005, 2009) schätzt in seiner Metaanalyse zu dieser Frage – basierend auf 14 angelsächsischen Studien aus den Jahren 1978 bis 2005 – den Effekt als existierend, aber zu gering ein. Für eine deutliche Reduktion der Klassengröße von 25 auf 15 Schülerinnen und Schüler gibt er über die Studien hinweg eine Effektstärke von .13 an. Weitere Beispiele für Einzelstudien mit demselben Schluss sind Shen und Konstantopoulos (2022, USA auf Basis der *Early Childhood Longitudinal Study*), Li und Konstantopoulos (2017, international vergleichend auf Basis von TIMSS 2011). Auch West und Wößmann (2002, auf Basis von TIMSS 1995) kommen zu einem ähnlichen Schluss und sehen positive Effekte kleinerer Klassen gar auf Länder begrenzt, die insgesamt schlechtere Ländermittelwerte haben, damit als wenig effiziente Systeme identifiziert werden. Nandrup (2014) zeigt für Dänemark förderliche Effekte kleinerer Klassen, insbesondere für die unteren Schulstufen, jedoch sind die Effekte gering: Eine Reduktion der Klassengröße um ein Kind der sechsten Schulstufe geht mit einer Steigerung der Leseleistung um 0,2 % einer Standarddeviation einher, in Mathematik sind es 0,9 %.

Paulus (2009) untersuchte für Grundschülerinnen und -schüler in 536 Klassen, die an Vergleichsarbeiten im deutschen Saarland teilnahmen, ob die Ergebnisse sich nach Klassengröße unterscheiden. Weder unkontrolliert noch nach Kontrolle von Migration oder sozioökonomischem Status (gemessen am Bücherbesitz) konnte er varianzanalytisch Unterschiede aufdecken. Allerdings zeigten sich signifikante Unterschiede in Lesen und Rechtschreibung nach einer Homogenisierung der Alterszusammensetzung.

Altrichter et al. (2009) beobachten für Österreich einen Mangel an Studien zum Thema, schließen aber aus Nebenergebnissen von österreichischen Studien zu anderen Themen darauf, dass die Wirkungen kleinerer Klassen auch in Österreich nur gering sind. Auch Neubacher et al. (2019) erkennen für Österreich auf Basis der Überprüfung der Bildungsstandards keinen Zusammenhang zwischen Klassengröße und Leseleistung, weder vor noch nach Kontrolle anderer Variablen. Lediglich in Volksschulen in dicht besiedelten Gemeinden gehen größere Klassen tendenziell mit etwas niedrigeren Testleistungen ein. Freunberger, Mayrhofer und Sauerwein (2019) finden im ländlichen Raum Österreichs Leistungsvorsprünge im Lesen bei Klein- und Kleinstschulen im Vergleich zu anderen Schulen, aber nicht in Mathematik.

Jepsen und Rivkin (2002) warnen sogar, dass eine Reduktion der Klassengröße trotz erwiesenen kleinen Nutzens langfristig negative Folgen haben kann, da der erhöhte Bedarf an Lehrerinnen und Lehrern dazu führt, dass weniger geeignetes Personal eingestellt wird, dessen Unterrichtsqualität geringer ist und potenzielle Gewinne der Klassenverkleinerung aufgezehrt werden.

Die Überblicksarbeit Bouguen, Grenet und Gurgand (2017) kommt hingegen zu einem anderen Schluss. Auf Basis von Studien aus den USA, Dänemark, Schweden und Norwegen schließen sie, dass eine Reduktion der Klassengröße deutlichen, die Kosten rechtfertigenden Nutzen haben kann.

Insbesondere die amerikanische STAR-Studie, auch von Bouguen et al. (2017) sowie Filges et al. (2018) herangezogen, wird vielfach als Beleg für den Nutzen von kleineren Klassen angeführt (Arnhold, 2004; im Brahm, 2006; Mishel & Rothstein, 2002; Nye, Hedges & Konstantopoulos, 2001). Die als randomisiertes Zufallsexperiment mit ca. 80 Schulen und 330 Klassen in Tennessee ab 1985 durchgeführte Studie zeigt deutliche positive Effekte auf Schülerinnen und Schüler zwischen dem Kindergartenjahr¹ und den ersten zwei Schulstufen. Hanushek, laut Mishel und Rothstein (2002, S. 1) „leading proponent of the prevailing view that money doesn’t make a difference“, interpretiert die Ergebnisse jedoch deutlich zurückhaltender und sieht allein den Effekt der kleinen Klassen im Kindergartenjahr als Ursache für die im Projekt gezeigten Effekte (Hanushek, 2002). Eine Sonderstellung nimmt die Studie von Mazrekaj et al. (2022) zu „Achievement Gaps“ ein, die keinen linearen Zusammenhang zwischen Klassengröße und Leistungslücken findet, aber bei nichtlinearer Betrachtung eine Klassengröße von 20 Schülerinnen und Schülern als Optimum identifiziert, d. h., Klassengröße wäre ein wichtiger Faktor für schulische Ergebnisse, aber nicht in dem Sinne, dass gilt, je kleiner, desto besser, sondern je näher am Optimum, desto besser. Bei Betrachtung der Verteilung der Klassengrößen in Österreich 2021/22 (Statistik Austria, 2023) wären damit zu kleine Klassen deutlich häufiger als zu große Klassen. Von

1 Hierbei bezieht sich der Begriff Kindergarten auf das erste Jahr im amerikanischen System der Schulstufen „K to 11“, d. h. einer Art Vorschulstufe.

den 18.389 Volksschulklassen liegen 46 % unter diesem vermeintlichen Optimum (1 bis 18 Schülerinnen und Schüler), 27 % lägen nahe am Optimum (19 bis 21 Schülerinnen und Schüler) und 26 % der Klassen wären zu groß, d. h. über dem Optimum.

Aufgrund der hohen Kosten, die mit kleinen Klassen einhergehen, stellt sich die Frage, ob diese Maßnahme gezielt für ausgewählte Schulen nützlich wäre. Es liegt nahe, anzunehmen, dass kleinere Klassen insbesondere für Schülerinnen und Schüler mit geringem sozioökonomischen Status positiven Einfluss auf den Lernerfolg haben könnten. Dies zeigt sich z. B. im STAR-Projekt im Vergleich der Effekte auf die Gruppen „White“ und „Minority“ (Arnhold, 2004), und Filges et al. zeigen dies an weiteren Studien. Nandrup (2014) findet stärkere Effekte für Schülerinnen und Schüler aus einkommensschwachen Familien. Auch Bouguen et al. (2017) schlagen aus Kostengründen vor, kleine Klassen gezielt für Schulen vorzusehen, deren Schülerinnen und Schüler geringeren sozioökonomischen Status haben. Gershenson und Langbein (2015) zeigen für North Carolina (USA) zumindest einen Zusammenhang zwischen steigender Schulgröße und abnehmenden Leistungen, insbesondere bei Schülerinnen und Schülern aus Familien mit niedrigerem sozioökonomischen Status.

Als Zwischenfazit lässt sich festhalten, dass in der Tendenz der Nutzen von kleineren Klassen in der Forschung zwar festgehalten wird, der Effekt sich aber meistens als sehr klein oder vernachlässigbar herausstellt. Auch für Österreich liegen erste Ergebnisse vor, die in diese Richtung deuten. Allerdings ist die nahezu 40 Jahre alte STAR-Studie, die sich als einzige echte Experimentalstudie auszeichnet, eine Ausnahme, indem sie positive Effekte zeigt. Die Studie Mazrekaj et al. (2022) deutet auf nichtlineare Zusammenhänge hin.

Wenn dieses Zwischenfazit halten sollte, stellt sich die Frage, warum die verbreitet erwartete Verbesserung der Ergebnisse in kleinen Klassen so häufig ausbleibt. Hattie (2009, S. 88) beobachtet Folgendes: „A typical response to this lower than expected effect of reducing class size is to note that many of the more powerful influences identified in this book could be more effective if the class size was lower.“ Die Annahme, dass kleinere Klassen eine andere Pädagogik bräuchten, um den Nutzen zu bringen, ist auch in der Literatur verbreitet (Altrichter et al., 2009; im Brahm, 2006; West & Wößmann, 2002). In diesem Zusammenhang müsste theoretisch beleuchtet werden, welche Wirkmechanismen dazu beitragen könnten, dass eine Reduktion der Klassengröße auch zu einer Verbesserung der Unterrichtsqualität und damit zu besseren Ergebnissen führt. Saldern (1993) und Altrichter et al. (2009) geben einen Überblick dazu. Wichtig ist somit, festzuhalten, dass die oben genannten Forschungsarbeiten, die keine oder nur wenig relevante Effekte zeigen, damit nicht ausschließen, dass bei entsprechenden begleitenden Reformen des Unterrichts Effekte eintreten könnten. Vielmehr deuten sie aber an, dass es den Akteuren selten gelingt, den Nutzen der teuren Maßnahme zu realisieren. Auch die gegenständliche Untersuchung fragt nicht danach, ob kleine Klassen in Österreich bei angepasster Pädagogik einen Nutzen haben könnten. Vielmehr will sie nur untersuchen, ob es aktuell Schulen bzw. Lehrerinnen und Lehrern in Österreich gelingt, bei einer Reduktion der mittleren Klassengröße bessere Ergebnisse zu erzielen, bzw. ob umgekehrt bei einer sich ergebenden Erhöhung der mittleren Klassengröße an einer Schule sich die Ergebnisse tendenziell verschlechtern.

2.2 Forschungsdesigns bisheriger Forschung

Der hier zusammengefasste Forschungsstand basiert auf Studien unterschiedlichster Designs. Bouguen et al. (2017) geben einen Überblick zu den methodischen Ansätzen. Als zentrale Herausforderung beurteilen sie es, beim Vergleich von großen und kleinen Klassen den Selection Bias zu „neutralisieren“, d. h., mit Leistung verbundene Faktoren, die bei der Zuteilung der Schülerinnen und Schüler in Klassen eine Rolle spielen, zu kontrollieren. Allerdings zeigen sich für Österreich bei Helm, George, Zintl und Weber (2025, im gleichen Band) keine systematischen und signifikanten Zusammenhänge zwischen Mechanismen der Klassenzuweisung und dem Schulerfolg.

Filges et al. (2018) beurteilen jedoch einen Großteil der von ihnen untersuchten Studien als methodisch nicht ausreichend. Von 127 ausgewerteten Studien konnten nur 10 die Kriterien für einen Einschluss in ihre Metaanalyse erfüllen. Allgemein als Königsweg angesehen ist das Zufallsexperiment, d. h. die randomisierte Zuteilung von Schulen bzw. Schülerinnen und Schülern zu Treatment- und Kontrollgruppen, wobei in der Treatmentgruppe die Klassengrößen reduziert werden. Bouguen et al. (2017) führen für dieses Design nur die STAR-Studie an und auch West und Wößmann (2002) und der Verfasser haben kein anderes Beispiel gefunden. Da Zufallsexperimente im Bildungsbereich nahezu ausgeschlossen sind, sollten, so Maag Merki (2014), quasi-experimentelle Designs für die empirische Bildungsforschung und in der Schuleffektivitätsforschung möglichst intensiv ausgelotet werden.

Alternativ bieten Quasi-Experimente eine Möglichkeit, sich der Frage anzunähern, solange die Zuordnung der Schülerinnen und Schüler in große und kleine Klassen von zufälligen, d. h. nicht beeinflussbaren Faktoren abhängt, die nicht mit wichtigen anderen Faktoren der Unterrichtsqualität oder Eigenschaften der Schülerinnen und Schüler zusammenhängen (z. B.: Jepsen & Rivkin, 2002, auf Basis des *California's class size reduction program*). Häufig werden auch zufällige, natürliche Schwankungen in der Anzahl der Schülerinnen und Schüler einer Schule genommen, die – auch aufgrund von Teilungszahlen – zu zufälligen Unterschieden in der Klassengröße führen. Hier werden somit keine Treatment- und Kontrollgruppen verglichen, sondern die natürliche Variation von Leistungen und Klassengröße in Schulen ohne besondere Maßnahme analysiert. Die Analyse bezieht sich dann auf den Vergleich zwischen Schulen mit verschiedenen Klassengrößen unter Kontrolle der anderen Faktoren. Neubacher et al. (2019) wenden diesen Ansatz für Österreich an. Bouguen et al. (2017) beschreiben die Untersuchung von Schulen mit Schülerinnen- und Schülerzahlen, die knapp unter bzw. über den Teilungszahlen liegen.

Bei den genannten Ansätzen wird der Effekt durch Unterschiede zwischen verschiedenen Schulen quantifiziert. West und Wößmann (2002) verfolgen einen anderen Ansatz, indem sie Unterschiede zwischen unterschiedlich großen Klassen innerhalb derselben Schulen analysieren. Die Datenbank zur TIMS-Studie von 1995 beinhaltet für 18 Länder verwertbare Daten zur 7. und 8. Schulstufe derselben Schulen. Durch die Analyse von *differences-in-differences* werden Unterschiede zwischen Schulen kontrolliert und es kann die Variation von Leistungen und Klassengrößen zwischen den Klassen innerhalb der Schulen analysiert

werden, während Schwankungen zwischen Schulen kontrolliert sind. Auch hier handelt es sich um ein Quasi-Experiment, da angenommen werden kann, dass Schwankungen der Klassengrößen zwischen Schulstufe 7 und Schulstufe 8 sich zufällig ergeben. Auch die vorliegende Studie zielt – auf Basis eines längsschnittlichen Schulpanels – auf Unterschiede zwischen Kohorten innerhalb von Schulen ab.

Eine Schwäche von Analysen auf Basis von Daten zu einzelnen Erhebungsjahren – wie West und Wößmann (2002) mit den Daten der TIMS-Studie oder Neubacher et al. (2019) mit Daten der BIST-Ü zeigen – ist, dass die Klassengröße nur für das letzte Jahr vor der Messung bekannt ist. Allerdings sollten bei Faktoren wie der Klassengröße langfristige Effekte berücksichtigt werden, da eine Senkung sich vermutlich nur über einen längeren Zeitraum bemerkbar machen kann. In dieser Studie werden daher die Klassengrößen der Schülerinnen und Schüler über die vier Schuljahre bis zur Messung ermittelt und im Mehrebenenmodell immer 2 Kohorten innerhalb einer Schule verglichen, deren Klassenstärke für beide Gruppen über die vier Jahre gemittelt verglichen wird.

Zwischenfazit: Mit der einen sehr prominenten Ausnahme, bei der ein Zufallsversuch eingesetzt wurde, arbeiten fast alle Studien mit Quasi-Experimenten, bei denen große und kleine Klassen verglichen werden und andere Faktoren und somit der Selection Bias kontrolliert werden sollen. Bei allen Studien, mit Ausnahme von West und Wößmann, müssen dazu Unterschiede zwischen Schulen kontrolliert werden. West und Wößmann verglichen Klassen unterschiedlicher Größe innerhalb von Schulen im selben Jahr zwischen Schulstufen. Dabei ist die Kontrolle der anderen Einflüsse ausreichend als Kontrolle der Unterschiede innerhalb der Schule.

Keine der genannten Studien nutzt eine längsschnittliche Betrachtung von Schulen. Damit unterscheidet sich die hier vorgelegte Studie von allen hier genannten. Das Schulpanel der Überprüfung der Bildungsstandards (BIST-Ü) erlaubt es, Veränderungen an Schulen für jeweils die gleichen Schulstufen in unterschiedlichen Jahren zu untersuchen. Damit ähnelt diese Studie dem Ansatz des deutschen PISA-Schulpanels von 2000 bzw. 2009 (Bischof, Hochweber, Hartig & Klieme, 2013) mit 59 Gymnasien. Das Panel der BIST-Ü erlaubt die längsschnittliche Analyse der Entwicklung aller Volksschulen in Österreich über einen Zeitraum von fünf Jahren (2013–2018). Wie bei West und Wößmann fokussiert die Analyse auf Änderungen in den Schulen, hier nun zwischen Kohorten gleicher Schulstufe im Abstand von 5 Jahren. Die vorliegende Studie verfolgt den gleichen analytischen Ansatz, den bereits Schmid, Trendtel, Bruneforth und Hartig (2019) zur Untersuchung der Wirkung von Schulentwicklungsplänen mithilfe des Panels der BIST-Ü vorgeschlagen haben. Der längsschnittliche Ansatz bringt eine bessere Kontrolle von nicht erfassten Merkmalen der Schulen, die den Kompetenzerwerb beeinflussen können, mit sich. Da Eigenschaften innerhalb einer Ebene über die Zeit vielfach konstant bleiben, d. h., Schulen in 2013 immer mit sich selbst in 2018 in Bezug gesetzt werden, werden bei der längsschnittlichen Analyse weit mehr Faktoren implizit kontrolliert, als im Querschnitt möglich wäre oder empirisch erfassbar, da z. B. Faktoren wie Nachbarschaft, vom Erhalter bereitgestellte Ressourcen oder Zubringer-Kindergärten für die Schulen meistens zwischen den Erhebungsjahren recht stabil sind.

3 Ziel der Studie und Forschungsfragen

Diese Studie verfolgt nicht das Ziel, aus theoretischen Überlegungen überprüfbare Hypothesen abzuleiten und diese Theorien zu prüfen. Vielmehr setzt sich die Studie mit der in der Einleitung skizzierten weit verbreiteten Annahme auseinander, dass kleinere Klassen mit besseren Unterrichtserfolgen einhergehen. Die Forschungsfragen sollen entsprechende Zusammenhänge unter den aktuellen Bedingungen im Schulsystem in Österreich untersuchen. Die Annahme von positiven Effekten der Senkung der Klassengröße liegt allen Fragestellungen mit Ausnahme von F5 zur Überprüfung zugrunde. Die Möglichkeiten dieser Studie erlauben es nicht, zu verstehen, welche Prozesse zwischen Klassengröße und Leistungen, d. h. Kompetenzerwerb, wirken. Es werden keine intervenierenden Variablen, wie Schul- oder Unterrichtsqualität oder Zufriedenheit der Lehrkräfte, untersucht. Damit kann nicht untersucht werden, warum bzw. warum nicht in kleinen Klassen höhere Leistungen erbracht werden. Auch bleibt die Untersuchung auf den Blick auf Kompetenzerwerb beschränkt, andere denkbare positive Effekte kleinerer Klassen werden nicht untersucht. Diese Studie untersucht für Volksschulen in Österreich folgende Forschungsfragen in drei Blöcken. Der erste Block hinterfragt die Zusammenhänge für die Volksschule allgemein. Der zweite Block bezieht sich auf die Annahme, dass in der Unterstützung von Schulen in sozial herausfordernder Lage die Klassengröße für die Steuerung ein Parameter sein könnte und setzt daher einen Fokus auf Schulen mit und ohne herausfordernde Lage. Der dritte Block bezieht sich auf die Annahme, dass Reduktionen von Klassengrößen im System nicht gleichermaßen alle Schulen treffen würden, sondern nur solche, die über einer zu reduzierenden Teilungszahl lägen, und unterscheidet daher Schulen nach der Prävalenz großer Klassen. Die Forschungsfragen sind:

- Zusammenhänge im gesamten System
 - F1: In einer Querschnittsbetrachtung zeigt sich, dass bei Kontrolle anderer Faktoren, wie z. B. Herkunftsmerkmale der Schülerinnen und Schüler und Standortfaktoren der Schule, in Schulen mit kleineren Klassengrößen Mathematikkompetenzen besser ausfallen.
 - F2: Ändert sich in einer Volksschule die mittlere Klassengröße zwischen zwei Erhebungsjahren, hängt die Änderung der mittleren Mathematikkompetenzen der Schulen zwischen den Erhebungen negativ mit der Änderung der Klassengröße zusammen, d. h., finden sich kleinere Klassen, geht dies mit tendenziell besseren Leistungen einher.
- Zusammenhänge spezifisch für Schulen in sozial herausfordernder Lage
 - F3: Der Zusammenhang zwischen Klassengröße und Mathematikkompetenzen ist für Volksschulen mit mehr Kindern mit niedrigerem sozioökonomischen Status stärker ausgeprägt, d. h., Schulen in benachteiligten Lagen profitieren stärker von einer Reduktion der Klassengröße.
- Zusammenhänge differenziert nach Schulstrukturen
 - F4: Der Vorteil zeigt sich stärker bei Änderungen der Klassengröße von (oder hin zu) sehr groß (für Österreich: >20) auf weniger groß als bei Änderungen von

(oder hin zu) mittel oder klein (für Österreich: >15) auf noch kleiner. Die getrennte Untersuchung von Schulen mit tendenziell größeren Klassen soll Aufschluss auf die Effekte geben, die einträten, würden die Regeln für maximale Klassengrößen geändert. Solche Änderungen würden nur bei Schulen schlagend, deren Klassen groß sind und über dem eventuellen neuen Maximum lägen.

- F5: Nicht Klassengröße ist entscheidend, sondern die Nähe zu einer idealen Größe von 20. Diese Fragestellung wird durch die Berechnungen zu F4 mituntersucht.

Da sich in Österreich die Mathematikkompetenzen zwischen dem ersten (2013) und dem zweiten Erhebungsjahr (2018) systemweit um 18 Punkte, d. h. ca. 18 % einer Standardabweichung, verbessert haben (Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Bildungswesens [BIFIE], 2019a), werden Unterschiede innerhalb einer Schule neben anderen Änderungen der Komposition insbesondere um diesen systemweiten Trend kontrolliert, d. h., als positiver Effekt der Klassengröße werden nur Effekte berücksichtigt, die über den Einfluss der anderen Faktoren und der allgemeinen Leistungssteigerung hinausgehen.

4 Datengrundlage und deskriptive Statistik

4.1 Datenquellen

Grundlage dieser Studie sind zum einen die Überprüfungen der Bildungsstandards in Mathematik, 4. Schulstufe Volksschule, in den Jahren 2013 und 2018 (BIFIE, 2019a; Schreiner & Breit, 2014) und – als administrative Quelle – die Gesamtevidenz der Schülerinnen und Schüler (§ 6 Bildungsdokumentationsgesetz, hier als BilDok bezeichnet), ein Ausschnitt der Daten der Schulstatistik (Statistik Austria, 2014). Letztere Datenquelle erlaubt es, Klassengrößen rückblickend für die gesamte Zeit von Schulstufe 1 bis Schulstufe 4 zu erfassen und aggregiert in das Model einzubringen. Die beobachtete Größe hat also langfristig gewirkt, bis in der 4. Schulstufe die Kompetenzen erhoben wurden. In der Folge beziehen sich alle Jahresangaben auf die Erhebungszeitpunkte der BIST-Ü, jeweils gegen Ende des Schuljahrs. Die Angaben 2013 bzw. 2018 beziehen sich somit auf die Schuljahre 2012/13 und 2017/18 in der BilDok. Alle Datenquellen können, mittels eines einheitlichen Schulindikators, auf Schulebene miteinander verknüpft werden. Damit liegen für alle berücksichtigten Schulen verknüpfbare Daten zu Mathematikkompetenzen in zwei Erhebungsjahren (2013, 2018) und zur Schulstatistik für alle Jahre (hier relevant 2009 bis 2018) vor. Der dadurch mögliche Längsschnitt der BIST-Ü bezieht sich auf die Schulen als Organisationen, aber nicht auf Schülerinnen und Schüler als Individuen. Auch BIST-Ü kann mit der BilDok nur auf Ebene der Schulen verknüpft werden. Allerdings erlaubt die BilDok eine Verknüpfung der Schülerinnen und Schüler über die Zeit, was es ermöglicht, pro Schülerin bzw. Schüler eine Historie der Klassengröße seit der ersten ermöglicht, zu ermitteln.

Die Datenbasis zur BIST-Ü umfasst die Schülerinnen und Schüler am Ende der 4. Schulstufe, die eine Schule mit auf Dauer verliehenem Öffentlichkeitsrecht besuchen. Aus-

genommen sind außerordentliche Schülerinnen und Schüler, Schülerinnen und Schüler, die in Mathematik nach ASO-Lehrplan unterrichtet werden, bzw. Schülerinnen und Schüler mit einer Körper- oder Sinnesbehinderung (Schreiner & Wiesner, 2019). Die BilDok umfasst alle Schülerinnen und Schüler, unabhängig von Schulform, Status oder Förderbedarf.

4.2 Definition der Zielgruppe und Auswahl der Analysedaten

Für diese Studie wird aus der Vollpopulation der Schulen mit Kindern der Schulform Volksschule laut Gesamtevidenz in den Jahren 2013 und 2018 eine Teilpopulation als Zielgruppe bestimmt. Diese Zielgruppe wird dann aufgrund von Aspekten der Datenverfügbarkeit und -konsistenz weiter reduziert. Abschließend wird auf die Analysedaten reduziert, wozu Schulen ausgeschlossen werden, bei denen Einflüsse auf die Änderungen vermutet werden können, die statistisch in diesem Design nicht kontrolliert werden können. Das Vorgehen und die entsprechenden Anzahlen von Schulen, Klassen sowie Schülerinnen und Schülern ist im Anhang 1 im Detail dokumentiert.

Ziel der Auswahl für die Analysedaten ist es, Schulen auszuschließen, bei denen die Klassenbildung und der Ablauf des Unterrichts durch strukturelle Faktoren beeinflusst ist bzw. sein könnte, die nicht durch das angewendete Model kontrolliert werden können. Insbesondere sollen die Analysedaten auf Schulen reduziert werden, die zwischen den beiden Erhebungsjahren keine strukturellen Veränderungen zeigen, die Änderungen des Zusammenhangs zwischen Klassengröße und Leistung an der Schule unkontrolliert überlagern würden.

Zusammengefasst lässt sich die Zielgruppe dieser Studie wie folgt definieren: Regulär unterrichtete Volksschülerinnen und -schüler der 4. Schulstufe in Schulen, die in beiden Erhebungsjahren mehr als fünf dieser Schülerinnen und Schüler geführt haben, nicht in Mehrstufenklassen unterrichten und keine schulinterne Organisation wählen, die zu stark unterschiedlich großen Klassen führt. Damit bezieht sich die Zielgruppe auf 81 % der Schülerinnen und Schüler der Vollpopulation.

Aus dieser Zielgruppe wurden aufgrund von Datenverfügbarkeit Schülerinnen und Schüler in Schulen ausgeschlossen, für die nicht sowohl Daten für beide Jahre in der BilDok als auch in der BIST-Ü verfügbar waren. Zusätzlich wurden aus inhaltlichen Überlegungen Schülerinnen und Schüler in Schulen ausgeschlossen, deren schulische Rahmenbedingungen sich in Bezug auf außerordentliche Schülerinnen und Schüler bzw. Inklusion zwischen den Erhebungen deutlich geändert haben. Nach Ausschluss dieser Schulen umfassen die Analysedaten der Studie 1.716 Schulen (57 % der Schulen, Basis 2018) mit 6.016 Klassen (ca. 60 % der Klassen) und ca. 70 % der Volksschülerinnen und -schüler der Vollpopulation.

Aus Sicht der Datenbasis der BIST-Ü stellt sich das Mengengerüst der Analysedaten wie folgt dar. Die Schülerinnen und Schüler in der Zielgruppe umfassen jeweils ca. 80 % der in der BIST-Ü getesteten Schulkinder. Mit der Reduktion auf das Schulpanel verbleiben

ca. 70 % der BIST-Ü-Vollpopulation bzw. 87 % der Zielgruppe in den Analysedaten. Die Daten umfassen für das Erhebungsjahr 2013 51.501 Schülerinnen und Schüler und für das zweite Erhebungsjahr 51.582 Schülerinnen und Schüler.

Da die BIST-Ü als Vollerhebung durchgeführt wurde, liegt dieser Studie keine Stichprobe zugrunde. Daher werden auch keine Stichprobengewichte verwendet und keine Stichprobenfehler berechnet. Einzig, um die Nichtteilnahme (Non-Response) bspw. wegen Krankheit auszugleichen, werden Gewichte berechnet, die auf Basis der klassenspezifischen Non-Response-Rate den Ausfall von Schülerinnen und Schülern kompensieren (George, Oberwimmer & Itzlinger-Bruneforth, 2016). Da nahezu eine Vollerhebung vorliegt, sollte die Interpretation der Ergebnisse anhand der Signifikanz nur mit Vorbehalt erfolgen: Es werden zwar Standardfehler angegeben, die MPLUS ausgibt, diese beziehen sich aber auf den Mess- und Imputationsfehler (erfasst mit den *Plausible Values and Multiple Imputations*) und eine theoretische Verallgemeinerung der vorhandenen Volldaten des Panels auf eine fiktive unendliche Superpopulation (Broscheid & Gschwend, 2003). Da die Standardfehler jeweils sehr klein ausfallen, schreibt eine Interpretation auf ihrer Basis vielen signifikanten Ergebnissen fälschlicherweise inhaltliche Relevanz zu. Daher wird in dieser Auswertung Signifikanz im Sinne einer Verallgemeinerung auf eine Superpopulation nur als Nebenkriterium für die Beurteilung der Ergebnisse verwendet.

Für die Interpretation der Daten ist vielmehr relevant, ob das Schulpanel (Analysedaten) systematisch von der Zielgruppe abweicht. Dies wird weiter unten in den deskriptiven Statistiken mituntersucht. Die Prüfung der Fragestellungen bezieht sich auf die Vollerhebung der Analysedaten, d. h. auf alle Schulen der Zielpopulation in Österreich mit vergleichbarer Situation in den Jahren 2013 und 2018.

4.3 Definition und Erfassung der mittleren Klassengröße

Für diese Studie könnten verschiedene Definitionen und Operationalisierungen von Klassengröße verfolgt werden. Dazu zählen ...

- *Aktuelle Klassengröße eines Jahrs laut den BIST-Ü-Erhebungen*
- *Aktuelle Klassengröße eines Jahrs laut der BilDok*
- *Mittlere Klassengröße der Schule eines Jahrs laut den BIST-Ü-Erhebungen*
- *Mittlere Klassengröße der Schule eines Jahrs laut der BilDok*
- *Längsschnittliche Klassengröße der Schule über vier Jahre laut der BilDok*

Eine detaillierte Darstellung und Diskussion dieser Varianten findet sich in Anhang 2.

Für diese Studie wurde die längsschnittliche Klassengröße der BilDok verwendet, da diese die Situation über die gesamten Volksschuljahre erfasst und langfristige Auswirkungen der Klassengröße auf die über vier Jahre gebildeten Mathematikkompetenzen erfassen kann. Diese wird wie folgt gebildet: Die Schulstatistik erlaubt, mit der persistenten Personen-

kennung laut Bildungsdokumentationsgesetz die Schullaufbahn aller Schülerinnen und Schüler von der ersten bis zur vierten Klasse Volksschule nachzuvollziehen. Das ermöglicht es, für jedes Kind die Klassengrößen der Klassen, die es bis zur 4. Schulstufe besucht hat, zu bestimmen und dann zu mitteln. Damit ergibt sich auf Individualebene die mittlere Klassengröße, in der das Kind auf den vier Schulstufen unterrichtet wurde. Die längsschnittliche Klassengröße für jedes Kind aus der BilDok lässt sich nicht individuell mit den BIST-Ü-Erhebungen verknüpfen. Für jede Schule wird darum für die 4. Schulstufe der Schulmittelwert der längsschnittlichen Klassengröße der Schülerinnen und Schüler ermittelt. Die längsschnittliche Klassengröße der BilDok lässt sich über die Schulkennung mit den Daten der BIST-Ü verknüpfen. Der Vorteil dieser Definition ist, dass nicht nur die Klassengröße der 4. Schulstufe berücksichtigt wird, sondern alle Schulstufen. Das ist insofern wichtig, als die gemessenen Mathematikkompetenzen im Sinne der Bildungsstandards als kumulativer Output der gesamten Zeit in der Volksschule interpretiert werden müssen. Zudem könnte angenommen werden, dass die Klassengröße auf verschiedenen Schulstufen einen unterschiedlichen Einfluss hat (Hanushek, 2002), womit vorteilhaft ist, alle Schulstufen zu berücksichtigen. Bei der Berechnung der längsschnittlichen Klassengröße für jedes Kind werden Schuljahre mit Wiederholung entsprechend geringer gewichtet, Schulwechsel werden berücksichtigt, d. h., dann werden die Werte aller vom Kind besuchten Schulen zusammengeführt.

Die mittlere Klassengröße bezieht sich immer auf alle Schülerinnen und Schüler (4. Schulstufe), inklusive jener mit außerordentlichem Status (ao. d. h., sie können dem Unterricht aufgrund der Sprache nicht folgen) bzw. mit sonderpädagogischem Förderbedarf (SPF). Es ist wichtig, bei der Klassengröße alle Schülerinnen und Schüler zu berücksichtigen, auch die, die von der Kompetenzmessung (siehe oben) ausgeschlossen sind.

In Österreich war die Klassengröße für die Volksschule bis 2018, dem zweiten Erhebungszeitpunkt dieser Studie, auf maximal 25 Kinder pro Klasse gedeckelt und durfte 10 nicht unterschreiten (§ 14 SchUG i. d. F. 31.8.2018). Seitdem ist die konkrete Klassengröße zwar nicht mehr gesetzlich gedeckelt, sondern liegt in der Autonomie des Schulleiters, allerdings richtet sich das zugeteilte Kontingent an Lehrpersonenwochenstunden pro Schule weiterhin nach der theoretischen Berechnungsgrundlage von 25 Schülerinnen und Schülern pro Schule (§ 8 SchUG i. d. F. 10.1.2024). Laut BilDok führen 2018 1.680 Schulen der Analysedaten genauso viele Klassen, wie sich aus der Regelung des SchUG ergeben würden, 29 Schulen führen mehr Klassen und nur 7 Schulen führen weniger Klassen. D. h., Schwankungen in der Klassengröße ergeben sich, vermittelt durch die gesetzliche Klassenschülerzahl nach § 14 SchUG, weitestgehend direkt aus Schwankungen in der Anzahl der Schülerinnen und Schüler. Somit kann man diese Studie als natürliches Experiment betrachten, das kaum von anderen Faktoren, die mit der Leistung zusammenhängen, abhängig ist.

Zwischen den Erhebungen 2013 und 2018 sind Mittelwert und Streuung der Klassengröße nahezu konstant, d. h., den Änderungen der Klassengröße für die einzelnen Schulen liegt kein allgemeiner Trend zugrunde, sondern es handelt sich um eine Fluktuation der einzelnen Schulen.

Alle weiteren statistischen Angaben im folgenden Text beziehen sich auf die längsschnittliche Klassengröße, ohne diese zu benennen, und auf die Schulen der Analysedaten.

4.4 Schwankungen in der Klassengröße zwischen 2013 und 2018

Bei 609 der Schulen (35,5 %) ergeben sich Schwankungen der Klassengröße um 3 oder mehr Schülerinnen und Schüler zwischen den Erhebungszeiträumen, wobei sich die Anzahl mit Zunahme von 2013 auf 2018 (294) und die Anzahl mit Abnahme (315) nahezu die Waage halten (Tabelle 1). Die Veränderung der Klassengröße korreliert mit der Änderung der Gesamtzahl der Schülerinnen und Schüler (4. Schulstufe) in den Schulen, d. h., Klassen werden größer bei steigender Schulgröße und umgekehrt. Deutlich seltener steigt die Klassengröße trotz einer Abnahme der Schulgröße, weil die Zahl der Klassen sinkt.

Änderungen der Klassengröße finden sich im ganzen Spektrum der Klassengrößen in der Ausgangssituation 2013, d. h., es finden sich sowohl Schulen mit sehr kleinen Klassen in 2013, die nun große Klassen zeigen, als auch umgekehrt. Die meisten Schulen zeigen zwar Änderungen der Klassengröße, aber die Anzahl der Klassen hat sich zwischen den beiden Erhebungen nicht geändert (1.349), während in 175 Schulen sich die Struktur durch Öffnung einer oder mehrerer neuer Klassen – bzw. in 192 durch Schließung – geändert hat (Abbildung 1).

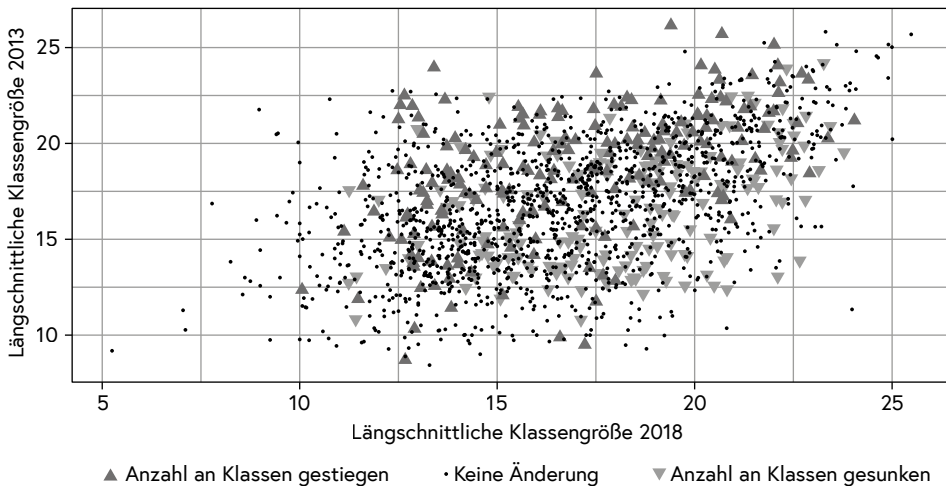


Abbildung 1: Vergleich der längsschnittlichen mittleren Klassengröße der Schulen (N = 1.716) 2013 und 2018.

4.5 Instrumente und Variablen

Mit Ausnahme der vorstehend eingeführten Schulebenenvariable „längsschnittliche Klassengröße“ stammen alle weiteren Variablen der Analyse aus den Daten der BIST-Ü 2013

Tabelle 1: Absolute Veränderung der Klassengröße und Änderung der Schulgröße an den Schulen (2013, 2018)

Änderung	Klassengröße hat zugenommen			Klassengröße hat abgenommen			Gesamt	%
	Gesamt	Schule: mehr Schü- rinnen und Schüler	Schule: weniger Schü- lerinnen und Schüler	Gesamt	Schule: mehr Schü- rinnen und Schüler	Schule: weniger Schü- lerinnen und Schüler		
[0,1]	231	154	77	211	100	111	442	25,8
(1,3]	321	243	78	344	108	236	665	38,8
(3,5]	178	141	37	190	45	145	368	21,4
(5,7]	83	64	19	80	18	62	163	9,5
(7,9]	28	24	4	29	4	25	57	3,3
(9,11]	4	3	1	12	4	8	16	0,9
(11,13]	1	1	0	4	0	4	5	0,3
(> 13]	0	0	0	0	0	0	0	0,0
Gesamt	846	630	216	870	279	591	1.716	100,0
Davon Änderungen größer als 3	294	233	61	315	71	244	609	35,5

und 2018. Die Almanache zu den BIST-Ü (BIFIE, 2017, 2019b) geben jeweils eine Übersicht der eingesetzten Items des Schülerinnen- und Schülerfragebogens. Für diese Analysen werden jeweils imputierte Datensätze verwendet, die im Rahmen der Forschungsdatenbibliothek des Instituts des Bundes für Qualitätssicherung im österreichischen Schulwesen (IQS) zur Verfügung stehen. Es wurde eine multiple Imputation mit MICE (10 Datensätze) eingesetzt, wobei die 10 *plausible values* für die Leistungsdaten und die Imputationen für die Kontextdaten in einen Prozess miteinander abgestimmt geschätzt wurden (Robitzsch, Pham & Yanagida, 2016). Die Daten zu Schülerinnen und Schülern beruhen – mit Ausnahmen – auf Angaben der Kinder, bei der Imputation der Variablen wurden dann auch Angaben der Eltern berücksichtigt. Daten aus dem Elternfragebogen wurden nur verwendet, wenn diese Angaben von den Schülerinnen und Schülern nicht erhoben wurden. Diese Präferenz für Angaben der Schülerinnen und Schüler ergibt sich aus der geringen Rücklaufquote des Elternfragebogens.

Folgende Variablen bzw. Kovariate werden verwendet.

- Ebene der Schülerinnen und Schüler:
 - *mathe*: Mathematikkompetenz. Die ursprünglichen Scores der BIST-Ü-Skala (mit Mittelwert für 2010 von 500) wurden, um die Interpretation der Parameter zu erleichtern, für diese Studie transformiert, sodass der Mittelwert der Schülerinnen und Schüler in 2013 0 Punkten entspricht, die SD wurde nicht verändert. Somit ergibt sich für 2013 MW = 0, SD = 98,4 und 2018 MW = 17,8, SD = 97,1, was die in BIFIE (2019a) beschriebene Verbesserung der Kompetenzen auch für die Analysedaten abbildet.
 - *jahr*: Erhebungszeitraum mit dichotomer Kodierung (0 = 2013, 1 = 2018).
 - *geschlecht*: dichotome Kodierung des Geschlechts mit Mädchen als Referenzgruppe und Knaben mit geschlecht = 1 als Vergleichsgruppe.
 - *buecher*: Anzahl der Bücher zuhause, ordinal skaliert von –2 bis 2 mit –2 = „0 bis 10 Buecher“, –1 = „11 bis 25 Buecher“, 0 = „26 bis 100 Buecher“ (Referenzgruppe), 1 = „101 bis 200 Buecher“ und 2 = „ueber 200 Buecher“.
 - *hisei*: der höchste sozioökonomische Status der Eltern der Schülerinnen und Schüler, kodiert nach Ganzebooms (2010) ISEI-08 auf Basis der Angaben zu den Berufen der Eltern. Allerdings wird der Mittelwert des hisei (ursprünglich in 2013 = 47,3) für die Analyse auf 0 gesetzt. Somit ergibt sich für 2013 MW = 0, SD = 20,7 und 2018 MW = 2,0, SD = 21,0.
 - *mig*: Migrationshintergrund der Schülerinnen und Schüler aufgrund der Angabe der Schülerinnen und Schüler zum Geburtsland der Eltern. Kinder mit zumindest einem in Österreich geborenen Elternteil werden als einheimisch kategorisiert (mig = 0), alle anderen als Kinder mit Migrationshintergrund (mig = 1).
 - *sprache*: Muttersprache des Kindes, d. h. die Sprache(n), in der das Kind zuerst sprechen gelernt hat, wobei auch Mehrfachnennungen möglich sind. Die Variable ist dichotom kodiert, mit Kindern, die Deutsch als die oder eine Muttersprache nennen, als Referenzgruppe und den Kindern ohne deutsche Muttersprache (sprache = 1) als Vergleichsgruppe.

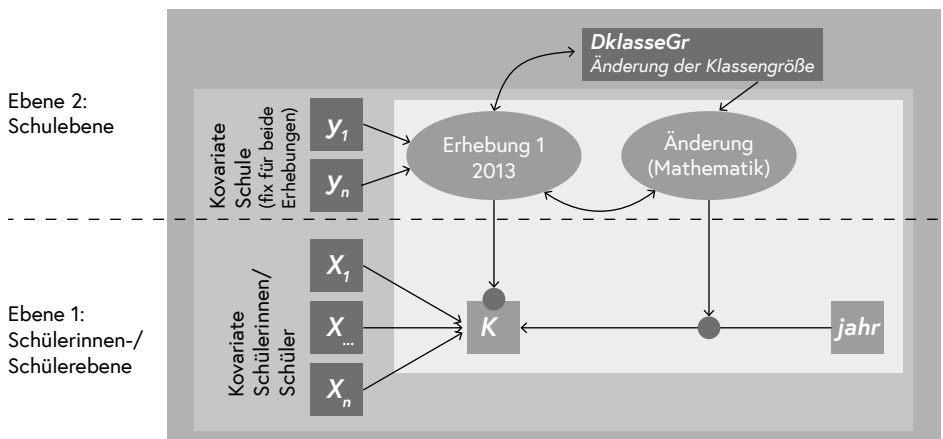
- *bildung*: Höchste Bildung der Eltern, ordinal kodiert mit -1 = „max. Pflichtschule“, 0 = „Berufsschule, Meisterausbildung, Schule fuer Gesundheitspflege“ (Referenzgruppe), 1 = „max. Schule mit Matura“, 2 = „Universität, Fachhochschule oder Pädagogische Akademie bzw. PH“.
- Schulebene
 - *land/stadt*: Zwei dichotome Variablen zur Besiedlungsdichte, d. h. zur Unterscheidung von Schulen in dünn besiedelten Gebieten (überwiegend ländlich) ($land = 1$), Schulen in Gebieten mittlerer Besiedlung ($land = 0$, $stadt = 0$) und Schulen in dicht besiedelten Gebieten (überwiegend städtisch).
 - *DklasseGr* (nur im Längsschnitt): Änderung der längsschnittlichen Klassengröße zwischen den Erhebungsjahren.
 - *klasseGr* (nur im Querschnitt): Längsschnittliche Klassengröße der Schülerinnen und Schüler der Schulen. Datenbasis Bildok (siehe oben), bezogen auf alle Schülerinnen und Schüler und nicht auf die Anzahl der BIST-Ü-Teilnehmerinnen und -Teilnehmer. Die Schülerinnen und Schüler innerhalb einer Schule bekommen je nach Erhebungsjahr die entsprechenden Werte zugewiesen (siehe oben).
 - *sesSchJ* (nur im Querschnitt): Zur Erfassung des Kompositionseffekts (Dumont, Neumann, Maaz & Trautwein, 2013) wird die Kovariate „Sozialstatus“ auf Ebene der Schülerinnen und Schüler gebildet und auf Schulebene gemittelt. Der Sozialstatus eines Kindes ergibt sich, angelehnt an George und Robitzsch (2014), als Mittelwert der z-standardisierten Variablen *hisei*, *buecher* und *bildung*.

5 Methoden

5.1 Berechnungen und Modellierung

Datenaufbereitung und deskriptive Auswertungen wurden mit der Software R, insbesondere mit dem Paket BIFIEsurvey (BIFIE, Robitzsch & Oberwimmer, 2022) durchgeführt. Zur Beantwortung der Fragestellungen wurden Mehrebenenmodelle mit Random Slopes mit MPLUS (Version 7.31) geschätzt, wobei das R-Paket MplusAutomation (Hallquist & Wiley, 2018) zur Unterstützung verwendet wurde. Fragestellung 1 wird mit einer Querschnittsanalyse zu den einzelnen Erhebungsjahren getrennt beantwortet. Diese Fragestellung wurde in die Analysen aufgenommen, um evtl. zeigen zu können, wie sich die inhaltliche Frage im Querschnitt anders beantwortet als aus der Längsschnittperspektive. Die Modellierungen zu den Fragestellungen 2 bis 5 folgen dem von Schmid et al. (2019) für das BIST-Ü-Schulpanel vorgeschlagenen Ansatz (siehe Abbildung 2). Hierbei handelt es sich um Slope-as-Outcome-Modelle (Schwippert, 2002), bei denen die kontrollierenden Variablen auf den Bevölkerungsmittelwert zentriert wurden. Dabei wird für jede Schule die Änderung der Mathematikleistungen ihrer Schülerinnen und Schüler zwischen den Erhebungen linear modelliert (*within*). Der Unterschied der Leistung in Abhängigkeit vom Erhebungsjahr (schulspezifische *random slope*) geht als erklärende Variable auf die nächsthöhere Analyseebene (between:

Schule unabhängig vom Erhebungsjahr) ein. Hier wird zwischen den Schulen analysiert, inwieweit Änderungen der im Längsschnitt erfassten durchschnittlichen Klassengröße zwischen 2013 und 2018 mit den Änderungen der Leistung zusammenhängen. Somit werden hier Leistungsänderungen von Schulen im Mehrebenenmodell analog behandelt, wie längsschnittliche Leistungsänderungen von Schülerinnen und Schülern im Mehrebenenmodell bei Ditton (1998). Wie bereits erwähnt, erlaubt die Kontrolle der Änderungen über die Zeit auf der *Within*-Ebene, d. h. jeweils getrennt für die Schulen, dass weit mehr Faktoren implizit kontrolliert werden, als im Modell explizit spezifiziert werden, z. B. Faktoren wie Nachbarschaft, vom Erhalter bereitgestellte Ressourcen, Zubringer-Kindergärten, die alle für die Schulen meistens zwischen den Erhebungsjahren recht stabil sind.



Anmerkungen: K = Mathematikkompetenz. Kovariate Schülerinnen und Schüler: geschlecht, buecher, hisei, mig, sprache und bildung. Kovariate Schule: land/stadt.

Abbildung 2: Schematische Darstellung des Mehrebenenmodells nach Schmid et al. (2019).

Im Model wird die Ebene Klasse nicht explizit geführt. Das hat zwei Gründe. Da verschiedene Kohorten verglichen werden, für die immer ganz neue Klassen gebildet werden, sind Klassen keine Einheiten, die zwischen Erhebungen verglichen werden können. Auf Unterschiede in der Größe zwischen Klassen eines Jahrs zielt die Analyse nicht ab. In Schulen, die unterschiedlich große Klassen führen, sollte es Gründe für diese Unterschiede geben, die hier nicht kontrolliert werden können. Es ist davon auszugehen, dass, wo immer aktiv steuernd Klassen unterschiedlicher Größe gebildet werden, dies aufgrund besonderer Bedürfnisse erfolgt.

5.2 Nullmodell, Pseudo-R² und Effektstärke

Um die inhaltlichen Modelle zur Überprüfung der Fragestellungen beurteilen zu können, werden zuerst zwei grundlegende Modelle berechnet. Beim Nullmodell ohne kontrollierende Kovariate werden die Varianzen auf den verschiedenen Ebenen für die Mathematikkompetenz geschätzt. Diese werden zwar nicht in Tabelle 2 aufgeführt, werden aber in folgenden Modellen herangezogen, um über ein Pseudo-R² (Lorah, 2018) abschätzen zu

können, wie viel Varianz durch die Modelle aufgeklärt wird. In einem zweiten Modell (Modell 0a) werden alle Kovariate, die bei der Beantwortung der Forschungsfragen kontrolliert werden sollen, d. h. alle außer denen zu Klassengröße, einbezogen. Es zeigt sich sowohl beim Querschnittsmodell als auch im Längsschnittmodell, dass die kontrollierenden Kovariate bereits ca. 22 % der ursprünglichen Varianz erklären. Die kontrollierenden Kovariate können jeweils Tabelle 2 (Querschnitt) und Tabelle 3 (Längsschnitt) entnommen werden. Für jedes weitere Modell wird die zusätzlich erklärte Varianz abgeschätzt, indem das Pseudo- R^2 des Modells mit dem Pseudo- R^2 des vorherigen verglichen wird.

Modell 0a zu den Jahren 2013 und 2018 zeigt, dass sich die kontrollierenden Kovariate in beiden Erhebungsjahren in ihrem Einfluss auf die Mathematikleistungen ähneln. Insbesondere auf dem Between-Level ergeben sich wenige Änderungen, während sich im Within-Level leichte Verschiebungen zwischen den Erhebungszeiträumen ergeben: Der Einfluss des Geschlechts nimmt zu, wie auch in BIFIE (2019a) berichtet, und es ergibt sich eine Verschiebung der Parameter bei SES und Herkunft auf der Ebene der Schülerinnen und Schüler. Trotzdem scheinen die Strukturen 2013 und 2018 ausreichend ähnlich, um ein Modell zu begründen, das beide Erhebungszeitpunkte in einem Modell (Längsschnitt: Tabelle 3) zusammenfasst, bei dem die Parameter für Variablen der Schülerinnen und Schüler zusammen mit je einem Parameter geschätzt werden.

Ein bisher ungelöstes Problem bei der Interpretation von Random-Slope-Modellen ist es, dass keine klaren Maßzahlen für die Effektgröße existieren (Lorah, 2018). Dies ist auch insofern problematisch, als damit die Chance für neue Ergebnisse sinkt, in Metastudien berücksichtigt zu werden. Lorah (2018) schlägt vor, die standardisierten Kovarianz-Terme, eine Korrelation, im Sinne von Effektstärken zu interpretieren. Hier werden somit die von MPLUS ausgegebenen Kovarianz-Terme für diesen Zweck herangezogen.

Zudem wird eine Effektstärke aus den Parametern des linearen Modells abgeleitet, wie groß eine Veränderung der mittleren Leistung wäre, würde eine Klasse um fünf Kinder kleiner bzw. größer. Die Wahl der Zahl der Schülerinnen und Schüler, die als Veränderung angenommen wird, bezieht sich somit auf ein fiktives Treatment, das auszuwählen in der Freiheit des Autors liegt. Diese Entscheidung kann direkt die zu berichtende Effektstärke vergrößern oder verkleinern und ist von ihrer Natur her willkürlich. Die fiktive „Treatment“-Größe von 5 wählt der Autor aus mehreren Gründen: In den Daten schwanken nur wenige Schulen zwischen den Erhebungen in der Klassengröße um mehr als 5 Schülerinnen und Schüler, daher sollte aus dem linearen Modell nicht auf viel größere Änderungen geschlossen werden. Aus Sicht des Systems ist eine Implementierung einer Maßnahme, die Klassengrößen um mehr als 5 ändert, höchst unwahrscheinlich. So gab es vor der Reduktion der Klassenschülerzahlen im Schuljahr 2007/08 nahezu keine Klassen mit mehr als 30 Schülerinnen und Schülern, d. h., die Reduktion auf maximal 25 ging mit einer angestrebten Reduktion der größten Klassen um 5 Schülerinnen bzw. Schüler einher (Rechnungshof Österreich, 2011).

Das Kriterium für die praktische Bedeutsamkeit entnimmt der Autor Salchegger und Bruneforth (2024), welches sich an einer Effektstärke von .13 orientiert und hier 13 Punkten

Tabelle 3: Modellparameter Längsschnittmodelle, Teil 1 (2013, 2018)

Ebene	Parameter	Modell LOa mit kontrollieren- den Kovariaten, Änderung der Klassengröße	(SE)	Modell L1	(SE)	Modell L2 (hoher SES)	(SE)	Modell L2 (mittlerer SES)	(SE)	Modell L2 (niedriger SES)	(SE)
Within: Mathe	BILDUNG	19,06	(0,35)	19,06	(0,35)	20,79	(0,70)	19,03	(0,54)	16,85	(0,71)
	BUECHER	14,29	(0,27)	14,29	(0,27)	14,42	(0,51)	14,06	(0,42)	14,02	(0,51)
	GENDER	19,30	(0,55)	19,31	(0,55)	23,14	(1,03)	17,55	(0,83)	18,17	(1,05)
	HISEI	0,63	(0,02)	0,63	(0,02)	0,59	(0,03)	0,62	(0,02)	0,66	(0,03)
	MIG	-3,90**	(1,13)	-3,90**	(1,13)	-8,47	(2,18)	-7,86	(1,78)	5,52**	(1,87)
	SPRACHE	-18,88	(1,01)	18,88	(1,01)	-13,32	(1,91)	-18,47	(1,48)	-24,46	(1,93)
	(Residual-) Variances	6.660,51	(31,08)	6.660,35	(32,91)	6.643,75	(65,55)	6.611,19	(45,94)	6.750,60	(64,12)
Between: Change:	LAND	9,96	(1,55)	9,94	(1,55)	3,92#	(3,12)	6,53**	(2,09)	23,50	(3,24)
	STADT	-5,52**	(2,06)	-5,50**	(2,07)	-1,99#	(3,20)	-9,71*	(3,88)	-9,57*	(3,93)
	Intercepts	-1,74#	(1,23)	-1,71#	(1,23)	26,90	(2,23)	-0,251#	(1,74)	-35,31	(2,48)
	(Residual-) Variances	563,34	(27,11)	563,38	(27,19)	509,42	(50,37)	491,09	(36,85)	607,14	(61,4)
	DKLASSEG	-	-	-0,70*	(0,27)	-1,60**	(0,50)	-0,41#	(0,39)	-0,26#	(0,57)
	Intercepts	17,86	(0,93)	17,09	(0,93)	14,31	(1,64)	19,06	(1,38)	16,49	(1,92)
	(Residual-) Variances	905,10	(47,6)	899,56	(47,63)	643,16	(71,81)	971,57	(71,74)	1.012,58	(102,76)
Pseudo- R ²	Vergleich mit Nullmodell	21,50		21,50		18,92		18,19		20,68	
	Vergleich mit vor- herigem Modell	-		0,002		0,05		-0,05		0,20	
Anmerkungen: N = 1.716 Schulen. Bei der Interpretation der Angaben zur Signifikanz ist zu bedenken, dass es sich bei den dargestellten Daten um eine Vollerhebung mit nicht randomisierten Ausschlüssen handelt. Wie bei einer derart großen Stichprobe erwartbar sind nahezu alle Angaben mit 0.001 signifikant, mit Ausnahme der wie folgt markierten: **mittlere Signifikanz: zwischen 0.05 und 0.01, *niedrige Signifikanz: zwischen 0.01 und 0.05, # keine Signifikanz, > = 0.05.											

entspricht. Hierbei wird nach Kraft (2020) die Standardabweichung der Schülerinnen und Schüler berücksichtigt (2018 = 97,1) und nicht die Standardabweichung der Leistungsmittelwerte der höheren Ebenen (Schule oder Klasse). Da sowohl die Wahl der Veränderungszahl als auch die Festlegung einer Schwelle der Effektstärke eine Interpretation des Autors sein muss, seien Leserinnen und Leser eingeladen, die Beurteilung der Ergebnisse eventuell auf Basis ihrer eigenen Interpretation vorzunehmen. Es stellt sich allerdings die Frage, ob es nicht klüger wäre, der Aufforderung von Götz, Gosling und Rentfrow (2024, S. 2) zu folgen: „No more benchmarks.“

6 Ergebnisse

6.1 Fragestellung 1: Klassengröße und Leistung im Querschnitt

In den Modellen Q1 wird für die Jahre 2013 und 2018 getrennt untersucht, ob Klassengrößen in den Querschnittsdaten einen Beitrag leisten können, Unterschiede zwischen Schulen in der Mathematikleistung zu erklären und so Fragestellung 1 zu beantworten. Die Fragestellung muss negativ beantwortet werden. In der Querschnittsbetrachtung zeigt sich bei Kontrolle anderer Faktoren kein Zusammenhang zwischen mittlerer Klassengröße einer Schule und der Mathematikleistung. Der entsprechende Parameter in Tabelle 2 zeigt mit $-0,37$ (2013) bzw. $-0,66$ (2018) Punkten Unterschiede in der mittleren Leistung der Schule pro Schülerin bzw. Schüler. Der Unterschied in der Klassengröße verläuft zwar in die erwartete Richtung, ist in der Größe aber marginal: Würden Klassen um 5 Schülerinnen und Schüler verkleinert, würde es eine Leistungssteigerung von 2 (2013) bis 3,5 (2018) Punkten bei einer Standardabweichung von ca. 100 Punkten geben. Bei Hinzuziehung der Signifikanz als Hilfe bei der Interpretation ist das Ergebnis – trotz der großen Anzahl der untersuchten Schulen – nicht signifikant von null abweichend. Die Erweiterung des Modells geht entsprechend nicht mit einer besseren Erklärung der Varianz (Pseudo- R^2) einher.

6.2 Fragestellung 2: Klassengröße und Leistung im Längsschnitt

Im nächsten Schritt wird zur Überprüfung der Fragestellung 2 in Modell L1 die Änderung der Klassengröße zwischen 2013 und 2018 (DKLASSEG) als Schulvariable auf Ebene 2 (Between) hinzugenommen, um Leistungsänderungen der Schulen zwischen den Jahren zu erklären. Es zeigt sich ein geringfügiger negativer Zusammenhang zwischen Leistung und Änderung der Klassengröße (DKLASSEG). Dieser ist mit $p = 0,011$ mäßig signifikant. Damit bestätigt sich die Vermutung, dass die Längsschnittstudie im Vergleich zum Querschnitt einen genaueren Blick auf die Fragestellung erlaubt. Es lässt sich allerdings keine zusätzliche Varianzaufklärung beobachten. Die Leistungszunahme ist mit 0,7 Punkten je Klassenverkleinerung um eine Schülerin bzw. einen Schüler auch nur marginal. Obwohl durch die Längsschnittperspektive mehr Faktoren implizit kontrolliert werden, zeigt sich,

anders als in Fragestellung 2 angenommen, der Effekt zwar signifikant, aber nicht stärker als bei der Querschnittsperspektive. Würde eine Schule die Klassengröße um 5 Schülerinnen oder Schüler reduzieren, würde die mittlere Leistung um 3,5 Punkte (bei einer Standardabweichung von 100) steigen. Dies entspricht einer Effektstärke von 0,3. Zwar gilt die Aussage der Fragestellung 2, dass bei Verkleinerung der Klassen die Leistung steigt, aber der Größe des Effekts kann keine praktische Bedeutung zugemessen werden. Die als Effektstärke interpretierbare Korrelation zur Änderung der Klassengröße (DKLASSG) wird in MPLUS auf $-0,046$ geschätzt, was auf eine nicht einmal kleine Effektstärke hindeutet.

6.3 Fragestellung 3: Unterschiede der Effekte nach schulischer Komposition

Für die in den rechten Spalten in Tabelle 3 dargestellten Modelle wurden die 1.716 Schulen nach dem mittleren sozioökonomischen Status (SES) ihrer Schülerinnen und Schüler aufgeteilt: a) Schulen, die im Schnitt der Jahre 2013 und 2018 im ersten Quartil der Schulen nach SES liegen (niedriger SES: 429 Schulen, $N_{2013} = 13.376$, $N_{2018} = 13.027$), b) 858 Schulen mit mittlerem SES (Quartil 2 und 4, $N_{2013} = 23.580$, $N_{2018} = 23.716$) und c) Schulen mit hohem SES (429 Schulen, $N_{2013} = 14.545$, $N_{2018} = 14.839$). Mit dieser Trennung kann Fragestellung 3 untersucht werden. Fragestellung 3 kann nicht bestätigt werden, Schülerinnen und Schüler in Schulen mit niedrigerem und mittlerem SES profitieren gar nicht von einer Reduktion der mittleren Klassengröße (DKLASSEG), der Effekt wird auf 0,3 bzw. 0,4 Punkte Änderung je Änderung der Klassengröße um eine Schülerin bzw. einen Schüler geschätzt und ist nicht signifikant. Vielmehr zeigt sich überraschenderweise ein der Annahme der Fragestellung entgegengesetztes Bild. Für Schulen mit hohem SES hat eine Verkleinerung von Klassen einen schwach signifikanten, positiven Zusammenhang mit Leistungen. Sinkt die Klassengröße um 5 Schülerinnen und Schüler, liegen die mittleren Leistungen um 8 Punkte höher, bei einer SD der Leistung von 96 Punkten. Dies entspricht einer Effektstärke von 0,8 und kann nicht als praktisch bedeutsam beurteilt werden.

Entgegen den Hoffnungen, dass eine gezielte Verringerung von Klassenstärken in sozial schwachen Schulen einen Beitrag zur Steigerung der Chancengerechtigkeit leisten könnte, stellt sich vielmehr die Frage, ob der Nutzen kleinerer Klassen in Schulen mit höherem SES einen Beitrag zur Chancenungleichheit darstellt. Dies wäre der Fall, wenn Volksschulen mit höherem SES auch kleinere Klassen hätten. Dies bestätigt sich allerdings 2013 und 2018 nicht, in den Schulen im oberen Quartil des mittleren Schul-SES liegen die Klassengrößen mit 19,2 Schülerinnen und Schülern um ca. 1,4 Schülerinnen und Schüler höher als in den Schulen mit niedrigem SES und um 1,8 höher als bei mittlerem SES. Trotzdem muss festgehalten werden, würden in Österreichs Volksschulen die Klassenschülerzahlen gesenkt, würden sich für Schülerinnen und Schüler in Schulen in benachteiligten Lagen keine positiven Effekte einstellen, während für Schülerinnen und Schüler in Lagen mit besseren sozioökonomischen Ausgangslagen leichte Leistungssteigerungen erwartet werden können. Die Chancengleichheit würde sinken, wobei die hohen Kosten ohnehin in keiner Relation zu dem geringen Nutzen stünden.

6.4 Fragestellung 4: Unterschiede der Effekte von Änderungen gruppiert nach Klassengrößen

Mit Fragestellung 4 soll geprüft werden, ob sich der etwaige Vorteil von verkleinerten Klassen vergleichsweise stärker zeigt, wenn Veränderungen weg von (oder hin zu) recht großen Klassen geschehen. Dazu werden die Schulen in zwei Gruppen geteilt: a) Schulen, in denen die Klassengröße trotz etwaiger Änderungen der Klassengröße zu beiden Erhebungszeitpunkten kleiner als 20 Schülerinnen und Schüler ist, also vor und nach der Änderung; b) Schulen, in denen die Klassengröße zu mindestens einem Erhebungszeitpunkt 20 Schülerinnen und Schüler oder mehr ist. Die Erwartung in Bezug auf diese Fragestellung ist, je größer die Klassen werden, desto stärker schlagen die Nachteile der Klassengröße für den Unterricht durch.

Tabelle 4: Modellparameter Längsschnittmodelle, Teil 2 (2013, 2018)

Ebene	Parameter	Modell L3 (Schulen mit kleineren Klassen)	(SE)	Modell L3 (Schulen mit größeren Klassen)	(SE)
Within: Mathe	BILDUNG	18,45	(0,49)	19,86	(0,57)
	BUECHER	13,91	(0,35)	14,80	(0,43)
	GENDER	17,76	(0,70)	21,46	(0,85)
	HISEI	0,62	(0,02)	0,65	(0,03)
	MIG	-5,80	(1,64)	-2,13#	(1,54)
	SPRACHE	-19,10	(1,40)	-18,21	(1,47)
	(Residual-)Variances	6.541,84	(41,55)	6.822,87	(54,77)
	LAND	9,67	(1,84)	8,51**	(3,09)
	STADT	-7,87*	(3,49)	-4,01#	(2,84)
	Intercepts	-2,59#	(1,52)	-0,43#	(2,10)
	(Residual-)Variances	575,12	(34,57)	539,30	(46,15)
Between: Change	DKLASSEG	-0,91*	(0,37)	-0,47#	(0,40)
	Intercepts	17,23	(1,14)	16,75	(1,61)
	(Residual-)Variances	923,67	(60,70)	860,14	(78,28)
Pseudo-R ²	Vergleich mit Nullmodell	26,21		17,45	
	Vergleich mit vorherigem Modell	0,09		-0,18	

Anmerkungen: N = 1.716 Schulen. Bei der Interpretation der Angaben zur Signifikanz ist zu bedenken, dass es sich bei den dargestellten Daten um eine Vollerhebung mit nicht randomisierten Ausschlüssen handelt. Wie bei einer derart großen Stichprobe erwartbar, sind nahezu alle Angaben mit 0.001 signifikant, mit Ausnahme der wie folgt markierten: **mittlere Signifikanz: zwischen 0.05 und 0.01, *niedrige Signifikanz: zwischen 0.01 und 0.05, # keine Signifikanz, > = 0.05.

Die Modelle L3 zu den beiden Gruppen sind in Tabelle 4 dargestellt. Es zeigt sich – ähnlich wie bei Fragestellungen 3 –, dass sich das gegenteilige Bild zur ursprünglichen Annahme zeigt. Entgegen der Intuition können Zusammenhänge zwischen Änderungen von Klassengröße und Leistung (DKLASSEG) nur in der Gruppe beobachtet werden, in der die Klassen sowohl vor als auch nach der Änderung zwischen den Erhebungen relativ klein waren, d. h. unter 20 Schülerinnen und Schülern. Für diese Gruppe würde eine Veränderung der Klassengröße um 5 Schülerinnen und Schüler mit einer Änderung der Leistung um 4,5 Punkte einhergehen. Diese Veränderung entspräche einer Effektstärke von knapp 0,05 und unterschreitet deutlich die Schwelle zur praktischen Bedeutsamkeit nach Salchegger und Bruneforth (2024).

Dieses überraschende Ergebnis zieht wichtige Konsequenzen nach sich. Da in Österreich die Anzahl der Klassen pro Schule und damit die Klassengrößen zumeist mechanisch über die Teilungszahlen gesteuert werden, würde eine Änderung der Teilungszahlen von jetzt 25 auf einen niedrigeren Wert nur in Schulen schlagend werden, in denen der Effekt kleinerer Klassen gar nicht auf die Leistungen wirken würde. Daher muss aus Sicht der vorgelegten Analyse von Maßnahmen zur Verkleinerung von Klassen abgeraten werden.

7 Diskussion

Die vorliegenden Ergebnisse zeigen für die Volksschule in Österreich, dass die mit einer Reduktion von Klassengrößen häufig verbundenen Erwartungen auf Qualitätsverbesserungen sich zumindest mit Blick auf Mathematikkompetenzen nicht bestätigen. Während die Querschnittsmodelle für die Jahre 2013 und 2018 trotz großer Stichprobe keine signifikanten Effekte zeigen (F1), weisen die Längsschnittauswertungen zwar auf signifikante Effekte hin (F2), jedoch sind diese nur sehr schwach ausgeprägt und praktisch ohne Bedeutung. Erst ab Änderungen der Klassengröße von über 10 Schülerinnen und Schülern kommen die Effekte in den Bereich praktische Bedeutsamkeit.

Von besonderer Bedeutung für die Bildungspolitik ist die Tatsache, dass Fragestellung 3, die Erwartung eines besonderen Nutzens der Reduktion von Klassengrößen für Schulen mit vielen sozial schwachen Schülerinnen und Schülern, nicht nur verworfen werden musste, sondern dass sich vielmehr ein wenn auch schwacher Nutzen bei Schulen mit Schülerinnen und Schülern mit eher günstigem sozioökonomischem Status zeigt. Eine gezielte Reduktion der Klassengröße für Schulen in herausfordernden Lagen wäre somit ohne Effekt auf die Kompetenzen der Schülerinnen und Schüler. Eine weitere Reduktion der Klassengröße für alle, d. h. unter die Klassenschülerzahl von 25, würde für Schulen in benachteiligten Lagen keine Effekte bringen, während für Schülerinnen und Schüler in Schulen mit besseren sozioökonomischen Ausgangslagen leichte Leistungssteigerungen zu erwarten wären. Die Chancengleichheit würde sinken, wobei die hohen Kosten ohnehin in keiner Relation zu dem geringen Nutzen stünden. Würde eine solche Maßnahme trotzdem in Betracht gezogen, evtl. dem Beispiel Hamburgs folgend, sollte genau geplant und

empirisch belegt werden, welche anderen Bedingungen angepasst werden müssen, damit diese Reduktion Wirkung zeigen könnte.

Irritierend ist der Befund, dass sich, wenn überhaupt, Effekte einer Reduktion der Klassengröße eher in Situationen zeigen, wenn die Änderungen der Klassengröße sich im mittleren Bereich der Klassenstärke, d. h. vor und nach der Änderung jeweils unter 20 Schülerinnen und Schüler pro Klasse, abspielen. Eine Reduktion der Klassengröße als Steuerungsmaßnahme würde vermutlich immer bei den größten Klassen ansetzen, durch die Herabsetzung der Klassenschülerzahl auf einen Wert unter 25. Die Schulen, die davon profitieren würden, hätten nach den vorliegenden Ergebnissen jedoch am wenigsten Vorteil davon. Auch dies deutet dahin, dass eine Reduktion der Klassengröße eine Investition ohne Effekt bliebe.

Die Ergebnisse dieser Studie ordnen sich in eine Reihe von Forschungsarbeiten ein, die nur geringe positive Effekte von kleineren Klassen finden (siehe Forschungslage 2.1) und widersprechen den wenigen Arbeiten, die einen Nutzen aufzeigen. Häufig stellt sich dann die Frage, warum die zusätzlichen Ressourcen verpuffen. Altrichter et al. (2009) verweisen darauf, dass vermutlich Lehrpersonen über entsprechende Qualifikationen und Konzepte verfügen müssen, um in der Lage zu sein, verbesserte Arbeitsbedingungen für die Weiterentwicklung von Unterricht zu nutzen und dass Reduktionen der Klassenschülerzahl nicht automatisch dazu führen, dass Lehrerinnen und Lehrer ihren Unterrichtsstil umstellen, um Potenziale der verbesserten Arbeitsbedingungen zu nutzen. Auch die hier präsentierten Ergebnisse bedeuten nicht, dass kleine Klassen in Österreich bei angepasster Pädagogik einen Nutzen haben könnten, doch folgt aus ihnen, dass die Frage nach der entsprechenden Pädagogik zuerst sehr gut beantwortet sein sollte und Implikationen für die Umsetzung geklärt werden müssten, bevor entsprechende Maßnahmen gesetzt werden. Bedingt durch die nicht vorhandenen Effekte im Volksschulsystem in Österreich scheint es jedoch gefährlich, weiter auf Reduktionen der Klassengröße zu drängen und darauf zu hoffen, flankierende Lehrerfortbildung könnte dann dafür sorgen, dass die Potenziale der Maßnahme genutzt werden. Vielmehr sollten vor jeder Maßnahme erst die Gelingensbedingungen gesichert werden, und dies ist in Österreich für kleinere Klassen offensichtlich nicht gegeben.

Abschließend muss betont werden, dass diese Studie sich nur mit der Klassengröße befasst und damit keine Aussagen über den Effekt des Einsatzes zusätzlicher Personalressourcen in anderen Zusammenhängen erlaubt. Aber sie betont damit die Notwendigkeit, zusätzliche Personalressourcen mit klaren pädagogischen Konzepten einzusetzen, anstatt dem populären Ruf nach kleineren Klassen zu folgen.

Literatur

- Altrichter, H., Trautmann, M., Wischer, B., Sommerauer, S. & Doppler, B. (2009). Unterrichten in heterogenen Gruppen: Das Qualitätspotenzial von Individualisierung, Differenzierung und Klassenschülerzahl. In W. Specht (Hrsg.), *Nationaler Bildungsbericht Österreich 2009, Band 2. Fokussierte Analysen bildungspolitischer Schwerpunktthemen* (S. 341–360). Graz: Leykam.
- Arnhold, G. (2004). Die Bedeutung der Klassengröße. Diskussionslinien für Schule und Unterricht. *Essener Unikate (Heft 24)*, 21–27. Verfügbar unter <https://duepublico2.uni-due.de/go/unikate>
- Arztmann, L. (2023, 29. August). Nur ein Drittel der Lehrenden in Österreich ist im Job zufrieden. *Der Standard*. <https://www.derstandard.at/story/3000000184128/nur-ein-drittel-der-lehrenden-in-oesterreich-ist-im-job-zufrieden>
- Bénabou, R., Kramarz, F. & Prost, C. (2009). The French zones d'éducation prioritaire. Much ado about nothing? *Economics of Education Review*, 28 (3), 345–356. doi:10.1016/j.econedurev.2008.04.005
- BIFIE, Robitzsch, A. & Oberwimmer, K. (2022). *BIFIEsurvey: Tools for survey statistics in educational assessment. R package version 3.4-15*. <https://CRAN.R-project.org/package=BIFIEsurvey>
- Bischof, L. M., Hochweber, J., Hartig, J. & Klieme, E. (2013). Schulentwicklung im Verlauf eines Jahrzehnts. Erste Ergebnisse des PISA-Schulpanels. In N. Jude & E. Klieme (Hrsg.), *PISA 2009 – Impulse für die Schul- und Unterrichtsforschung* (Zeitschrift für Pädagogik, 59. Beiheft, S. 172–199). Weinheim: Beltz Pädagogik.
- Bouguen, A., Grenet, J. & Gurgand, M. (2017). *Does class size influence student achievement?* (IPP Policy Brief Nr. 28). Paris: Institut des Politiques Publiques. Verfügbar unter <https://www.ipp.eu/wp-content/uploads/2018/07/n28-noteIPP-Sept2017.pdf>
- Bremische Bürgerschaft (2017). *Ermittlung von Sozialindikatoren und Sozialstufen für die allgemeinbildenden Schulen im Land Bremen*. Drucksache 19/1446. Verfügbar unter https://www.bremische-buergerschaft.de/drs_abo/2017-12-13_Drs-19-1446_991be.pdf
- Broscheid, A. & Gschwend, T. (2003). *Augäpfel, Murmeltiere und Bayes: Zur Auswertung stochastischer Daten aus Vollerhebungen*. (MPIfG Working Paper 3/7). Köln: Max-Planck-Institut für Gesellschaftsforschung. Verfügbar unter: <https://www.econstor.eu/bitstream/10419/44278/1/644405333.pdf>
- Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (BIFIE). (2017). *Almanach zur Standardüberprüfung 2013 Mathematik, 4. Schulstufe, Schülerebene, M413I, nicht imputierter Datensatz*. Verfügbar unter <https://iqs.gv.at/fdb>.
- Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (BIFIE). (2019a). *Standardüberprüfung 2018. Mathematik, 4. Schulstufe. Bundesergebnisbericht*. Salzburg. Verfügbar unter https://www.bifie.at/wp-content/uploads/2019/02/BiSt_UE_M4_2018_Bundesergebnisbericht.pdf
- Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (BIFIE). (2019b). *Almanach zur Standardüberprüfung 2018 Mathematik,*

4. Schulstufe, Schülerebene, M418I, nicht imputierter Datensatz. Verfügbar unter <https://iqs.gv.at/fdb>.
- Bürgerschaft der freien und Hansestadt Hamburg (2012). *Schriftliche Kleine Anfrage Robert Heinemann (CDU) vom 16.04.12 und Antwort des Senats. KESS-Faktoren im Wandel*.
- Ditton, H. (1998). *Mehrebenenanalyse. Grundlagen und Anwendungen des Hierarchisch Linearen Modells*. Weinheim: Juventa.
- Dumont, H., Neumann, M., Maaz, K. & Trautwein, U. (2013). Die Zusammensetzung der Schülerschaft als Einflussfaktor für Schulleistungen. *Psychologie in Erziehung und Unterricht*, 60 (3), 163–183.
- European Commission. (2022). *Interim report of the Commission expert group on quality investment in education and training*. Brüssel.
- Filges, T., Sonne-Schmidt, C. S. & Nielsen, B. C. V. (2018). Small class sizes for improving student achievement in primary and secondary schools. A systematic review. *Campbell systematic reviews*, 14 (1), 1–107. doi:10.4073/csr.2018.10
- Freunberger, R., Mayrhofer, L. & Sauerwein, J. (2019). Die Situation von Klein- und Kleinstschulen in der Volksschule. In A. C. George, C. Schreiner, C. Wiesner, M. Pointinger & K. Pacher (Hrsg.), *Fünf Jahre flächendeckende Bildungsstandardüberprüfungen in Österreich. Vertiefende Analysen zum Zyklus 2012 bis 2016* (S. 89–102). Münster, New York: Waxmann.
- Ganzeboom, H. (2010). *Questions and answers about ISEI-08*. Verfügbar unter <http://www.harryganzeboom.nl/isco08/qa-isei-08.htm>
- George, A. C. & Klinglmair, R. (2023). Schule 2040 – Ein Bericht mit Perspektive(n). *Erziehung und Unterricht* (7–8), 634–647.
- George, A. C., Oberwimmer, K. & Itzlinger-Bruneforth, U. (2016). Stichprobenziehung. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandard-Überprüfung* (S. 51–82). Wien: facultas.
- George, A. C. & Robitzsch, A. (2014). *Fairer Vergleich. Technische Dokumentation – BIST-Ü Mathematik, 4. Schulstufe, 2013*, Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (BIFIE).
- Gershenson, S. & Langbein, L. (2015). The Effect of Primary School Size on Academic Achievement. *Educational Evaluation and Policy Analysis*, 37 (1_suppl), 135S–155S. doi:10.3102/0162373715576075
- Götz, F. M., Gosling, S. D. & Rentfrow, P. J. (2024). Effect sizes and what to make of them. *Nature Human Behaviour*, 8 (5), 798–800. doi:10.1038/s41562-024-01858-z
- Hallquist, M. N. & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Structural Equation Modeling*, 25 (4), 621–638. doi:10.1080/10705511.2017.1402334.
- Hanushek, E. A. (2002). Evidence, politics, and the class size debate. In L. Mishel & R. Rothstein (Hrsg.), *The Class Size Debate* (S. 37–65). Washington: Economic Policy Institute.
- Hattie, J. (2005). The paradox of reducing class size and improving learning outcomes. *International Journal of Educational Research*, 43 (6), 387–425. doi:10.1016/j.ijer.2006.07.002
- Hattie, J. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. doi:10.4324/9780203887332
- Helm, C., George, A., Zintl, R. & Weber, Ch. (2025). Mechanismen der Klassenzusammensetzung und ihre Effekte auf die Schulleistung. In A. C. George, R. Zintl & C. Helm

- (Hrsg.), *10 Jahre Bildungsstandarderhebungen – Bedingungen (langfristiger) Kompetenzentwicklungen auf Schul- und Systemebene* (Kompetenzmessungen im österreichischen Schulsystem: Analysen, Methoden & Perspektiven, Band 4, S. 237–272). Münster: Waxmann. doi:10.31244/9783818850593
- Im Brahm, G. (2006). Klassengröße: eine wichtige Variable von Schule und Unterricht? *Bildungsforschung*, 3 (1). Verfügbar unter <https://bildungsforschung.org/ojs/index.php/bildungsforschung/article/view/28/26>
- Jepsen, C. (2015). Class size: does it matter for student achievement? *IZA World of Labor*. doi:10.15185/izawol.190
- Jepsen, C. & Rivkin, S. G. (2002). What is the tradeoff between smaller classes and teacher quality? *NBER WORKING PAPER SERIES: 9205*. Verfügbar unter https://www.nber.org/system/files/working_papers/w9205/w9205.pdf
- Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, 49 (4), 241–253. doi:10.3102/0013189X20912798
- Krüger, A. B. (2002). Understanding the magnitude and effect of class size on student achievement. In L. Mishel & R. Rothstein (Hrsg.), *The Class Size Debate* (S. 7–35). Washington: Economic Policy Institute.
- Levin, B. (2008). *Large-Scale Change in Education. Paper presented to the Canadian Society for the Study of Education*. Verfügbar unter <http://lib-ocs.lib.sfu.ca:8087/fedcan/index.php/csse/csse2008/paper/download/402/273>
- Li, W. & Konstantopoulos, S. (2017). Does class-size reduction close the achievement gap? Evidence from TIMSS 2011. *School Effectiveness and School Improvement*, 28 (2), 292–313. doi:10.1080/09243453.2017.1280062
- Lorah, J. (2018). Effect size measures for multilevel models. Definition, interpretation, and TIMSS example. *Large-scale Assessments in Education*, 6 (1), 1. doi:10.1186/s40536-018-0061-2
- Maag Merki, K. (2014). Das quasi-experimentelle Design in der Educational Governance-Forschung? Herausforderungen, Möglichkeiten und Grenzen am Beispiel der Analyse der Wirksamkeit der Einführung zentraler Abiturprüfungen. In K. Maag Merki, R. Langer & H. Altrichter (Hrsg.), *Educational Governance als Forschungsperspektive. Strategien, Methoden, Ansätze* (S. 51–83). Wiesbaden: Springer VS.
- Mazrekaj, D., Witte, K. D. & Triebs, T. P. (2022). Mind the Gap. Measuring Academic Underachievement Using Stochastic Frontier Analysis. *Exceptional Children*, 88 (4), 442–459. doi:10.1177/00144029211073524
- Mishel, L. & Rothstein, R. (Hrsg.). (2002). *The Class Size Debate*. Washington: Economic Policy Institute. Verfügbar unter <http://hanushek.stanford.edu/sites/default/files/publications/Hanushek%202002%20ClassSizeDebate.pdf>
- Muthén, L. K. & Muthén, B. O. (2015). *Mplus User's Guide* (7. Auflage). Los Angeles, CA: Muthén & Muthén.
- Nandrup, A. B. (2014). Do class size effects differ across grades? *Education Economics*, 24 (1), 83–95. doi:10.1080/09645292.2015.1099616
- Neubacher, M., Freunberger, R., Schreiner, C., Vogtenhuber, S., Oberwimmer, K., Höller, I., Mayrhofer, L., Baumegger, D., Steiger, A., Gurtner-Reinthal, S., Toferer, B. & Wallner-Paschon, C. (2019). Indikatoren D: Output – Ergebnisse des Schulsystems. In K. Oberwimmer, S. Vogtenhuber, L. Lassnigg & C. Schreiner (Hrsg.), *Nationaler Bildungs-*

- bericht Österreich 2018, Band 1: Das Schulsystem im Spiegel von Daten und Indikatoren.* doi:10.17888/nbb2018-1.2
- Nusche, D., Radinger, T., Theisens, H. & Busemeyer, M. R. (2016). *OECD Reviews of School Resources: Austria 2016*. Paris: OECD Publishing. doi:10.1787/9789264256729-en
- Nye, B., Hedges, L. V. & Konstantopoulos, S. (2001). The Long-Term Effects of Small Classes in Early Grades. Lasting Benefits in Mathematics Achievement at Grade 9. *The Journal of Experimental Education*, 69 (3), 245–257. doi:10.1080/00220970109599487
- OECD (2012). *Equity and quality in education. Supporting disadvantaged students and schools*. Paris: OECD Publishing. doi:10.1787/9789264130852-en
- Paulus, C. (2009). *Der Einfluss der Klassengröße auf Schülerleistungen an Grundschulen*. doi:10.23668/psycharchives.10897
- R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. Wien, Österreich: R Foundation for Statistical Computing. <https://www.R-project.org>
- Radisch, F. & Vatterrott, A. (2024). Klassengröße als Steuerungselement gegen den Lehrkräftemangel. Eine Ad-hoc-Analyse am Beispiel Mecklenburg-Vorpommerns. *DDS – Die Deutsche Schule*, 2024 (3), 310–323. doi:10.31244/dds.2024.03.09
- Rechnungshof Österreich (2011). *Klassenschülerhöchstzahl 25* (Reihe BUND 2011/1). Wien: Rechnungshof. Verfügbar unter https://www.parlament.gv.at/dokument/XXIV/III/205/imfname_203990.pdf
- Robitzsch, A., Pham, G. & Yanagida, T. (2016). Fehlende Daten und Plausible Values. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandard-Überprüfung* (S. 293–259). facultas.
- Salchegger, S. & Bruneforth, M. (2024). *Zur Beurteilung der praktischen Bedeutsamkeit von Gruppenunterschieden in Erstberichten zu internationalen Studien* (IQS-Report 2/2024). Salzburg: IQS. doi:10.17888/iqsreport-2024-2
- Schmid, C., Trendtel, M., Bruneforth, M. & Hartig, J. (2019). Effectiveness of a governmental action to improve Austrian primary schools – results of multilevel analyses based on repeated cycles of educational standards assessments. *School Effectiveness and School Improvement*, 31 (2), 149–171. doi:10.1080/09243453.2019.1620294
- Schreiner, C. & Breit, S. (Hrsg.). (2014). *Standardüberprüfung 2013 Mathematik, 4. Schulstufe – Bundesergebnisbericht*. Salzburg. Verfügbar unter <https://www.bifie.at/node/2489>
- Schreiner, C. & Wiesner, C. (2019). Die Überprüfung der Bildungsstandards in Österreich: der erste Zyklus als Meilenstein für die Schul- und Unterrichtsentwicklung – eine gelungene Innovation im österreichischen Schulsystem. In A. C. George, C. Schreiner, C. Wiesner, M. Pointinger & K. Pacher (Hrsg.), *Fünf Jahre flächendeckende Bildungsstandardüberprüfungen in Österreich. Vertiefende Analysen zum Zyklus 2012 bis 2016* (Kompetenzmessungen im österreichischen Schulsystem: Analysen, Methoden & Perspektiven, S. 13–53). Münster, New York: Waxmann. Verfügbar unter <https://www.waxmann.com/index.php?eID=download&buchnr=3925>
- Schwippert, K. (2002). *Optimalklassen: Mehrebenenanalytische Untersuchungen*. Münster, New York: Waxmann.
- Shen, T. & Konstantopoulos, S. (2022). Are class size and teacher characteristics associated with cognitive outcomes in early grades? *School Effectiveness and School Improvement*, 33 (3), 333–359. doi:10.1080/09243453.2021.2023585

- Statistik Austria (2014). *Standard-Dokumentation Metainformationen (Definitionen, Erläuterungen, Methoden, Qualität) zur Schulstatistik*. Bearbeitungsstand: 07.03.2014, Statistik Austria. Zugriff am 10.12.2014. Verfügbar unter https://www.statistik.at/fileadmin/shared/QM/Standarddokumentationen/B_2/std_b_schulstatistik.pdf
- Statistik Austria. (2023). *Bildung in Zahlen 2021/22 – Tabellenband*. Wien: Statistik Austria. Verfügbar unter https://www.statistik.at/fileadmin/user_upload/BiZ-2021-22_Tabellenband.pdf
- Saldern, M. von (1993). *Klassengröße als Forschungsgegenstand*. Landau: Verl. d. Univ. Landau. Verfügbar unter <https://pub-data.leuphana.de/frontdoor/deliver/index/docId/145/file/VVN358CF.pdf>
- West, M. R. & Wößmann, L. (2002). *Class-Size Effects in School Systems Around the World: Evidence from Between-Grade Variation in TIMSS* (Kiel Working Paper No. 1099). Kiel: Kiel Institute of World Economics (IfW).

Anhang 1: Auswahl der Analysedaten

Dieser Anhang beschreibt, wie in mehreren Schritten die Analysedaten der Studie gebildet werden. Zuerst wird die Zielgruppe der Studie, d. h. die Gruppe der Schulen, für die die Studie Aussagen treffen soll, definiert, dann wird die Auswahl reduziert auf Schulen, für die sowohl Leistungsdaten in beiden Erhebungen vorliegen, als auch Konsistenz zwischen Erhebungsjahren besteht, abschließend werden die Analysedaten gebildet, wozu Schulen ausgeschlossen werden, bei denen Einflüsse auf die Änderungen vermutet werden können, die statistisch in diesem Design nicht kontrolliert werden können.

Bestimmung der Schulen der Zielpopulation

Für die Studie wird in folgenden Auswahlritten nach und nach die Menge der Schulen von der Vollpopulation der Zielsetzung der Studie gemäß reduziert. Zuerst finden die BilDok-Daten Anwendung, für die Schritte a bis d noch getrennt für die Jahre 2013 und 2018.

- a. Vollpopulation: Schülerinnen und Schüler in Schulen und Klassen mit mindestens einem Kind auf der 4. Schulstufe, das in der Schulform „Volksschule“ gemeldet ist und das weder außerordentlichen Status (ao. Status, Kinder, die aufgrund mangelnder Deutschkenntnisse dem Unterricht nicht folgen können) noch sonderpädagogischen Förderbedarf (SPF) hat. Hier sind bereits Statutschulen ausgeschlossen und Sonderschulen (ASO), die keine regulären Volksschulkinder führen (d. h. ohne die sogenannte „umgekehrte Inklusion/Integration“). Diese Auswahl entspricht etwa der Grundgesamtheit der BIST-Ü mit laut BIFIE (2019a) 76.447 zu testenden Schülerinnen und Schülern in 2018.
- b. Reduktion auf Schulen mit mehr als 5 Volksschülerinnen und -schülern (laut Schulform), die weder ao. Status noch SPF haben. In diesem Schritt entfallen sowohl sehr

- kleine Schulen als auch Schulen, die (nahezu) nur Schülerinnen und Schüler mit SPF unterrichten (weitere ASO).
- c. Ausschluss der Schulen mit Schülerinnen und Schülern der 4. Schulstufe in Mehrstufenklassen. Schulen mit Mehrstufenklassen werden ausgenommen, weil angenommen werden muss, dass der Zusammenhang zwischen Unterrichtsqualität und Klassengröße ein anderer ist als für Einstufen-Klassen. Insbesondere, wenn für die beiden Erhebungsjahre nicht die gleiche Organisationsform vorliegt, kann diese Änderung nicht kontrolliert werden.
 - d. Zielgruppe: Ausschluss von Schulen, die gleichzeitig mehr als eine Klasse führen und stark ungleiche aktuelle Klassengröße zwischen den Klassen derselben Schulstufe zeigen. Es ist anzunehmen, dass innerhalb der Schule ein Selection Bias bei der Zuordnung der Schülerinnen und Schüler zu den Klassen vorliegt, da die Schulleitung offensichtlich Gründe hat, die Klassen unterschiedlich zu führen. Ausgeschlossen werden Schulen, bei denen die kleinste Klasse weniger als $\frac{2}{3}$ der Schülerinnen und Schüler der zweitkleinsten Klasse hat.

Konsistenz zwischen Erhebungsjahren und mit der BIST-Ü

Ab hier werden beide Erhebungsjahre zusammen betrachtet und der Datenbestand zu den BIST-Ü berücksichtigt. Die nächsten Schritte e bis h könnten als Ausfall im Schulpanel betrachtet werden.

- e. Reduktion auf Schulen, für die die Jahre 2013 und 2018 in der BilDok verknüpft werden können. Es entfallen Schulen, für die nur in einem der beiden Jahre Daten vorliegen, z. B. aufgrund von Schulschließungen, Neugründungen oder Zusammenlegungen. Im selben Schritt werden ferner nur Schulen beibehalten, für die für beide Jahre (2013, 2018) Daten in der BIST-Ü vorliegen. Änderungen der Schulkennzahl zwischen 2013 und 2018 wurden berücksichtigt und führen nicht zum Ausschluss.
- f. Zuletzt entfallen Schulen, für die zwar die Ausschlussgründe in den Schritten b bis d in einem Jahr nicht zutreffen, aber in dem anderen Jahr. Es verbleiben nur Schulen, die in beiden Jahren keines der Ausschlusskriterien erfüllen.

Ausschluss aufgrund bedeutsamer Änderungen in der Schulstruktur bzw. den Rahmenbedingungen

Da das Ziel der folgenden Analysen ist, jeweils im Längsschnitt die Veränderungen der mittleren Kompetenzen der Schülerinnen und Schüler in Zusammenhang mit sich ändernden Klassengrößen ihrer Schulen zu untersuchen, sollen andere Faktoren, die mit der Leistung zusammenhängen, möglichst kontrolliert werden. Dazu werden im linearen Mehrebenenmodell wichtige demografische Faktoren kontrolliert. Allerdings finden sich Schulen mit bedeutsamen Änderungen in der Schulstruktur bzw. den Rahmenbedingungen,

die sich einer statistischen Kontrolle im Modell entziehen. Daher werden in den nächsten Schritten Schulen mit wichtigen Änderungen zwischen den Erhebungen identifiziert und auch ausgeschlossen:

- g. Schulen, bei denen sich zwischen den Erhebungsjahren die Rahmenbedingungen aufgrund von Einflüssen durch Migration deutlich unterscheiden. Ausgeschlossen werden Schulen, die auf der 4. Schulstufe, in einem Jahr keine ao. Schülerinnen und Schüler hatten, aber im anderen Jahr mehr als 4, und Schulen, bei denen sich die Anzahl der ao. Schülerinnen und Schüler zwischen den Erhebungsjahren um mehr als 5 veränderte.
- h. Schulen, bei denen sich zwischen den Erhebungsjahren die Rahmenbedingungen aufgrund von Änderungen in der Inklusion deutlich unterscheiden. Eine Erweiterung oder Reduktion inklusiven Unterrichts geht häufig mit der Änderung von Klassenstrukturen einher bzw. beeinflusst sie bei unveränderten Klassenstrukturen die Unterrichtssituation (wobei hier nebensächlich ist, in welcher Art). Diese systematischen Effekte würden den Einfluss der sich ändernden Klassengröße überlagern, können aber nicht kontrolliert werden. Ausgeschlossen werden daher Schulen, die in einem Jahr keine Schülerinnen und Schüler mit SPF hatten, aber im anderen Jahr mehr als 5, und Schulen, bei denen sich die Anzahl der Schülerinnen und Schüler mit SPF zwischen den Erhebungsjahren um mehr als 7 veränderte.

Tabelle A1.1: Anzahl der Schulen, Schülerinnen und Schüler sowie Klassen in der Vollpopulation, Zielgruppe und Analysedaten und weitere Auswahlsschritte (2013, 2018)

Auswahl- schritt (Erklärung siehe im Text)	2013			2018		
	N-Schulen	N-Klassen	N-Schülerinnen und -Schüler	N-Schulen	N-Klassen	N-Schülerinnen und -Schüler
a (Vollpopulation)	3.052	5.013	75.823	2.976	5.020	76.369
B	2.757	4.717	74.796	2.699	4.742	75.363
C	2.030	3.530	61.930	2.003	3.559	62.427
d (Zielgruppe)	2.013	3.487	61.273	1.988	3.524	61.976
E	1.975	3.426	60.235	1.952	3.468	61.005
F	1.770	3.143	55.897	1.770	3.179	56.358
G	1.743	3.065	54.317	1.743	3.091	54.630
h (Analysedaten)	1.716	2.997	52.998	1.716	3.019	53.284

Anmerkung: N-Schülerinnen und -Schüler bezieht sich auf alle regulären Schülerinnen und Schüler, die für die 4. Schulstufe gemeldet sind, d. h. exkl. solcher mit ao. Status bzw. SPF, sowie Schülerinnen und Schüler in Mehrstufenklassen auf anderen Schulstufen.

Aus Sicht der Datenbasis der BIST-Ü stellt sich das Mengengerüst wie folgt dar.

Tabelle A1.2: Mengengerüst zu den Kompetenzdaten: Anzahl der Schulen, getestete Schülerinnen und Schüler sowie Klassen in der Vollpopulation, Zielgruppe und Analysedaten (2013, 2018)

Auswahl- schritt (Erklärung siehe Text)	2013			2018		
	N-Schulen	N-Klassen	N-Schülerinnen und -Schüler	N-Schulen	N-Klassen	N-Schülerinnen und -Schüler
Vollpopulation	3.048	4.910	73.655	2.961	4.804	73.780
Zielgruppe	1.975	3.426	58.528	2.174	3.489	59.473
Analysedaten	1.716	2.997	51.501	1.716	3.017	51.582

Anmerkung: N-Schülerinnen und -Schüler bezieht sich auf die in der BIST-Ü getesteten Schülerinnen und Schüler. N-Klassen auf die in der BIST-Ü getesteten Klassen. Die Schulauswahl ist auf Basis der BilDok-Daten erfolgt.

Anhang 2: Alternative Definitionen der mittleren Klassengröße

Für diese Studie wurden die folgenden fünf Definitionen und Operationalisierungen von Klassengröße zur Verwendung in Betracht gezogen.

Längsschnittliche Klassengröße der BilDok: Diese Definition wird in der Studie verwendet. Die Definition findet sich daher im Abschnitt 4.3.

Aktuelle Klassengröße der BIST-Ü-Erhebungen: Die BIST-Ü erfasst für jede Schülerin und jeden Schüler eine Kennung der Klasse. Aus den Angaben der Schulleitungen ergibt sich die Anzahl der gemeldeten Schülerinnen und Schüler jeder Klasse, inklusive der nicht zur Teilnahme verpflichteten und abwesenden Schülerinnen und Schüler. Da der vorliegende Datensatz nur die Schülerinnen und Schüler umfasst, die zur Teilnahme an den BIST-Ü verpflichtet waren und auch teilgenommen haben, ist eine simple Auszählung der Anzahl der Schülerinnen und Schüler mit Daten keine Quelle für die Klassengröße.

Aktuelle Klassengröße der BilDok: Die Erhebung zur BilDok durch die Statistik Austria erfolgt direkt aus den Schulverwaltungsprogrammen der Schulen und erlaubt über eine schulspezifische Klassenkennung die Bestimmung der Klassengrößen. Diese schließt alle Schülerinnen und Schüler ein mit der Möglichkeit, die Zahl der Schülerinnen und Schüler mit Förderbedarf bzw. außerordentlichem Status zu bestimmen. Allerdings können die Klassenkennungen der BilDok nicht mit den Klassenkennungen der BIST-Ü verknüpft werden. Daher kann dieser Indikator nicht für Analysen verwendet werden.

Mittlere Klassengröße der BIST-Ü-Erhebungen: Da anzunehmen ist, dass bei Schulen mit mehr als einer Klasse Unterschiede in der Klassengröße das Ergebnis bewusster Steuerung sind, sind in diesen Fällen die Klassengrößen vermutlich nicht unabhängig von anderen leistungsrelevanten Faktoren. Bouguen et al. (2017) vermuten, dass Schulleitungen häufig die

leistungsstärkeren Schülerinnen und Schüler den größeren Klassen zuteilten und damit ein systematischer Fehler entstünde, wenn Klassengrößen mit Leistungen verglichen würden. Um diesen Zuteilungsfehler innerhalb der Schulen auszublenden, wird für jede Schule der Mittelwert der Klassengröße an der Schule als Indikator verwendet. Mit der mittleren Größe wird Klassengröße als Schulebenenfaktor behandelt.

Mittlere Klassengröße der BilDok: Wie für die BIST-Ü ausgeführt, wird aus der aktuellen Klassen-/Gruppengröße der BilDok für jede Schule der Mittelwert der Klassengröße berechnet. Die mittlere Klassengröße der BilDok lässt sich über die Schulkennung mit den Daten der BIST-Ü verknüpfen. Da die Klassengröße der BilDok standardisiert jährlich seit 2006 über die Schulverwaltung erhoben wird, ist denkbar, dass diese Zahlen im Längsschnitt vergleichbarer und verlässlicher sind.

Die folgende Tabelle zeigt die Verteilung der Klassengrößen für Klassen mit Schülerinnen und Schülern der 4. Schulstufe. In den Analysedaten, für die die Analysen im Ergebnisteil gelten, besuchte 2013 eine durchschnittliche Schülerin bzw. ein durchschnittlicher Schüler in Österreich eine Klasse mit ca. 19 Schülerinnen und Schülern, wobei die Standardabweichung 3,4 betrug. Bis 2018 ist die mittlere Klassengröße leicht angestiegen und die Streuung leicht erhöht. Die mittlere Größe der Klassen ohne Schülergewichtung ist entsprechend kleiner, mit 17,9 in 2018 und einer SD von 3,8. Die kleinsten 10 % der Klassen haben 13 oder weniger Schülerinnen und Schüler, während die größten 10 % der Klassen 23 oder mehr Schülerinnen und Schüler haben.

Tabelle A2.1: Verteilung der mittleren Klassengröße (BilDok, jährlich) aus Sicht der Schülerinnen und Schüler (gewichtet nach Schülerinnen und Schülern) und aus Sicht der Klassen in der Vollpopulation, Zielgruppe, Analysedaten (2013, 2018)

Auswahl bzw. Teilpopulation	Jahr	MW	SD	Perzentile				
				10.	25.	50.	75.	90.
				Gewichtet nach Schülerinnen und Schülern (4. Schulstufe)				
Vollpopulation	2013	19,0	3,6	14,0	16,5	19,5	22,0	23,4
	2018	19,4	3,7	14,0	17,0	20,0	22,5	24,0
Zielgruppe	2013	19,0	3,5	14,0	16,5	19,3	22,0	23,3
	2018	19,4	3,6	14,0	17,0	19,7	22,5	24,0
Analysedaten	2013	19,0	3,4	14,0	16,5	19,3	21,7	23,3
	2018	19,3	3,6	14,0	16,7	19,5	22,3	23,8
		Nach Klassen						
Vollpopulation	2013	17,5	4,1	12,0	14,5	17,7	20,7	23,0
	2018	17,6	4,1	12,0	14,5	18,0	21,0	23,0
Zielgruppe	2013	17,6	3,8	13,0	14,5	18,0	20,7	23,0
	2018	17,8	3,9	13,0	15,0	18,0	21,0	23,0
Analysedaten	2013	17,8	3,7	13,0	15,0	18,0	21,0	23,0
	2018	17,9	3,8	13,0	15,0	18,0	21,0	23,0

Anmerkung: MW = Mittelwert, SD = Standardabweichung. Für Mehrstufenklassen in der Vollpopulation zählen alle Schülerinnen und Schüler zur Klassengröße, aber nur Schülerinnen und Schüler der 4. Schulstufe tragen zur Gewichtung bei.

Tabelle A2.2: Längsschnittliche Klassengröße (Bildok) aus Sicht der Schülerinnen und Schüler (gewichtet nach Schülerinnen und Schülern), Zielgruppe, Analysedaten (2013, 2018)

Auswahl bzw. Teilpopulation	Jahr	MW	SD	Perzentile				
				10.	25.	50.	75.	90.
				Gewichtet nach Schülerinnen und Schülern (4. Schulstufe)				
Vollpopulation	2013	17,6	4,1	12,0	14,9	18,0	20,5	22,5
	2018	17,7	4,1	12,3	15,0	18,0	21,0	22,8
Zielgruppe	2013	17,9	3,9	13,0	15,5	18,3	20,8	22,8
	2018	18,0	3,9	13,0	15,4	18,3	21,0	23,0
Analysedaten	2013	18,0	3,8	13,0	15,5	18,5	20,8	22,5
	2018	18,0	3,9	13,0	15,3	18,3	21,0	22,8

Anmerkung: MW = Mittelwert, SD = Standardabweichung. Für Mehrstufenklassen in der Vollpopulation zählen alle Schülerinnen und Schüler zur Klassengröße, aber nur Schülerinnen und Schüler der 4. Schulstufe tragen zur Gewichtung bei.

Die verschiedenen Teilpopulationen und Definitionen von Klassengröße unterscheiden sich nicht substanziell (Tabelle 8). Auf Schulebene betrachtet, korrelieren mittlere Klassengröße (4. Schulstufe) und längsschnittliche Klassengröße mit .92. Auch Änderungen zwischen 2013 und 2018 korrelieren für die Anwendung unterschiedlicher Definitionen hoch, mit .88. Aus genannten inhaltlichen Gründen wird trotzdem für die Analysen die komplexere längsschnittliche Klassengröße verwendet, da so der gesamten Bildungslaufbahn der Kinder Rechnung getragen wird.

Anhang 3: Mplus-Beispielsyntax: Model L1

```

TITLE: Model L1 (based on Schmid et al.) ;

DATA:
    FILE IS bist_imp_list.dat; !List of 10 Imputations
    TYPE IS IMPUTATION;

! Die Daten liegen in 10 Dateien im Langformat vor.

VARIABLE:  NAMES ARE
    school tsidjahr tsidcl tsidstud wgtstud mathe jahr gender buecher hisei
    mig sprache bildung land stadt klasseGr DklasseG ses sesSch sesSchJ sesG
    sesXDklG sesGfix Seslow Seshigh;

USEVARIABLES ARE
    jahr gender buecher sprache mig bildung hisei
    mathe land stadt DKLASSEG ;

MISSING IS ALL (-9999);

WITHIN = gender buecher hisei mig sprache bildung jahr;
BETWEEN = land stadt DKLASSEG;
CLUSTER = school;

DEFINE:
    CENTER gender buecher hisei mig sprache bildung (GRANDMEAN); !sesXklg

ANALYSIS:
    TYPE = TWOLEVEL RANDOM; ESTIMATOR = ML;
    MODEL:
        %WITHIN%
            mathe ON gender buecher hisei mig sprache bildung; ! regression
            student covariates
            change | mathe ON jahr; ! random slope
        %BETWEEN%
            mathe ON land stadt ;
            change ON DklasseG ;

```


Warum schneiden Mädchen bei der Mathematik-Zentralmatura schlechter ab als Burschen? Ein Erklärungsversuch zu Geschlechterungleichheiten im österreichischen Bildungssystem

1 Einleitung

Der Rechnungshofbericht zur Zentralmatura stellte fest, „dass bei der schriftlichen Mathematik-Matura die Kandidatinnen schlechter abschnitten als die Kandidaten“ (Rechnungshof Österreich, 2020, S. 9). Die Schlussempfehlung 22 lautet dementsprechend: „Der Gender-Gap in Mathematik wäre einer umfassenden Untersuchung zu unterziehen, um langfristig mit geeigneten, innerhalb der eigenen Ingerenz gelegenen Maßnahmen (z. B. Unterricht, Gewährleistung genderfairer Testungen) gegensteuern zu können“ (S. 103).

Das vorliegende Kapitel geht der Frage nach, welche vorgelagerten Gründe eine Rolle für das schlechtere Abschneiden der Kandidatinnen im Vergleich zu den Kandidaten bei der standardisierten kompetenzorientierten Reife- und Diplomprüfung (SRDP)² in Mathematik spielen können. Helbig (2012) stellt ein Modell vor, in dem verschiedene Bildungsindikatoren in Form von individuellen Leistungsmerkmalen oder zu bewältigenden Übergängen für das Erlangen der Matura auf dem klassischen schulischen Weg³ relevant sind (Abbildung 1). Dabei spielen Kompetenzen und Noten eine wesentliche Rolle als Prädiktoren für den Übertritt in eine höhere Schule, deren erfolgreicher Abschluss wiederum zum Antreten bei der Matura berechtigt.

1 Die Autorinnen danken Markus Bönisch und Martin Hofer für ihre Anmerkungen zu einer früheren Version dieses Beitrags.

2 Die Ausdrücke „SRDP“ und „Zentralmatura“ werden im vorliegenden Kapitel synonym verwendet. „SRDP“ umfasst damit – analog z. B. zum Bundesministerium für Bildung, Wissenschaft und Forschung (BMBWF, 2023h) – nicht nur die standardisierte Reife- und Diplomprüfung an berufsbildenden höheren Schulen (BHS), sondern auch die standardisierte Reifeprüfung an allgemeinbildenden höheren Schulen (AHS). Die Berufsreifeprüfung (BRP) wird hier explizit ausgeklammert.

3 Auf die Wege zur Berufsreifeprüfung wird hier nicht eingegangen.

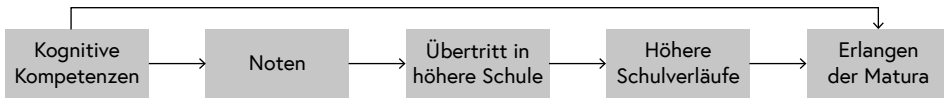


Abbildung 1: Schematische Darstellung der Bildungsindikatoren auf dem Weg zur Matura (übernommen von Helbig, 2012; adaptiert an das österreichische Bildungssystem durch die Autorinnen)

Für diese Prädiktoren untersuchen wir im vorliegenden Kapitel, ob geschlechtsspezifische Unterschiede bestehen, die zum unterschiedlichen Abscheiden von Mädchen und Burschen bei der schriftlichen Mathematiklausur der SRDP beitragen könnten.

Die vorliegende Studie folgt einem Pseudopanel und untersucht die eben angeführten Indikatoren bei Personen, die 2002 bzw. 2003 geboren sind, über unterschiedliche Erhebungsmeilensteine hinweg. Im Jahr 2013 waren Kinder dieser Geburtskohorten in der Regel 10 Jahre alt und nahmen auf der 4. Schulstufe an der Bildungsstandardüberprüfung (BIST-Ü) 2013 in Mathematik (M4) teil, vier Jahre später als 14-Jährige bei der BIST-Ü M8 2017. Im Jahr darauf nahm ein repräsentativer Teil von ihnen am Programme for International Student Assessment (PISA) 2018 teil und schließlich traten diejenigen unter ihnen, die die Oberstufe einer allgemeinbildenden höheren Schule (AHS) absolviert haben, im Jahr 2021 zur Zentralmatura an und jene, die eine berufsbildende höhere Schule (BHS) absolviert haben, im Jahr 2022. Somit ist es mit diesen Geburtskohorten möglich, die Selektionsprozesse von Mädchen und Burschen bis zum Erreichen der Matura nachzuzeichnen.

1.1 Geschlechterunterschiede in Mathematikkompetenzen und -noten

Groß angelegte Metaanalysen, in denen eine Vielzahl von Ländern und Erhebungen berücksichtigt wurde, kamen zu dem Konsens, dass Leistungsunterschiede zwischen Mädchen und Burschen in Mathematik meist sehr gering sind (z. B. Else-Quest, Hyde & Linn, 2010; Lindberg, Hyde, Petersen & Linn, 2010; Reilly, Neumann & Andrews, 2015; Salchegger, 2015). Österreich scheint im internationalen Vergleich oft als Land mit vergleichsweise großen Geschlechterunterschieden in der Mathematikkompetenz zugunsten von Burschen auf (z. B. Brandmair & Schmoller, 2023; Glaeser et al., 2019; Oberwimmer et al., 2016; Suchań, Schmich, Toferer & Höller, 2016). Bemerkenswert ist in diesem Zusammenhang, dass bei den Bildungsstandardüberprüfungen auf der achten Schulstufe Geschlechterunterschiede in Mathematik gering sind (Cohen's $d = 0,07$ im Jahr 2012 [Oberwimmer et al., 2016] und $d = 0,08$ im Jahr 2017 [Tabelle 4]), aber bei PISA (in der Regel neunte oder zehnte Schulstufe) wesentlich größer ausfallen. Bei PISA 2018 betrug der standardisierte Unterschied $d = 0,14$ (Glaeser et al., 2019; Tabelle 4), bei PISA 2015 $d = 0,28$ (Suchań et al., 2016). Der größere Geschlechterunterschied bei PISA 2015 im Vergleich zu anderen PISA-Erhebungen kann großteils dadurch erklärt werden, dass durch eine spätere Testung (im Herbst statt im Frühjahr) insbesondere leistungsschwache Burschen nach Absolvierung von neun Jahren Schulpflicht häufiger das Schulsystem verlassen haben und damit bei dieser Testung unterrepräsentiert waren (Salchegger & Suchań, 2018).

Dieses Ergebnis weist darauf hin, dass eine unterschiedliche Populationsabdeckung von Mädchen und Burschen Einfluss auf die Größe von Geschlechterunterschieden haben kann. Dieses Phänomen wurde bereits von Reilly (2012) beschrieben, der sich auf den amerikanischen Scholastic Assessment Test (SAT) bezieht, bei dem Burschen in Mathematik stets deutlich besser abschneiden als Mädchen. Dieses Ergebnis lässt sich Reilly (2012) zufolge allerdings nicht auf die Population dieses Alters generalisieren, da am SAT nur Jugendliche teilnehmen, die eine weitere Ausbildung am College anstreben und dies sind stets mehr Mädchen als Burschen. Damit ist beim SAT die Gruppe der Burschen stärker selektiert als die der Mädchen und folglich werden Geschlechterunterschiede zugunsten der Burschen überschätzt. Um Kompetenzen zwischen den Geschlechtern fair vergleichen zu können, muss demnach darauf geachtet werden, dass nicht eine Gruppe stärker selektiert ist (Burschen) als die andere (Mädchen).

In Bezug auf die österreichische Zentralmatura wies Neuwirth (2016) darauf hin, dass die medial diskutierten Geschlechterunterschiede, wonach Mädchen bei der Matura unerwartet schlecht abschnitten (Nimmervoll, 2016), unter anderem damit erklärt werden könnten, dass deutlich mehr Mädchen als Burschen zur Matura antreten und die antretenden Burschen stärker auf Basis ihrer Kompetenzen bzw. Noten ausgesiebt sind. In der Folge werden die Schullaufbahnentscheidungen von Mädchen und Burschen genauer beleuchtet, die zu unterschiedlichen Geschlechteranteilen in maturaführenden Schulen und schlussendlich bei der Matura führen könnten.

Aufnahmevoraussetzung für höhere Schulen sind nicht die in standardisierten Tests gemessenen Kompetenzen, sondern Noten, die die Schülerinnen und Schüler von ihren Lehrkräften erhalten. So können nur Schülerinnen und Schüler ohne Aufnahmeprüfung von der Sekundarstufe I in eine höhere Schule der Sekundarstufe II übertreten, wenn sie die Pflichtgegenstände Deutsch, Mathematik, lebende Fremdsprache in einer Neuen Mittelschule⁴ (NMS) nach dem Leistungsniveau „Vertiefte Allgemeinbildung“ oder in einer AHS-Unterstufe positiv abgeschlossen haben (Bundesministerium für Bildung, Wissenschaft und Forschung [BMBWF], 2018). Übersteigt die Zahl der Bewerberinnen und Bewerber mit diesen Notenvoraussetzungen die Zahl der vorhandenen Plätze, erfolgt eine Aufnahme nach zusätzlichen schulautonom festgelegten Reihungskriterien, bei denen die Zeugnisnoten der achten Schulstufe oft eine wesentliche Rolle spielen (z. B. HTL Krems, 2020).

Obwohl schulische Noten und Kompetenzen zusammenhängen, ist dieser Zusammenhang nicht perfekt. So identifizierte Schmid (2018) systematische Abweichungen in Abhängigkeit von Geschlecht, sozialem Status und Migrationshintergrund: Mädchen, Schülerinnen und Schüler aus Elternhäusern mit einem höheren Sozialstatus sowie Schülerinnen und Schüler mit Migrationshintergrund erzielten bei gleichen Testleistungen bessere Noten in Mathematik als Burschen, Schülerinnen und Schüler aus Elternhäusern mit einem

4 Erst mit dem Schuljahr 2020/21 ersetzte die Mittelschule (MS) die Neue Mittelschule (NMS) als Pflichtschule für die 10- bis 14-Jährigen. Demnach war im Jahr 2017, als die 2002/03 Geborenen auf der achten Schulstufe getestet wurden, noch die NMS die aktuelle Schulform. Wir beziehen uns in diesem Kapitel daher stets auf die NMS inkl. der zugehörigen gesetzlichen Regelungen (z. B. zur Notenskala).

niedrigeren Sozialstatus bzw. Schülerinnen und Schüler ohne Migrationshintergrund. Ähnliche Befunde zeigen sich z. B. bei Maaz, Baeriswyl und Trautwein (2013) sowie Moser, Oostlander und Tomasik (2017). Ein Vorteil der Mädchen bei der Benotung wurde auch auf internationaler Ebene gefunden (Voyer & Voyer, 2014). Dieser Vorteil wird u. a. damit begründet, dass nicht nur fachliche Kompetenzen, sondern auch Faktoren wie Anstrengungsbereitschaft, Gewissenhaftigkeit oder geringes Störverhalten in die Notengebung einfließen (Brookhart et al., 2016; Duckworth & Seligman, 2006; Schreiner, Breit & Haider, 2008).

1.2 Geschlechterunterschiede in der Bildungswahl und im Bildungsverlauf

Im Kontext des österreichischen Bildungswesens zeigen Befunde aus dem Nationalen Bildungsbericht 2021 (Wimmer & Oberwimmer, 2021), dass sich beim Übergang von der Volksschule in die Sekundarstufe I 40 % der Mädchen und 36 % der Burschen für die AHS-Unterstufe entscheiden. Fast alle Schülerinnen und Schüler, die die Unterstufe in einer AHS absolviert haben, besuchen am Beginn der Sekundarstufe II eine maturaführende Schule (97 % der Mädchen und 94 % der Burschen). Größere Geschlechterdifferenzen zeigen sich beim Übertritt von der NMS in die Sekundarstufe II: 53 % der Mädchen und 36 % der Burschen wechseln nach dem Abschluss einer NMS in eine maturaführende Schule. Insgesamt treten 67 % der Mädchen und 52 % der Burschen von der Sekundarstufe I in eine maturaführende Schule über (oder bleiben an einem Gymnasium der Langform). Auch Bacher (2005) sowie Bacher, Beham und Lachmayr (2008) berichten, dass Burschen weniger häufig in eine maturaführende Schule übertreten als Mädchen und dieser Unterschied auch nach Kontrolle der Noten, Kompetenzen und des sozialen Hintergrunds bestehen bleibt. Ähnlich zeigten Benischek, Oberwimmer und Höller (2022) anhand der BIST-Ü-M8-2017-Daten, dass Mädchen mit NMS-Karriere, die die Bildungsstandards in Mathematik erreichen (aber nicht übertreffen), zu 82 % planen, in eine höhere Schule überzutreten, bei den Burschen sind es hingegen nur 63 %.

Werden die unterschiedlichen maturaführenden Schulen differenzierter betrachtet, ergeben sich starke geschlechtsspezifische Unterschiede nach Schulform. Beispielsweise treten Mädchen selbst bei gleicher Leistung wesentlich seltener in eine höhere technische Lehranstalt (HTL) über als Burschen (Salchegger, Glaeser, Widauer & Bitesnich, 2017). Geschlechtsspezifische Schulwahlen haben zudem zur Folge, dass die Schulformen der oberen Sekundarstufe in Österreich stark geschlechtersegregiert sind. So werden Bildungsinstitutionen für Elementarpädagogik bzw. Sozialpädagogik fast ausschließlich von Mädchen besucht, technisch gewerbliche Schulen hingegen nur selten (Wimmer & Oberwimmer, 2021). Diese Geschlechtersegregation nach Schulformen folgt der Regel: *Je mehr Mathematik, desto weniger Mädchen* (Salchegger et al., 2017). Ein Effekt dieser Segregation ist unter anderem, dass Mädchen auf der Sekundarstufe II in Österreich deutlich weniger Mathematikunterricht

erhalten als Burschen (Organisation für wirtschaftliche Zusammenarbeit und Entwicklung [OECD] 2015, Tab. 3.9). Bei PISA 2022 betrug der Nachteil der Mädchen 46 Minuten pro Woche und war einer der größten OECD-weit (Lang & Glaeser, 2023).

Der Großteil der wenigen Forschung zu akademischem Wahlverhalten stammt aus dem amerikanischen Raum. Studien zur Sekundarstufe II fokussieren auf das Kurswahlverhalten von Schülerinnen und Schülern (z. B. Palmer, Burke & Aubusson, 2017; Wang, Eccles & Kenny, 2013), insbesondere die Wahl von Mathematikkursen (z. B. Guo, Parker, Marsh & Morin, 2015; Perez-Felkner, Nix & Thomas, 2017). Diese Studien zeigen, dass das Kurswahlverhalten bereits die Entscheidungen für die Wahl eines späteren (mathematisch-naturwissenschaftlichen) Universitätsstudiums vorhersagt. Das unterschiedliche Kurswahlverhalten von Mädchen und Burschen wird am stärksten durch vorangegangene Kompetenzen/Noten sowie motivationale Variablen vorhergesagt (z. B. Hemming, Grootenboer & Kay, 2011; Lazarides & Lauermaun, 2019; Parker, Nagy, Trautwein & Lüdtke, 2014). Motivationale Variablen spielen laut dem Erwartung-Wert-Modell von Wigfield und Eccles (2000) eine zentrale Rolle für Bildungsentscheidungen. Dieses Modell besagt, dass Entscheidungen auf einem komplexen Zusammenspiel zwischen Fähigkeiten, Erfolgserwartungen und Werten bezogen auf Aufgaben bzw. Entscheidungen basieren (z. B. Kampa, Krämer & Hannover, 2020). Die Wertkomponente wird noch einmal unterteilt in intrinsische Werte (z. B. Freude an Mathematik), den Nutzen (z. B. für eigene Bildungsziele), die Erreichbarkeit (z. B. Wichtigkeit guter Leistungen) und die Kosten (z. B. subjektive Kosten des Besuchs einer Schulform; Wigfield & Eccles, 2000).

Im Zusammenhang mit akademischem Wahlverhalten (z. B. Entscheidung für oder gegen ein MINT-Studium) werden wiederum geschlechtsspezifische Unterschiede näher untersucht, z. B. beleuchteten Gaspard, Dicke, Flunger und Brisson (2015) und Kampa et al. (2020) die Wahl mathematisch-naturwissenschaftlicher Fächer für die Abiturprüfungen (vs. mathematisch-naturwissenschaftsvermeidendes Wahlverhalten) und konnten zeigen, dass die Wahl der Prüfungen von beiden Geschlechtern sehr stark vom (naturwissenschaftlichen) Selbstkonzept und etwas schwächer von relativen Noten (relativ zu weiteren wählbaren Fächern für die Prüfungen) geprägt war.

Für die österreichische Mathematik-Zentralmatura ist die Schulform von Bedeutung, in der angetreten wird. Einerseits gibt es für die Mathematik Klausur unterschiedliche Varianten je nach BHS-Schulform, andererseits ist es in allen Schulformen außer den AHS und HTL möglich, nicht zur jeweiligen Variante der standardisierten schriftlichen Mathematik Klausur anzutreten, sondern auf eine individuelle mündliche auszuweichen (vgl. Abschnitt 2.4; BMBWF, 2023g). Im vorliegenden Kapitel wird daher neben geschlechtsspezifischen Unterschieden bei der Entscheidung für oder gegen eine maturaführende Schule auch genauer untersucht, welche Merkmale zur Wahl einer bestimmten maturaführenden Schulform beitragen. An bestehende Literatur anknüpfend, werden die leistungsbezogenen Merkmale Mathematikkompetenz sowie Mathematiknote einbezogen.

1.3 Geschlechterunterschiede bei der Reifeprüfung

Wir gehen zunächst darauf ein, wie viele Mädchen und Burschen die Reifeprüfung ablegen. Die Reifeprüfungsquote⁵ liegt bei den Mädchen aktuell wesentlich höher als bei den Burschen. So haben im Schuljahr 2020/21 39,4 % der Burschen und 54,6 % der Mädchen aller 18-/19-Jährigen die Reifeprüfung erfolgreich abgelegt (Statistik Austria, 2023). Die historische Entwicklung zeigt, dass bis in die 1970er-Jahre die Reifeprüfungsquote bei den Burschen höher lag als bei den Mädchen, in den 1980er-Jahren ein ziemlich ausgewogenes Geschlechterverhältnis vorlag und ab den 1990er-Jahren Mädchen häufiger eine Reifeprüfung absolvierten als Burschen (Neubacher & Wimmer, 2021, S. 294). Seit dem Schuljahr 2016/17 absolviert mehr als die Hälfte der 18-/19-jährigen Mädchen die Matura. Der Höchststand an Maturantinnen und Maturanten wurde im Jahr 2019/20 verzeichnet. Hier betrug die Maturaquote bei den Mädchen 57,9 % und bei den Burschen 42,4 %.

Seit 2012 liegt die Reifeprüfungsquote bei den Mädchen ca. um 14–15 Prozentpunkte höher als bei den Burschen. In den Jahren 2000 bis 2010 betrug der Unterschied in der Reifeprüfungsquote zwischen Burschen und Mädchen durchschnittlich 12 Prozentpunkte und in den 1990ern lag er zwischen 3,5 Prozentpunkten (1990/91) und 10,5 (1998/1999). Bemerkenswerte Veränderungen in der allgemeinen Reifeprüfungsquote in den letzten 10 Jahren sind ein Abfall in den beiden Schuljahren 2014/15 und 2015/16, der auf die Einführung der standardisierten Reife- und Diplomprüfung (2014/15 an AHS, 2015/16 an BHS) zurückzuführen ist und ein starker Anstieg im Schuljahr 2019/20, als während der Corona-Pandemie gelockerte Maturaregelungen eingeführt wurden (Statistik Austria, 2023).

Nicht nur in den Reifeprüfungsquoten, sondern auch in den Reifeprüfungsnoten zeigen sich Geschlechterunterschiede. Mit Bezug auf den Rechnungshofbericht von 2020, der empfiehlt, die Geschlechterunterschiede in Mathematik genauer zu untersuchen, wird hier auf die Noten von Mädchen und Burschen bei der schriftlichen Mathematiklausur eingegangen. Im Tabellenband der Statistik Austria (Hirt, Basas & Meisterhofer, 2023)⁶ zeigt sich, dass zum Haupttermin 2022 an den AHS bei den Mädchen 13 % der Mathematiklausuren mit Nicht genügend bewertet wurden, bei den Burschen waren es 8 %, in den BHS stehen 16 % Nicht genügend bei den Mädchen 7 % bei den Burschen gegenüber. Auf der anderen Seite stehen 17 % Sehr gut bei den AHS-Mädchen und 11 % Sehr gut bei den BHS-Mädchen 23 % Sehr gut bei den AHS-Burschen und 16 % Sehr gut bei den BHS-Burschen gegenüber. In allen Terminen seit 2017 (der erstmaligen Verfügbarkeit eines Tabellenbandes) schneiden Burschen (sowohl der AHS als auch der BHS) bei der Mathematiklausur besser ab als Mädchen (Bönisch, Maldet, Mayerweck & Zheng, 2019a, 2019b; Bönisch, Maldet & Zheng, 2020; Bönisch, Ernst, Hirt, Basas, & Zheng, 2021; Bönisch, Ernst, Hirt, Basas & Dentler, 2022; Hirt et al., 2023). Aus den Tabellenbänden sind jedoch weder Durchschnittsnoten noch Effektstärken ersichtlich, weshalb wir Sekundäranalysen durchgeführt haben, die zeigen, dass der standardisierte Geschlechterunterschied bei den Noten auf die Mathematik-

5 Bestandene Reifeprüfungen gemessen am arithmetischen Mittel der 18- und 19-jährigen Wohnbevölkerung.

6 Zum Zeitpunkt der Abfassung dieses Beitrags war dies der neueste Tabellenband.

klausur zum Haupttermin (ohne Kompensationsprüfung und Jahresnoteneinrechnung) zwischen $d = -0,22$ (AHS 2022) und $d = -0,38$ (BHS 2017) schwankt.⁷ Demnach ist der Notenschnitt der Burschen stets niedriger (d. h. besser) als jener der Mädchen und praktisch durchaus relevant. Mit der vorliegenden Studie untersuchen wir, inwiefern dieser relevante Unterschied durch unterschiedliche Selektionsmechanismen induziert ist.

2 Methoden

Wie in der Einleitung erwähnt, basiert die vorliegende Untersuchung auf einem Pseudopanel-Design und verfolgt wesentliche leistungsbezogene Bildungsindikatoren bei Personen, die 2002 bzw. 2003 geboren wurden, über mehrere Erhebungsmeilensteine hinweg. Tabelle 1 (siehe nächste Seite) zeigt, wann diese Erhebungen erfolgten und wie alt die 2002/03 Geborenen zu den unterschiedlichen Erhebungszeitpunkten waren. Tabelle 1 zeigt, dass der Burschenanteil bei jenen, die zur Zentralmatura antreten, wesentlich niedriger liegt (mit 40,7 % in den AHS und 45,6 % in den BHS) als zuvor bei der BIST-Ü oder bei PISA (mit jeweils 50,8 %), als noch fast alle der im Jahr 2002/2003 Geborenen abgedeckt waren (vgl. Tabelle 3 zur Größe der Wohnbevölkerung für diese Kohorte).

In der Folge werden die Studien BIST-Ü M8 2017, BIST-Ü M4 2013, PISA 2018 sowie die SRDP vorgestellt. Für PISA 2018 und die BIST-Ü erfolgt nur eine kurze Beschreibung, weil diese anderweitig bereits detailliert dargestellt wurden (OECD, 2020a; Schreiner et al., 2018; Schreiner & Breit, 2014; Schreiner & Wiesner, 2019). Der Ablauf und die Hintergründe der SRDP sowie insbesondere der Mathematiklausur werden genauer beschrieben, da unseres Wissens aktuell noch keine zusammenfassenden Überblicke veröffentlicht wurden.

2.1 Bildungsstandardüberprüfungen (BIST-Ü) M8 2017 und M4 2013

Bei BIST-Ü M8 2017 handelt es sich um eine österreichweite Bildungsstandardüberprüfung Mathematik auf der achten Schulstufe (jahrgangsbezogen) basierend auf dem österreichischen Lehrplan. Die Erhebung zeichnet sich dadurch aus, dass (fast) alle Schülerinnen und Schüler der achten Schulstufe in Österreich getestet werden. Ausgenommen sind nur Schülerinnen und Schüler mit sonderpädagogischem Förderbedarf (somit keine Testung an Sonderschulen und von Schülerinnen und Schülern, die in Mathematik nach dem Sonderschul-Lehrplan unterrichtet werden), außerordentliche Schülerinnen und Schüler sowie Schülerinnen und Schüler an Schulen ohne Öffentlichkeitsrecht. Die Mathematikskala wurde bei der Baseline-Erhebung im Jahr 2009 auf einen Österreich-Mittelwert von 500 und eine Standardabweichung von 100 verankert. Zusätzlich zur Mathematik-Überprüfung

7 Daten für den Haupttermin 2020 konnten nicht berücksichtigt werden, da keine Ergebnisse VOR Noteneinrechnung vorliegen. Für den Haupttermin 2022 ist im Internet nur ein Tabellenband ohne Ergebnisse bei der Mathematiklausur nach Geschlecht verfügbar. Eine (nicht barrierefreie) Langversion mit diesen Ergebnissen ist auf Anfrage bei der Statistik Austria erhältlich.

Tabelle 1: Schülermerkmale bei den einbezogenen Erhebungen

Erhebung	N (getestet)	N (gewichtet)	geboren (bei regulärer Einschulung ohne Laufbahnverzögerung)		Testzeitpunkt	Durchschnittsalter zum Testzeitpunkt	weiblich (%)	männlich (%)
BIST-Ü M4 2013	73.655	76.250	1. Sep. 2002	31. Aug. 2003	Mai 2013	10,3	49,1	50,9
BIST-Ü M8 2017	72.704	76.810	1. Sep. 2002	31. Aug. 2003	Mai 2017	14,4	49,2	50,8
PISA 2018	6.802	75.077	1. Jan. 2002	31. Dez. 2002 ^a	April/Mai 2018	15,8	49,2	50,8
SRDP 2021 – AHS (Haupttermin)	19.148	d	1. Sep. 2002	31. Aug. 2003	Mai 2021 ^c	18,4 ^b	59,3	40,7
SRDP 2022 – BHS (Haupttermin)	20.752	d	1. Sep. 2002	31. Aug. 2003	Mai 2022 ^c	19,5 ^b	54,4	45,6

BIST-Ü M4 Bildungsstandardüberprüfung Mathematik 4. Schulstufe, BIST-Ü M8 Bildungsstandardüberprüfung Mathematik 8. Schulstufe, PISA = Programme for International Student Assessment, SRDP = Standardisierte kompetenzorientierte Reifeprüfung bzw. Reife- und Diplomprüfung.

^a Zielpopulation sind Schülerinnen und Schüler, die im gegebenen Zeitraum geboren wurden und mind. die 7. Schulstufe besuchen.

^b Median des Abschlussalters laut Statistik Austria (2023, S. 69).

^c Standardisierte schriftliche Klausuren.

d = Nicht relevant, weil Getestete und Zielpopulation ident sind.

wurden Kontextfragebögen für Schülerinnen und Schüler und Schulleiter/innen eingesetzt, in denen eine Reihe an Hintergrundinformationen erhoben wurden. Sehr ähnlich, nur auf der vierten Schulstufe (4. Klasse Volksschule) war die BIST-Ü M4 2013 angelegt. Bei dieser gab es darüber hinaus Fragebögen für Lehrkräfte und Eltern. Genauere Informationen zur BIST-Ü M4 2013 finden sich bei Schreiner und Breit (2014), Details zur BIST-Ü M8 2017 bei Schreiner et al. (2018) sowie allgemeine Informationen zur Konzeption der BIST-Überprüfungen bei Schreiner und Wiesner (2019).

2.2 Programme for International Student Assessment (PISA) 2018

PISA ist eine von der OECD organisierte internationale Erhebung und folgt einem altersbasierten Erhebungsdesign. Zielpopulation sind Schülerinnen und Schüler des Geburtsjahrgangs 16 Jahre vor der Erhebung. Bei PISA 2018 war dies der Geburtsjahrgang 2002. Ausgenommen sind Schülerinnen und Schüler, die sich noch unter der siebten Schulstufe befinden. Ausgeschlossen werden können außerordentliche Schülerinnen und Schüler aufgrund mangelnder Sprachkenntnisse sowie Schülerinnen und Schüler mit körperlichen, kognitiven oder emotionalen Beeinträchtigungen, wenn ihnen selbst der eigens für Sonderschülerinnen und -schüler konzipierte Test nicht zumutbar ist (Suchaň & Höller, 2019).

In Österreich wurden bei PISA 2018 6802 Schülerinnen und Schüler getestet. Zur Zeit der Testung im April und Mai 2018⁸ befanden sich die meisten dieser Schülerinnen und Schüler auf der neunten (44,5 %) oder zehnten Schulstufe (48,1 %) und damit im ersten oder zweiten Jahr der Sekundarstufe II. Jene Jugendlichen, die nach neun Jahren Schulpflicht keine Schule mehr besuchen, werden bei PISA nicht berücksichtigt.

PISA folgt einem Dreijahresrhythmus. Es werden zu jeder Erhebung die Kompetenzen der Schülerinnen und Schüler in Lesen, Mathematik und Naturwissenschaften getestet. PISA erhebt lebensnahe Kompetenzen, die nicht mit den in nationalen Lehrplänen definierten Fähigkeiten übereinstimmen müssen (Höller & Toferer, 2019; OECD, 2019). Für das vorliegende Kapitel ist Mathematik relevant. Die Mathematikskala wurde im Jahr 2003 auf einen OECD-weiten Mittelwert von 500 und eine Standardabweichung von 100 verankert. Diese Verankerung gilt für alle nachfolgenden Erhebungen.

2.3 Die standardisierte Reife- und Diplomprüfung (SRDP)

Die *teilstandardisierte kompetenzorientierte Reifeprüfung*⁹ ist an AHS seit dem Schuljahr 2014/2015 flächendeckend eingeführt. An den BHS ist die *teilstandardisierte kompetenz-*

⁸ In saisonalen Berufsschulen war bereits eine Testung ab März möglich.

⁹ Obwohl es an den AHS keine Diplomprüfung gibt, wird in diesem Kapitel auch für die AHS die Abkürzung SRDP verwendet und damit der Vorgehensweise des BMBWF (z. B. 2023b) gefolgt.

orientierte Reife- und Diplomprüfung seit dem Schuljahr 2015/2016 flächendeckend eingeführt (BMBWF, 2023a).

Die frühere Form der Reife- und Diplomprüfung (vor Einführung der SRDP) war ausschließlich auf den jeweiligen Schulstandort bezogen und trug der zunehmenden Kompetenzorientierung im Schulwesen sowie internationalen Entwicklungen in Richtung (teil-)standardisierter Prüfungsform nicht ausreichend Rechnung (Rechnungshof Österreich, 2020). Ziel der „Teilstandardisierung“ ist es, die Klausurprüfungen unabhängig von einzelnen Lehrenden auf ein gemeinsames österreichisches Qualitätsniveau zu bringen und damit die Objektivität von Maturaergebnissen zu steigern. Die Grundidee der standardisierten Reife- und Diplomprüfung umfasst demnach die Standardisierung (der schriftlichen Klausuren) und die Kompetenzorientierung, durch die Kandidatinnen und Kandidaten ihre Teilprüfungen hinsichtlich klar definierter Anforderungen ablegen (BIFIE, 2013). Das BMBWF (2023a) geht davon aus, dass die SRDP mehr Fairness, gleiche Bedingungen für alle Maturantinnen und Maturanten und eine leichtere Vergleichbarkeit der Bildungsabschlüsse für weiterführende Bildungseinrichtungen bringt.

Strukturell besteht die Zentralmatura aus drei Säulen: die abschließende *schriftliche Arbeit* (vorwissenschaftliche Arbeit bzw. Diplomarbeit), die in den meisten Gegenständen zentral erstellte *Klausurprüfung* (auch „schriftliche Prüfung“ genannt) und die standortbezogene *mündliche Prüfung* (BMBWF, 2023c, 2023d). Analog dazu werden im vorliegenden Kapitel die Ausdrücke „Klausur“, „Klausurprüfung“ und „schriftliche Prüfung“ bedeutungsgleich verwendet. Wenn nur von „Prüfungen“ die Rede ist, bezieht dies sowohl schriftliche als auch mündliche Prüfungen mit ein.

Zentral erstellte Klausurprüfungen werden in den vier Fächergruppen Deutsch, lebende Fremdsprachen, klassische Fremdsprachen sowie Mathematik für die AHS und die Angewandte Mathematik für die BHS eingesetzt. Weitere schriftliche Klausuren, wie z. B. aus Fachtheorie oder Betriebswirtschaft/Rechnungswesen (vgl. Tabelle 1), werden von den Klassenlehrpersonen erstellt. Die Lehrpersonen der Abschlussklassen korrigieren sämtliche schriftlichen – auch zentral erstellten – Klausuren (Rechnungshof Österreich, 2020). Die Korrektur und Beurteilung der standardisierten Klausuren erfolgt dabei nach einem zentral vorgegebenen Korrektur- und Beurteilungsschlüssel (z. B. BMBWF, 2023f). Es besteht die Möglichkeit, im Fall einer negativ benoteten Klausur eine zentral erstellte mündliche Kompensationsprüfung zum gleichen Termin abzulegen oder, wenn das Ergebnis weiterhin negativ ist, an drei weiteren Terminen anzutreten – jeweils mit zentral erstellten schriftlichen Klausuren.

In den Jahren 2020–2021 gab es Corona-bedingte Änderungen bei der Matura. Unter anderem war die mündliche Matura freiwillig, es wurde die Jahresnote in die Gesamtnote eingerechnet, die Anzahl der verpflichtend zu absolvierenden Klausurprüfungen wurde auf drei beschränkt und die Arbeitszeit bei den Klausuren wurde um eine Stunde verlängert (BMBWF, 2020a, 2021). Diese Änderungen wurden in den Jahren 2022 und 2023 größtenteils wieder zurückgenommen. Auch im Jahr 2023 beibehalten wurde allerdings die

Einbeziehung der Jahresnote in die Gesamtnote (BMBWF, 2022a). Diese Regelung schützt Kandidatinnen und Kandidaten mit einem Befriedigend oder einer besseren Note im Jahreszeugnis vor einem Durchfallen bei der Matura, sofern zumindest 30 Prozent der Punkte bei der schriftlichen Prüfung erreicht wurden.

Im vorliegenden Kapitel wird auf die Ergebnisse bei der schriftlichen Mathematiklausur fokussiert, weil der Rechnungshofbericht speziell in dieser Domäne Erklärungsbedarf bezüglich der Geschlechterunterschiede ortet. In der Folge wird die schriftliche Prüfung in Mathematik genauer beschrieben.

2.4 Die Klausur (schriftliche Prüfung) in Mathematik der SRDP

Die Mathematiklausur (d. h. die schriftliche Prüfung in Mathematik) wird immer zentral erstellt (für alle BHS und AHS). Allerdings gibt es Unterschiede dahingehend, ob ein Antreten bei der Mathematiklausur verpflichtend ist. Während in den AHS und den HTL alle Kandidatinnen und Kandidaten zur Mathematiklausur antreten müssen, gibt es in den weiteren BHS Wahlmöglichkeiten (vgl. Tabelle 2 auf der nächsten Seite; BMBWF, 2023d). So können die Schülerinnen und Schüler der höheren land- und forstwirtschaftlichen Schulen, der Handelsakademien sowie der Bildungsanstalten für Elementar- bzw. Sozialpädagogik zwischen einem schriftlichen Antreten in Angewandter Mathematik oder in einer lebenden Fremdsprache entscheiden. Bei den humanberuflichen Schulen können zwei Klausuren aus folgenden drei gewählt werden: (a) lebende Fremdsprache und/oder (b) Angewandte Mathematik und/oder (c) Angewandte Betriebswirtschaft und Rechnungswesen.

Demnach besteht in allen BHS ausgenommen den HTL die Möglichkeit, die Mathematiklausur abzuwählen. Für alle Schulformen mit Wahlmöglichkeit gilt: Der Gegenstand, der nicht schriftlich gewählt wurde, muss mündlich absolviert werden. Die mündliche Prüfung wird von den Lehrkräften am Schulstandort erstellt und ist demnach nicht standardisiert (BMBWF, 2023g). Im vorliegenden Kapitel wird unter anderem beleuchtet, ob Mädchen die Abwahlmöglichkeit der Mathematiklausur und damit das Antreten zu einer nicht standardisierten mündlichen Prüfung in Mathematik häufiger nutzen als Burschen.

Inhaltlich beruhen die Mathematiklausuren für AHS und BHS auf unterschiedlichen Grundkonzepten (Aue et al., 2021; BMBWF, 2019). Damit geht einher, dass alle AHS-Schülerinnen und -Schüler dieselbe Mathematiklausur erhalten, es in den BHS allerdings schulformspezifische Unterschiede gibt. So gliedern sich dort die standardisierten schriftlichen Klausurprüfungen in Angewandter Mathematik in einen Teil A, der schulformübergreifend ist, und einen Teil B, der schulform- bzw. clusterspezifisch ist (BMBWF, 2023e). Teil A und Teil B gehen jeweils ungefähr zur Hälfte in die Gesamtpunktezahl ein. Zum Haupttermin 2022 konnten in Angewandter Mathematik im A-Teil 25 Punkte erreicht werden und im B-Teil 23.¹⁰

10 Die Autorinnen danken Martin Hofer für diese Information.

Tabelle 2: Übersicht über die Teilprüfungen der SRDP an BHS (Quelle: BMBWF, 2023d)

Höhere technische Lehranstalten (HTL)	Handelsakademien (HAK)	Humanberufliche Schulen (HUM)	Höhere land- und forstwirtschaftliche Schulen (HLFS)	Bildungsanstalten für Elementar- bzw. Sozialpädagogik (BAFEP/BASOP)
1. Diplomarbeit ¹⁾				
2. Schriftliche Klausur				
1. Deutsch und/oder Englisch ³⁾ 2. Mathematik ²⁾ 3. Fachtheorie	1. Deutsch 2. Lebende Fremdsprache und/oder Mathematik ^{2,3)} 3. Mathematik ^{2,3)} 4. Betriebswirtschaftliche Fachklausur	1. Deutsch 2. Lebende Fremdsprache, Ang. Mathematik ²⁾ oder BWL/RW ³⁾ (2 aus 3 Fächern)	1. Deutsch 2. Lebende Fremdsprache und/oder Mathematik ^{2,3)} 3. Betriebswirtschaft und RW	1. Deutsch 2. Lebende Fremdsprache und/oder Ang. Mathematik ^{2,3)} 3. Fachtheorie ³⁾
3. Mündliche Prüfung				
1. Technischer Schwerpunkt ⁴⁾ 2. Englisch oder Deutsch ⁵⁾ 3. Wahlfach ⁴⁾	1. Betriebswirtschaftliches Kolloquium 2. Mathematik oder lebende Fremdsprache ⁵⁾ 3. Wahlfach ⁴⁾	1. Fachkolloquium ⁶⁾ 2. Lebende Fremdsprache, Ang. Mathematik oder BWL/RW ⁵⁾ 3. Wahlfach ⁴⁾	1. Fachkolloquium ⁶⁾ 2. Mathematik oder lebende Fremdsprache ⁵⁾ 3. Wahlfach ⁴⁾	1. Fachtheorie ⁵⁾ 2. Lebende Fremdsprache oder Ang. Mathematik ⁵⁾ 3. Wahlfach oder Berufsspezifisches Prüfungsgebiet ⁶⁾
Prinzip: 3 schriftliche + 3 mündliche oder 4 schriftliche + 2 mündliche Teilprüfungen				
¹⁾ Schriftliche Arbeit inkl. deren Präsentation und Diskussion. ²⁾ Angewandte Mathematik (typenspezifisch). ³⁾ Nicht gewählter Unterrichtsgegenstand (mündlich). ⁴⁾ Von den Prüfungskandidatinnen und Prüfungskandidaten gemäß Vorgaben der Prüfungsordnung BMHS wählbar. ⁵⁾ Schriftlich nicht gewählter Unterrichtsgegenstand. ⁶⁾ Umfasst einen oder mehrere fachtheoretische Pflichtgegenstände. RW = Rechnungswesen; BWL = Betriebswirtschaftslehre; Ang. = Angewandte.				

Teil B liegt in fünf unterschiedlichen Varianten (sogenannten Clustern) vor, wobei es zwei Cluster für die HTL (HTL 1 und HTL 2) gibt, in einem weiteren werden humanberufliche Schulen und höhere land- und forstwirtschaftliche Schulen zusammengefasst (W 1), ein Cluster umfasst die Handelsakademien (W 2) und einer die Bildungsanstalten für Elementarpädagogik bzw. Sozialpädagogik sowie die Berufsreifeprüfung (Cluster P; BMBWF, 2022b).

Zusammenfassend kann festgestellt werden, dass es für die AHS-Mathematik Klausur nur eine Variante gibt, die Mathematik Klausur an allen AHS verpflichtend ist und nicht abgewählt werden kann. In den BHS gibt es fünf unterschiedliche Varianten der Klausur in Angewandter Mathematik, das Antreten zur Mathematik Klausur ist nur in den HTL verpflichtend, in den weiteren BHS-Schulformen kann ausgewichen werden auf eine nicht standardisierte mündliche Prüfung in Angewandter Mathematik.

2.5 Fragestellung und Datengrundlage

Im vorliegenden Kapitel wird der Frage nachgegangen, inwiefern dieser Gender-Gap bei der Mathematik Klausur der SRDP darauf zurückzuführen ist, dass weniger Burschen als Mädchen eines Jahrgangs zur Matura antreten und es sich somit bei den Burschen um eine stärker leistungsselektierte Gruppe handelt als bei den Mädchen. Insbesondere gehen wir darauf ein, inwiefern diese Unterschiede durch ungleiche Verteilungen von Schülerinnen und Schülern auf Schulformen der Sekundarstufe II begründet werden können. Des Weiteren beleuchten wir, welche Rolle sowohl die Mathematikkompetenz als auch die Mathematiknote für den Übertritt in eine maturaführende Schulform spielen.

Datenbasis für die vorliegenden Analysen sind (a) der imputierte Schülerdatensatz der BIST-Ü M8 2017, (b) der nationale (österreichische) Schülerdatensatz von PISA 2018, der auch nationale Zusatzvariablen wie z. B. Schulsparten inkludiert, (c) der Tabellenband zur SRDP 2021 der Statistik Austria (Bönisch et al., 2022; Hirt et al., 2023) sowie (d) Sonderauswertungen des BMBWF zu Maturanoten und -quoten getrennt nach HTL und weiteren BHS zur SRDP 2022 und (e) der STATcube (Statistische Datenbank von Statistik Austria) für Daten zur Größe der Wohnbevölkerung nach Geschlecht. Umfang und Merkmale der bei der BIST-Ü M8, PISA 2018 und der SRDP 2021 (AHS) sowie der SRDP 2022 (BHS) getesteten Schülerinnen und Schüler können Tabelle 1 entnommen werden.

Einige Variablen aus den Originaldatensätzen wurden für die vorliegenden Analysen rekodiert. Diese werden hier genauer beschrieben. Genauere Informationen zu den nicht rekodierten Variablen finden sich bei BIFIE (2018) für die BIST-Ü M8 2017, bei BIFIE (2017) für die BIST-Ü M4 2013 sowie bei der OECD (2020b) für die internationalen Variablen bei PISA 2018.

Schulform nach der Sekundarstufe I: Bei der BIST-Ü M8 2017 wurden die Schülerinnen und Schüler gefragt „In welche Schule wirst du nach diesem Jahr gehen?“. Ihnen standen

insgesamt acht Optionen zur Verfügung, die zu den zwei Kategorien maturaführende Schule (AHS, HTL und weitere BHS) und nicht maturaführende Schule (Polytechnische, Berufs-, berufsbildende mittlere Schule, keine weiterführende Schule) zusammengefasst wurden. Die Kategorie „weiß noch nicht“ (3,7 %) wurde der letzteren Kategorie zugeordnet. Demnach werden 67 % der Schülerinnen und Schüler eine maturaführende und 33 % keine maturaführende Schule besuchen.

Mathematiknote (BIST-Ü): Informationen zu den Noten stammen ebenfalls aus dem BIST-Ü-M8-2017-Kontextfragebogen für Schülerinnen und Schüler und gliedern sich in zwei Fragen. Einerseits wird die Mathematiknote im letzten Semesterzeugnis erhoben, andererseits, auf welcher Grundlage diese Benotung erfolgte. In Anlehnung an die Vorgaben des BMBWF (2020b) sowie an Bacher et al. (2008) wurden Notenwerte ermittelt, um die Noten über unterschiedliche Schulformen und Leistungsgruppen vergleichbar zu machen. Noten auf Basis „grundlegender Allgemeinbildung“ der NMS sowie der zweiten Leistungsgruppe Hauptschule¹¹ wurden um zwei Notenpunkte herabgesetzt und Noten der dritten Leistungsgruppe einer Hauptschule um vier. Dadurch ergibt sich als bester Notenwert eine 1 (entspricht einem „Sehr gut“ in der AHS oder der NMS in „Vertiefter Allgemeinbildung“ oder der ersten Hauptschul-Leistungsgruppe) und als schlechtester Notenwert eine 9 (entspricht einem „Nicht genügend“ in der dritten Hauptschul-Leistungsgruppe). Ein Übertritt in eine höhere Schule ohne Aufnahmeprüfung ist nur möglich, wenn der Notenwert im Jahreszeugnis der achten Schulstufe in Deutsch, Mathematik und Englisch 4 oder besser ist (vgl. BMBWF, 2018). Die BIST-Schülerinnen und -Schüler berichten den Notenwert des Semesterzeugnisses. Da eine Verbesserung des Notenwerts um eine Stufe von Semester- auf Jahreszeugnis durchaus möglich ist, könnten auch noch Schülerinnen und Schüler mit Notenwert 5 im Semesterzeugnis die Notenvoraussetzungen für den Besuch einer höheren Schule schaffen. Eine Verbesserung um zwei Notenstufen ist allerdings unwahrscheinlich. Für die vorliegenden Analysen werden daher alle Notenwerte von 6 oder höher zu einer Kategorie zusammengefasst.

Note bei der Mathematiklausur der SRDP: Die Angaben aus dem SRDP-Tabellenband 2021 (Bönisch et al., 2022) zur Anzahl an Burschen und Mädchen mit unterschiedlichen Noten bei der Klausur in Mathematik (AHS) wurden in ein SPSS-Datenfile überführt, um Notenmittelwerte und Standardabweichungen zu berechnen. Konkret wurden die Noten zum Haupttermin (Mai/Juni) vor einer eventuellen Kompensationsprüfung und ohne Jahresnoteneinrechnung entnommen. Für die unterschiedlichen BHS-Schulformen wurden die entsprechenden Angaben für den Haupttermin 2022 vom BMBWF in Form einer Sonderauswertung zur Verfügung gestellt.

Alle Analysen zur BIST-Ü und zu PISA erfolgen mit gewichteten Daten. Ausgefallene und nicht getestete Schülerinnen und Schüler werden durch eine Höhergewichtung von verbleibenden Schülerinnen und Schülern mit ähnlichen Merkmalen kompensiert (Bruneforth,

11 Bei der BIST-Ü M8 2017 wurden nur mehr wenige Schülerinnen und Schüler (9,3 %) nach dem Hauptschul-Lehrplan unterrichtet.

Oberwimmer & Robitzsch, 2016; OECD, 2009). Die gewichtete Anzahl entspricht damit der gesamten Schüleranzahl in der Zielpopulation (Tabelle 1 zur Gegenüberstellung von gewichteter und ungewichteter Schüleranzahl).

3 Ergebnisse

3.1 Anteile an Mädchen und Burschen bei der Wahl maturaführender Schulen sowie bei der Matura

Bezogen auf die 14-jährige Wohnbevölkerung gaben bei der BIST-Ü 2017 67 % der Mädchen und 54 % der Burschen an, dass sie nach dem laufenden Schuljahr in eine maturaführende Schule gehen werden. Bei PISA befinden sich bezogen auf die 16-jährige Wohnbevölkerung 61 % der Mädchen und 48 % der Burschen tatsächlich an einer maturaführenden Schule (Summe AHS & BHS in Tabelle 3 auf der nächsten Seite). Demnach sind bei PISA, bei dessen Testung sich die Schülerinnen und Schüler in der Regel im ersten oder zweiten Jahr der Sekundarstufe II befinden, um ca. sechs Prozentpunkte weniger Schülerinnen und Schüler an einer maturaführenden Schule anzutreffen, als vorher bei der BIST-Ü angegeben, dorthin zu wechseln. Eine weitere Selektierung¹² ist im Verlauf der Karriere an den höheren Schulen erkennbar, was schließlich dazu führt, dass 52 % der 2002/2003 geborenen Mädchen und 37 % der Burschen als 18-jährige AHS-Schülerinnen und -Schüler beim Haupttermin 2021 oder als 19-jährige BHS-Schülerinnen und -Schüler beim Haupttermin 2022 zur Matura angetreten sind.¹³

Darüber hinaus zeigt Tabelle 3, dass unter allen Maturakandidatinnen und -kandidaten Mädchen (mit 90 %) seltener zur schriftlichen Klausur in Mathematik antreten als Burschen (95 %), was vor allem damit erklärt werden kann, dass an allen BHS-Schulformen (außer den HTL) die schriftliche Mathematik Klausur abgewählt werden kann und der Mädchenanteil (73 %) an diesen Schulen besonders hoch ist. So treten an den BHS (ausgenommen die HTL) nur 80 % der Maturakandidatinnen zur Mathematik Klausur an. Dass auch in den AHS und HTL nicht alle, die zur Matura antreten, auch zur Mathematik Klausur antreten (sondern nur ca. 97 %), könnte damit zusammenhängen, dass manche Schülerinnen und Schüler bewusst einzelne Gegenstände auf einen Nebetermin der Matura (meist Herbst) verschieben.¹⁴

Tabelle 3 zeigt zudem, dass die angetretenen Mädchen die Matura zum Haupttermin etwas häufiger erfolgreich bestehen (zu 92 %) als die angetretenen Burschen (90 %). Die Zahl der

12 Vgl. Statistik Austria (2023, S. 61) zu den Abbruchsquoten an maturaführenden Schulen.

13 Diese Quoten unterscheiden sich von den bei Statistik Austria (2023) publizierten, weil hier die Angetretenen und nicht die Erfolgreichen (= Bestandenen) einbezogen werden. Darüber hinaus rechnet die Statistik Austria erfolgreiche Reifeprüfungen über das ganze Jahr ein (auch zu Nebeterminen) und es wird auch die Berufsreifeprüfung einbezogen.

14 Die Autorinnen danken Markus Bönisch für diesen Hinweis.

Tabelle 4: Mathematikkompetenz bei standardisierten Tests von Schülerinnen und Schülern (geboren 2002/2003), die in eine maturaführende Schule gehen werden (BIST-Ü), eine solche besuchen (PISA) bzw. zur Mathematiklausur der SRDP angetreten sind

	AHS		BHS: HTL		BHS: weitere		AHS & BHS gesamt		Österreich gesamt	
	M	SD	M	SD	M	SD	M	SD	M	SD
BIST-Ü M4 2013^a										
weiblich	573	86,92							526	98,25
männlich	591	91,17							540	100,96
gesamt	582	89,46							533	99,90
Differenz (d)	18	(0,20)							14	(0,14)
BIST-Ü M8 2017^b										
weiblich	588	82,52	569	84,13	537	80,25	563	85,26	538	91,84
männlich	603	86,45	583	85,85	542	84,00	582	88,49	545	96,23
gesamt	594	84,49	580	85,69	538	81,35	572	87,30	542	94,16
Differenz (d)	15	(0,18)	14	(0,16)	6	(0,07)	19	(0,22)	7	(0,08)
PISA 2018^c										
weiblich	539	74,96	546	65,80	508	69,15	525	73,45	492	88,84
männlich	565	79,57	570	71,00	521	77,97	557	78,54	505	97,35
gesamt	550	78,06	565	70,71	512	71,90	540	77,51	499	93,50
Differenz (d)	26	(0,33)	24	(0,34)	13	(0,18)	33	(0,42)	13	(0,14)
SRDP 2021 (AHS) bzw. 2022 (BHS)^e										
weiblich	2,82	1,34	2,84	1,18	3,24	1,24	2,99	1,31	Ident mit AHS & BHS gesamt.	
männlich	2,51	1,27	2,57	1,10	3,15	1,21	2,65	1,22		
gesamt	2,70	1,32	2,63	1,12	3,22	1,23	2,84	1,28		
Differenz (d)	-0,31	(-0,23)	-0,26	(-0,24)	-0,09	(-0,07)	-0,33	(-0,26)		

Anmerkungen: BIST-Ü M8 = Bildungsstandardüberprüfung Mathematik 8. Schulstufe, PISA = Programme for International Student Assessment, AHS = Allgemeinbildende höhere Schulen, BHS = Berufsbildende höhere Schulen, HTL = Höhere technische Lehranstalten, SRDP = Standardisierte kompetenzorientierte Reifeprüfung bzw. Reife- und Diplomprüfung, Differenz = männlich-weiblich, Cohen's d steht in Klammern, ^a Kompetenzwerte von Schülerinnen und Schülern, deren Eltern angeben, dass sie in die entsprechende Schulform übertreten werden; ^b Kompetenzwerte von Schülerinnen und Schülern, die angeben, dass sie in die entsprechende Schulform übertreten werden; ^c Kompetenzwerte von Schülerinnen und Schülern, die sich in der entsprechenden Schulform befinden, ^d Noten auf die Mathematiklausur der SRDP 2021 (AHS) bzw. auf die Klausur in Angewandter Mathematik der SRDP 2022 (BHS).

Quelle: BIST-Ü M4 2013 Datensatz; BIST-Ü M8 2017 Datensatz; PISA 2018 Österreich-Datensatz; SRDP 2021 (AHS); Bönisch et al. (2022); SRDP 2022 (BHS); Sonderauswertung der Statistik Austria.

„erfolgreich Bestanden“ beinhaltet auch jene Kandidatinnen und Kandidaten, die einzelne standardisierte Klausuren nicht bestanden haben, aber dies durch Noteneinrechnung bzw. mündliche Kompensationsprüfung(en) ausgleichen konnten, und bezieht sich auf alle Maturagegenstände (nicht nur Mathematik).

3.2 Geschlechterunterschiede bei standardisierten Mathematiktests

Tabelle 4 zeigt, dass es auf der vierten Schulstufe Geschlechterunterschiede in der Mathematikkompetenz zugunsten von Buben gibt, die etwa gleich hoch ausfallen wie bei PISA ($d = 0,14$). Zwischenzeitlich (auf der achten Schulstufe) fallen Geschlechterunterschiede etwas niedriger aus ($d = 0,08$). Dies könnte daran liegen, dass Mädchen häufiger als Buben nach der Volksschule in eine AHS übertreten (vgl. Abschnitt 1.2) und somit mehr Mädchen als Buben anspruchsvolleren AHS-Unterricht erhalten. Bemerkenswert ist, dass es schon bei 14-jährigen Schülerinnen und Schülern, die eine maturaführende Schule (AHS oder BHS) anstreben, deutliche Geschlechterunterschiede in der Mathematikkompetenz gibt und die Geschlechterunterschiede in dieser Gruppe mit $d = 0,22$ wesentlich größer ausfallen als in der Population der Schülerinnen und Schüler der achten Schulstufe ($d = 0,08$). Bei PISA (meist neunte oder zehnte Schulstufe) zeigt sich eine weitere Verschärfung des Geschlechterunterschieds bei Schülerinnen und Schülern an maturaführenden Schulen ($d = 0,42$) in Relation zur Gesamtpopulation ($d = 0,14$). Es fällt auf, dass der Geschlechterunterschied in der Mathematikkompetenz bei PISA in den maturaführenden Schulen größer ausfällt als später bei der Matura. Dies könnte einerseits damit zusammenhängen, dass die Mathematikkompetenz bei PISA differenzierter bewertet wird als bei der Matura (500er-Skala vs. 5-stufige Notenskala), womit auch kleinere Unterschiede sichtbar werden. Andererseits könnte es auch damit zu tun haben, dass PISA nicht lehrplanbasiert testet, sondern die funktionelle Anwendung mathematischen Wissens in unterschiedlichen Situationen in den Vordergrund stellt (d.h., inwiefern Schülerinnen und Schüler ihre Kenntnisse in lebensnahen Aufgaben anwenden können; vgl. Höller & Toferer, 2019; van den Ham, Ehmke, Hahn, Wagner & Schöps, 2016). Mädchen könnten sich mit stärker funktionell orientierten Aufgaben schwerer tun, weil sie in ihren täglichen Aktivitäten weniger mit Mathematik konfrontiert sind bzw. weniger in der Lage sein könnten, unterschiedliche Wissensquellen zu integrieren und theoretisches Wissen auf unterschiedliche Situationen und Kontexte anzuwenden (vgl. Else-Quest et al., 2010). Allerdings liegen die Geschlechterunterschiede bei PISA in Österreich auch im internationalen Vergleich im Spitzenfeld (z. B. Brandmair & Schmoller, 2023). Dies deutet darauf hin, dass sich österreichische Mädchen mit lebensnahen Aufgaben besonders schwer tun und/oder österreichische Burschen besonders leicht.

Im Vergleich der dargestellten Schulformen wird deutlich, dass Schülerinnen und Schüler mit durchschnittlicher Mathematikkompetenz eine weitere BHS anstreben, während die AHS und HTL im Schnitt von Schülerinnen und Schülern mit relativ hoher Mathematikkompetenz angestrebt werden. Ganz ähnlich zeigt sich bei PISA, dass sich die Schülerinnen und Schüler mit besonders hoher Mathematikkompetenz an AHS und HTL befinden und die Mathematikkompetenz der Schülerinnen und Schüler an weiteren BHS deutlich unter

jenen von AHS- und HTL-Schülerinnen und -Schülern liegt. Wie in Tabelle 3 dargestellt, ist in den weiteren BHS der Mädchenanteil mit 73 % besonders hoch. Es sind also vor allem Mädchen, die sich von den weiteren BHS angezogen fühlen und solche besuchen.

Wie Tabelle 5 zeigt, ist die Mathematikunterrichtszeit in den weiteren BHS wesentlich geringer als in den AHS oder HTL. Die weiteren BHS ziehen damit nicht nur Mädchen mit – im Vergleich zu den übrigen Schülerinnen und Schülern höherer Schulen – relativ schwachen Mathematikkompetenzen an, Mathematik wird an diesen Schulen auch deutlich weniger unterrichtet. Zudem kann an diesen Schulen die standardisierte schriftliche Mathematikklausur bei der Matura abgewählt und durch eine nicht standardisierte mündliche Prüfung ersetzt werden. Und obwohl diese Schülerinnen und Schüler, die Mathematik schriftlich abgewählt haben, nicht in den SRDP-Analysen zu Tabelle 4 enthalten sind, zeigt sich trotzdem für die Kandidatinnen/Kandidaten der weiteren BHS ein deutlich schlechterer Notenschnitt bei der Mathematikklausur (3,22) als für die Kandidatinnen/Kandidaten der AHS (2,70) oder HTL (2,63). In Tabelle 4 ist zu beachten, dass bei der BIST-Ü und PISA höhere Werte für eine höhere Mathematikkompetenz stehen, bei der SRDP aber ein niedrigerer Notenwert eine höhere Kompetenz indiziert. Obwohl also die Differenzwerte (Burschen minus Mädchen) bei der BIST-Ü und PISA positiv sind und bei der Matura negativ, bedeutet es in allen Fällen, dass Burschen besser abgeschnitten haben als Mädchen.

Insgesamt bewegen sich bei der Matura die Geschlechterunterschiede in der Mathematikleistung in einem Bereich, wie er aufgrund der bei der BIST-Ü M8 2017 sowie der bei PISA 2018 gemessenen Geschlechterunterschiede zu erwarten war. Dieses Ergebnis lässt vermuten, dass Geschlechterunterschiede bei der Mathematikklausur zu einem wesentlichen Teil durch Selektionsprozesse und vorgeschaltete Entscheidungsprozesse entstehen und entkräftet damit zu einem gewissen Teil die Kritik bezüglich der Genderfairness der Zentralmatura in Mathematik. Die vorliegende Untersuchung zeigt vielmehr, dass diese Geschlechterunterschiede bis auf die Sekundarstufe I zurückreichen, und damit in eine Zeit, in der Entscheidungen über den weiteren Schulweg getroffen werden. Das österreichische Schulsystem eröffnet auf der Sekundarstufe II ein breites Feld unterschiedlicher Schulformen mit sehr unterschiedlichem Fokus auf Mathematik. Am Ende der Sekundarstufe I entscheiden sich Schülerinnen und Schüler mit durchschnittlicher Mathematikkompetenz, die die Matura anstreben, häufig für eine weitere BHS, jene mit hoher Mathematikkompetenz für eine AHS oder HTL. In weiterer Folge erhalten Erstere an den weiteren BHS weniger Mathematikunterricht (Tabelle 5 auf der nächsten Seite) und schneiden schlussendlich auch bei der Matura schlechter ab. Dadurch, dass Mädchen über Schulformen mit unterschiedlicher Mathematikintensität hinweg unterschiedlich stark vertreten sind (hoher Anteil mit über 70 % in den weiteren BHS, sehr geringer in den HTL mit ca. 20 %), sind es häufiger Mädchen, die bei der Mathematikklausur schwach abschneiden. Doch das schlechtere Abschneiden der Mädchen ist nicht nur auf die unterschiedliche Verteilung nach Schulformen zurückzuführen, denn innerhalb der einzelnen Schulformen zeigt sich ebenso ein Vorteil der Burschen. Dieser ist innerhalb der weiteren BHS klein und vernachlässigbar ($d = +/ - 0,07$ bei der BIST-Ü und der SRDP, $0,18$ bei PISA), aber in den AHS ($d = -0,23$ bei der SRDP) und HTL ($d = -0,24$ bei der SRDP) deutlicher ausgeprägt. In den HTL könnte

Tabelle 5: Anzahl der Unterrichtsminuten in Mathematik pro Woche in Schulen der Sekundarstufe II laut Angaben der PISA-Schülerinnen und -Schüler. Quelle: PISA-2018-Österreich-Datensatz

	AHS		BHS: HTL		BHS: weitere		AHS & BHS gesamt		Österreich gesamt	
	M	SD	M	SD	M	SD	M	SD	M	SD
weiblich	169	28	149	32	106	24	138	41	137	66
männlich	175	43	169	52	108	31	158	52	165	84
gesamt	172	35	165	50	107	26	147	47	150	76
Differenz (d)	5	(0,15)	20	(0,41)	2	(0,09)	20	(0,43)	28	(0,37)

Anmerkungen: PISA = Programme for International Student Assessment, AHS = Allgemeinbildende höhere Schulen, BHS = Berufsbildende höhere Schulen, HTL = Höhere technische Lehranstalten, Differenz = männlich–weiblich, Cohen's d steht in Klammern; die von den Schülerinnen und Schülern angegebenen Unterrichtseinheiten wurden mit 50 multipliziert, um Unterrichtsminuten zu erhalten.

der Geschlechterunterschied unter anderem daran liegen, dass es eine große Anzahl unterschiedlicher Formen mit unterschiedlichem Mathematikfokus¹⁵ gibt und Mädchen eher jene Formen mit geringerer Mathematikintensität besuchen. Dafür sprechen auch die Ergebnisse in Tabelle 5, wonach Mädchen an den HTL im Durchschnitt 20 Minuten pro Woche weniger Mathematikunterricht erhalten als Burschen, während sich innerhalb der weiteren BHS bzw. der AHS Geschlechterunterschiede in der Unterrichtszeit auf maximal fünf Minuten pro Woche belaufen.

3.3 Die Bedeutung von Geschlecht, Mathematiknote und -kompetenz für die Wahl einer maturaführenden Schule

Der Übertritt in die Sekundarstufe II erweist sich den bisherigen Ergebnissen zufolge als wesentlich für den weiteren Schulverlauf und determiniert bereits in einem großen Umfang Geschlechterunterschiede bei der Matura. In der Folge wird genauer auf diese wesentliche Schnittstelle Sekundarstufe I/II eingegangen und untersucht, wie Leistungsmerkmale (Kompetenzen und Noten) und Geschlecht bei der Wahl einer bestimmten Schulform zusammenspielen.

Tabelle 6: Mathematiknote nach Geschlecht, Kompetenz und Schulübertritt. Quelle: BIST-Ü-M8-2017-Datensatz.

	Notenwert (Mathematik)						
	1	2	3	4	5	≥6	gesamt
weiblich %	56	52	49	48	46	44	49
männlich %	44	48	51	52	54	56	51
Differenz	12	4	-2	-4	-8	-12	-2
Mathematikkompetenz (M)							
weiblich %	632	581	538	509	469	432	538
männlich %	643	597	554	524	484	446	545
Differenz	11	15	15	16	15	14	7
Übertritt maturaführende Schule (%)							
weiblich %	95	90	80	70	43	20	73
männlich %	90	83	69	58	31	13	61
Differenz	-5	-7	-11	-12	-13	-8	-12
<i>Anmerkungen: Differenz = männlich – weiblich.</i>							

15 Es gibt Lehrpläne für 29 unterschiedliche HTL-Formen. Relativ geringe Mathematikintensität weist z. B. die HTL für Art and Design auf, relativ hohe z. B. die HTL für Maschinenbau (BGBl. II Nr. 262/2015).

Tabelle 6 zeichnet folgendes Bild. Unter den Schülerinnen und Schülern mit den Notenwerten 1 und 2 sind Mädchen überrepräsentiert, bei der 3 liegt eine ungefähre Gleichverteilung vor und ab 4 sind Burschen überrepräsentiert. Generell gilt: Je schlechter die Note, desto stärker sind Burschen überrepräsentiert. Beispielsweise macht der Mädchenanteil in der BIST-Population 49 % aus, unter den Einserschülerinnen und -schülern¹⁶ liegt ihr Anteil bei 56 %, bei den Schülerinnen und Schülern mit den schlechtesten Noten (Notenwert ab 6¹⁷) beträgt ihr Anteil 44 %. Der zweite und dritte Abschnitt der Tabelle weist die Kompetenzpunkte und die Übertrittshäufigkeit in eine maturaführende Schule getrennt für Schülerinnen und Schüler mit unterschiedlichen Notenwerten aus. Die beiden Abschnitte machen deutlich, dass Mädchen bei gleicher Mathematiknote eine um 11 bis 15 Punkte niedrigere Mathematikkompetenz aufweisen als Burschen. Darüber hinaus planen Mädchen aller Notenstufen häufiger, in eine maturaführende Schule zu wechseln als Burschen (siehe auch Abbildung 2). Besonders groß ist dieser Unterschied bei Schülerinnen und Schülern mit Notenwerten von 3 bis 5. Für die Notenstufen 1–4 gilt, dass Mädchen etwa so häufig in eine maturaführende Schule übertreten wie Burschen der darüberliegenden Notenstufe (z. B. 69 % der Burschen mit Notenwert 3 treten in eine maturaführende Schule über und 70 % der Mädchen mit Notenwert 4). Diese Gegenüberstellung verdeutlicht den doppelten Nachteil der Burschen. Burschen erhalten bei besserer Mathematikkompetenz im Vergleich zu Mädchen schlechtere Mathematiknoten und treten selbst bei gleichen Noten seltener in eine maturaführende Schule über.

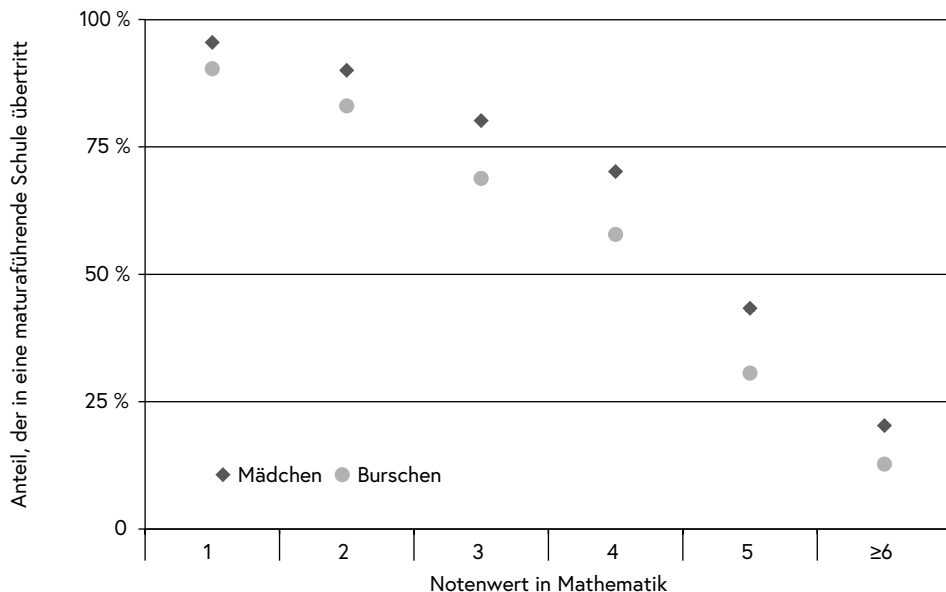


Abbildung 2: Geplanter Übertritt in eine maturaführende Schule auf der Sekundarstufe II getrennt nach Mathematiknote und nach Geschlecht (Quelle: BIST-Ü-M8-2017-Datensatz)

16 „Sehr gut“ in einer AHS bzw. in einer NMS mit Beurteilung nach vertiefter Allgemeinbildung bzw. in einer Ersten Leistungsgruppe der Hauptschule.

17 6 entspricht einem „Genügend“ in einer NMS mit Beurteilung nach grundlegender Allgemeinbildung oder in der Zweiten Hauptschul-Leistungsgruppe oder einem „Gut“ in der dritten Hauptschul-Leistungsgruppe.

In Tabelle 7 sind die Ergebnisse von logistischen Regressionsanalysen dargestellt. In diesen Analysen wird der Besuch einer maturaführenden Schule auf der Sekundarstufe II vorhergesagt. Odds Ratios (OR) geringer 1 bedeuten eine geringere Wahrscheinlichkeit für einen Übertritt in eine maturaführende Schule der Sekundarstufe II (im Vergleich zu keinem Übertritt in eine maturaführende Schule), OR über 1 eine höhere Wahrscheinlichkeit. Bei nicht signifikanten OR um den Wert 1 eignet sich der Prädiktor nicht zur Vorhersage des Besuchs einer maturaführenden Schule. Zunächst zeigt das Modell 1a, dass sich die Wahrscheinlichkeit des Besuchs einer maturaführenden Schule bei einer höheren (und damit schlechteren) Note verringert. Dieser Effekt bleibt im Übrigen in allen Modellen und somit auch unter Kontrolle von Mathematikkompetenz und Geschlecht bestehen (Modell 1b und Modell 2b). Im Modell 1b wird sichtbar, dass das Geschlecht auch bei Kontrolle der Mathematiknote die Wahl der Schulform der Sekundarstufe II vorhersagt. Mädchen haben eine höhere Wahrscheinlichkeit, eine maturaführende Schule zu wählen. In Modell 1c haben wir den Interaktionseffekt zwischen der Mathematiknote und dem Geschlecht in die Regression aufgenommen. Die OR besagen, dass sich für Burschen bei einer schlechteren Mathematiknote die Wahrscheinlichkeit des Besuchs einer maturaführenden Schule stärker verringert als bei Mädchen. Für die Entscheidung, nach der Sekundarstufe I in eine maturaführende Schule überzutreten, ist die Mathematiknote für Burschen somit von wesentlich höherer Relevanz als für Mädchen.

Tabelle 7: Vorhersage des Besuchs einer maturaführenden Schule (vs. keine maturaführende Schule) durch die Mathematiknote, die Mathematikkompetenz und das Geschlecht (inklusive Interaktionen mit Geschlecht), logistische Regression mit Angabe von Odds Ratios

Prädiktor	Modell					
	1a	1b	1c	2a	2b	2c
Mathematiknote	0,44	0,44	0,43	0,56	0,57	0,56
Mathematikkompetenz				1,01	1,01	1,01
Geschlecht (männlich)		0,56	0,46		0,47	0,41
Mathematiknote*Geschlecht			1,05			1,04
Mathematikkompetenz*Geschlecht						1,00
R ²	0,31	0,33	0,33	0,39	0,42	0,42

Anmerkungen: Eingetragen sind Odds Ratios. Alle Effekte außer Mathematikkompetenz*Geschlecht sind statistisch signifikant ($p < .001$).

Die Modelle 2a–2c zeigen, dass die Mathematiknote auch bei Kontrolle der Mathematikkompetenz weiterhin bedeutsam zum Übertritt in eine maturaführende Schule beiträgt. Bei Kontrolle von Mathematiknote und -kompetenz haben Mädchen weiterhin eine höhere Wahrscheinlichkeit des Besuchs einer maturaführenden Schule als Burschen. Das finale Modell 2c mit den Interaktionseffekten macht deutlich, dass sich für Burschen bei einer schlechteren Note die Wahrscheinlichkeit des Besuchs einer maturaführenden Schule weiterhin stärker verringert als bei Mädchen. Dieser Effekt ist jedoch gering (OR = 1.04).

Die Abbildung 2 zeigt in diesem Zusammenhang, dass dieser Effekt nicht linear verläuft und insbesondere bei den Noten 3 und 4 vorzufinden ist. Hier haben Mädchen mit dem Notenwert 4 sogar eine minimal höhere Übertrittsquote (70 %) als Burschen mit Notenwert 3 (69 %). Die Ergebnisse der Regressionsanalysen zeigen somit, dass sowohl die Mathematiknote als auch das Geschlecht ausschlaggebend für den Besuch einer maturaführenden Schule ist. Die Interaktionseffekte verdeutlichen wiederum, dass die Mathematiknote für Burschen eine höhere Relevanz hat als für Mädchen. Somit handelt es sich in maturaführenden Schulen bei den Burschen um eine positiver selektierte Gruppe mit Hinblick auf Mathematiknoten als bei Mädchen.

4 Diskussion und Ausblick

In diesem Kapitel wurde untersucht, ob sich tatsächlich nur die in Mathematik leistungsstärksten Mädchen und Burschen für den Besuch einer maturaführenden Schule entscheiden und wie hoch Geschlechterunterschiede in der Mathematikleistung bei jenen Schülerinnen und Schülern ausfallen, die eine maturaführende Schule besuchen. Dies sollte insgesamt klären, ob es sich bei den Burschen, die schlussendlich zur Matura antreten, um eine leistungsmäßig stärker selektierte Gruppe handelt als bei den Mädchen und inwiefern diese ungleiche Selektion zu Geschlechterunterschieden in den Noten der Mathematiklausur der SRDP beitragen. Diese Fragen wurden mittels eines Pseudopanel-Designs untersucht, bei dem die Kohorte der 2002/2003 geborenen Schülerinnen und Schüler über unterschiedliche standardisierte Testungen beobachtet wurde – von der BIST-Ü M4 im Jahr 2013 über die BIST-Ü M8 im Jahr 2017 und PISA im Jahr 2018 bis hin zur SRDP im Jahr 2021 (AHS) bzw. im Jahr 2022 für die BHS-Absolventinnen und -Absolventen.

Die Ergebnisse der BIST-Ü-M8-Daten zeigen, dass Mädchen am Ende der Sekundarstufe I selbst unter Kontrolle der Mathematiknoten und der Mathematikkompetenz wesentlich häufiger eine maturaführende Schule wählen als Burschen. Mädchen, die in maturaführende Schulen eintreten, weisen von Anfang an eine niedrigere durchschnittliche Mathematikkompetenz auf als Burschen. Dies könnte unter anderem daran liegen, dass generell weniger Burschen (54 % eines Altersjahrgangs) als Mädchen (67 %) in eine maturaführende Schule übertreten. Zudem wird bei Burschen dieser Übertritt viel stärker durch ihre Mathematiknoten (und Mathematikkompetenz) mitbestimmt als bei Mädchen. So verringert sich bei schlechteren Noten die Wahrscheinlichkeit für den Übertritt in eine maturaführende Schule bei Mädchen weniger stark als bei Burschen. Burschen entscheiden sich in der Regel dann für eine maturaführende Schule, wenn sie in Mathematik vergleichsweise leistungsstark sind, Mädchen auch dann, wenn sie eher durchschnittliche Mathematikleistungen aufweisen. Eine Studie im österreichischen Kontext zu geschlechtsbezogenen Notenunterschieden zeigte bereits, dass Burschen in Mathematik bei gleichen Kompetenzständen etwas schlechtere Noten als Mädchen erlangen (Schmid, 2018). Dieses Ergebnis geht zudem damit einher, dass Mädchen häufiger in eine maturaführende Schule übertreten und dort häufiger die Matura bestehen. Somit scheinen bezogen auf die Selektivität im Übergang von der Sekundarstufe I zur Sekundarstufe II multiple Benachteiligungsmechanismen für Burschen vorzuliegen.

Ein bedeutender Unterschied in den Mathematikkompetenzen zwischen Burschen und Mädchen ist nicht nur beim Übertritt in maturaführende Schulen ersichtlich, sondern wird dort weiter zementiert und verschärft, wie die PISA-Daten zeigen. Erstens wählen gerade Mädchen mit schwächerer Mathematikkompetenz häufig die weiteren BHS (darunter fallen z. B. höhere Lehranstalten für wirtschaftliche Berufe oder Bildungsanstalten für Elementarpädagogik). In diesen Schulformen erhalten sie wiederum wesentlich weniger Mathematikunterricht als die leistungsstarken Burschen, die sich häufig an einer HTL befinden. Diese Mädchen und Burschen mit völlig unterschiedlichen Voraussetzungen erhalten schließlich bei der Mathematiklausur der SRDP alle denselben A-Teil (der etwa 50 % zur Gesamtnote beiträgt). Folglich ist auch der Notendurchschnitt bei der SRDP-Mathematiklausur in den weiteren BHS wesentlich schlechter (mit 3,22 im Jahr 2022) als in den HTL (2,63). Hier muss zudem berücksichtigt werden, dass in den weiteren BHS die Möglichkeit besteht, die standardisierte Mathematiklausur abzuwählen und zu einer nicht standardisierten mündlichen Prüfung anzutreten. Es liegt die Vermutung nahe, dass eher Kandidatinnen (und auch Kandidaten) mit relativ geringer Mathematikkompetenz davon Gebrauch machen und der Notendurchschnitt in den weiteren BHS noch schlechter ausfallen würde, wenn auch hier das Antreten zur Mathematiklausur verpflichtend wäre. Dieser Vermutung sollte in zukünftigen Studien genauer nachgegangen werden.

Mädchen schneiden zwar bei der Mathematiklausur schlechter ab als Burschen, bestehen aber die Matura zum Haupttermin (in der Regel das erste Antreten) häufiger als Burschen. Dies könnte daran liegen, dass seit 2020 Noten des Jahreszeugnisses in das Maturaergebnis eingerechnet werden sowie dass negative Ergebnisse durch eine mündliche Kompensationsprüfung ausgeglichen werden können. Darüber hinaus fließen in das Matura-Gesamtergebnis auch alle anderen Maturagegenstände ein und z. B. in Deutsch schneiden Mädchen deutlich besser ab als Burschen. Somit wird der Nachteil der Mädchen, der durch weniger Mathematikunterricht entsteht, durch Noteneinrechnung und Kompensationsprüfung wieder ausgeglichen. Dieser Schlussfolgerung könnte mit einer weiterführenden Studie zu geschlechtsspezifischen Unterschieden in den Noten der Maturaklausuren und den Jahreszeugnisnoten nachgegangen werden.

Mädchen bestehen damit trotz durchschnittlich schwächerer Mathematikkompetenz als Burschen die Matura insgesamt sogar etwas häufiger als Burschen. Dies mag zunächst den Anschein eines positiven Effekts für die Mädchen erwecken, ist in ihrem weiteren Lebensverlauf aber zum Teil wieder mit negativen Konsequenzen verbunden. Etwa ist eine Reihe an Studienfächern mit Aufnahme- bzw. Eignungsverfahren oder schweren Prüfungen insbesondere zu Studienbeginn (sogenannten *Knock-out-Prüfungen*) verbunden.¹⁸ Oft spielen dabei Mathematikkompetenzen eine wichtige Rolle (u. a. bei den Aufnahmeverfahren für Wirtschaft, Psychologie, Informatik¹⁹, aber auch für Medizin²⁰). Insbesondere beim Aufnahmetest für das Medizinstudium sorgte das schlechtere Abschneiden der

18 <https://www.aufnahmepruefung.at/uni/>

19 <https://studieren.univie.ac.at/aufnahmeverfahren/testdetails/>

20 <https://www.medcampus.at/medat/>

Frauen wiederkehrend für Aufsehen und wurde auch wissenschaftlich untersucht (Spiel, Schober & Litzenberger, 2008). Beim Aufnahmetest für das Medizinstudium spielen naturwissenschaftliche Kompetenzen eine entscheidende Rolle. Diese werden bei der Matura nicht standardisiert geprüft. Die Ergebnisse von PISA (Brandmair & Riegler, 2023) zeigen allerdings auch für diese Domäne (wie schon für Mathematik) einen Vorteil der Burschen, der in Österreich zu den größten im OECD- und EU-Raum zählt.

Eine Anhebung der Kompetenzen von Mädchen in Naturwissenschaft und Mathematik wäre wichtig, um mehr Frauen und Mädchen in die MINT-Fächer (Mathematik, Informatik, Naturwissenschaft, Technik) zu bewegen. Kompetenzen spielen hierfür eine wesentliche Rolle (Hemmings et al., 2011), aber nicht die einzige. Darüber hinaus könnte der kulturelle Kontext (z. B. Geschlechterstereotype in der Gesellschaft; Nosek et al., 2009) oder motivationale Merkmale der Schülerinnen und Schüler (z. B. akademisches Selbstkonzept; Salchegger, 2015) zur Wahl mathematikintensiver Bildungs- bzw. Berufswege beitragen. Motivationale Merkmale und Kompetenzkonstellationen erklären in einem wesentlichen Umfang, warum sich Mädchen und Frauen selbst bei gleichen Kompetenzen seltener für MINT-bezogene Bildungswege und Berufe entscheiden als Burschen und Männer (Eccles, 1994; Wang et al., 2013; Parker et al., 2014). Bemerkenswert ist, dass insbesondere für die MINT-Fächer eine Leaky Pipeline (d. h. ein sinkender Frauenanteil mit steigender Qualifizierungsstufe) festgestellt wird (z. B. Köszegi, Günther, Haas, Keinert & Zedlacher, 2011; Speer, 2023).

Das BMBWF (2023a) geht davon aus, dass die SRDP mehr Fairness, gleiche Bedingungen für alle Maturantinnen und Maturanten und eine leichtere Vergleichbarkeit der Bildungsabschlüsse für weiterführende Bildungseinrichtungen bringt. In Bezug auf die Fairness wirft dieses Kapitel eine Reihe an Fragen auf, die sich zum Teil auf die SRDP direkt beziehen, zum Teil auch auf vorgelagerte Ungerechtigkeit im Schulsystem, die für die SRDP relevant sind. So werfen unsere Ergebnisse die Frage auf, ob Fairness gegeben ist, wenn ...

- Schülerinnen und Schüler der weiteren BHS (das sind zum Großteil Mädchen) weniger Mathematikunterricht erhalten (um ca. 60 Minuten pro Woche weniger als HTL-Schülerinnen und -Schüler), aber denselben A-Teil bei der Mathematik Klausur der SRDP erhalten.
- ein wesentlich geringerer Anteil an Burschen als Mädchen zur Matura antritt, aber die Bestehensquote bei den Mädchen trotzdem höher ist.
- an manchen Schulformen (mit wenig Mathematikunterricht und hohem Mädchenanteil) die standardisierte Mathematik Klausur abgewählt werden kann und auf eine nicht standardisierte mündliche Prüfung ausgewichen werden kann.

Der Rechnungshof hatte empfohlen, die Gründe für Geschlechterunterschiede in der Mathematik Klausur genauer zu untersuchen (Rechnungshof Österreich, 2020). Wir konnten zeigen, dass die Unterschiede teilweise mit Selektionsmechanismen begründet werden können. So beziehen Burschen die Mathematikleistung am Ende der Sekundarstufe I stärker in ihre Entscheidung ein, ob sie in eine maturaführende Schule übertreten. Sie erhalten in

weiterer Folge durch ihre Schulwahl mehr Mathematikunterricht auf der Sekundarstufe II und sind im Endeffekt eine stärker ausgesiebte Gruppe, weil wesentlich weniger Burschen als Mädchen zur Matura antreten.

Da sich gegensätzliche Unterschiede in Deutsch finden (Mädchen schneiden bei der Deutschklausur der SRDP besser ab; z. B. Bönisch et al., 2022), liegt die Vermutung nahe, dass auch im sprachlichen Bereich Selektionsmechanismen eine Rolle spielen. Geschlechterfairness bei der SRDP sollte daher in zukünftigen Untersuchungen nicht nur für Mathematik beleuchtet werden, sondern auch für Deutsch und Englisch. Mit diesen Studien könnte untersucht werden, ob ein ähnlicher Zusammenhang besteht, also die Entscheidung für den Übertritt in eine maturaführende Schule bei Burschen stärker nach Leistung erfolgt als bei Mädchen oder ob es in diesen Fächern eventuell umgekehrt ist und bei Mädchen Noten in sprachlichen Fächern bedeutender für den Übertritt in eine maturaführende Schule sind als bei Burschen.

Das vorliegende Kapitel hat gezeigt, dass es in Bezug auf das Abschneiden bei der Mathematik-Zentralmatura bedeutsam ist, in welcher höheren Schulform die Sekundarstufe II absolviert wurde. So erreichen Maturakandidatinnen und -kandidaten, die in einer HTL antreten, einen wesentlich besseren Notenschnitt bei der Mathematik Klausur (2,63) als jene, die in einer weiteren BHS antreten (3,22). Zukünftige Studien sollten genauer untersuchen, welche Faktoren im österreichischen Kontext zur Wahl einer bestimmten maturaführenden Schulform beitragen. Dabei wäre es wichtig, nicht nur leistungsbezogene Aspekte in den Blick zu nehmen, sondern auch motivationale, wie Selbstkonzept, Einstellungen oder Interesse, da diese über die Kompetenzen hinaus wesentlich zu Bildungs- und Berufswahlen beitragen (z. B. Eccles, 1994; Wang, Eccles & Kenny, 2013; Parker et al., 2014). Da weitere Forschung zeigt, dass die Unterstufen-Schulkarriere eine wesentliche Rolle für die Performanz in der Oberstufe spielt (z. B. Steiner, 2024), sollte in zukünftigen Studien auch genauer untersucht werden, welchen Einfluss die Unterstufen-Schulkarriere auf die Oberstufen-Schulkarriere und letztendlich auf das Abschneiden bei der Matura hat.

Literatur

- Aue, V., Frebort, M., Hohenwarter, M., Liebscher, M., Sattlberger, E., Schirmer, I., Siller, H.-S. et al. (2021). *Die standardisierte schriftliche Reifeprüfung in Mathematik (AHS)*. Verfügbar unter: <https://www.matura.gv.at/index.php?eID=dumpFile&t=f&f=4826&token=4574fed24b889f914a68a7411172dbce06459c69>
- Bacher, J. (2005). Bildungsungleichheit und Bildungsbenachteiligung im weiterführenden Schulsystem Österreichs: eine Sekundäranalyse der PISA 2000-Erhebung. *SWS-Rundschau*, 45 (1), 37–62.

- Bacher, J., Beham, M. & Lachmayr, N. (Hrsg.). (2008). *Geschlechterunterschiede in der Bildungswahl*. Wiesbaden: Verlag für Sozialwissenschaften.
- Benischek, I., Oberwimmer, K. & Höller, I. (2022). Zu Einflussfaktoren auf Bildungserwartungen von Schülerinnen und Schülern nach der Sekundarstufe I. In A. C. George, S. Götz, M. Illetschko & E. Süss-Stepancik (Hrsg.), *Empirische Befunde zu Kompetenzen im Mathematikunterricht der Sekundarstufe I und Folgerungen für die Praxis* (S. 159–186). Münster: Waxmann. doi:10.31244/9783830995586
- BIFIE (2013). *Standardisierte kompetenzorientierte Reifeprüfung | Reife- und Diplomprüfung. Grundlagen – Entwicklung – Implementierung*. Verfügbar unter: <https://www.matura.gv.at/index.php?eID=dumpFile&t=f&f=1786&token=6d36c3b29ea588d46f540972351fed64cc783ae8>
- BIFIE (2017). *Almanach zur Standardüberprüfung 2013 Mathematik, 4. Schulstufe. Schüler-ebene, M413I. Forschungsdatenbibliothek (FDB). Nicht-imputierter Datensatz, v2.0-E*. Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/ad560a65744addeb705ba028fb3638c99dc315c7/M413I_EXTERN_2_0_Almanach.pdf
- BIFIE (2018). *Almanach zur Standardüberprüfung 2017 Mathematik, 8. Schulstufe. Schüler-ebene, M817I. Forschungsdatenbibliothek (FDB). Nicht-imputierter Datensatz, v1.0-E*. Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/ee2f6128a40025fb4bcb2a48886c1d14e807f62f/M817I_EXTERN_1_0_Almanach.pdf
- BM BW F (2018). *Aufnahme in berufsbildende mittlere und höhere Schulen*. Verfügbar unter: https://www.bmbwf.gv.at/Themen/schule/schulsystem/sa/bmhs/bmhs_aufnahme.html
- BM BW F (2019). *Standardisierte kompetenzorientierte schriftliche Reife- und Diplomprüfung in Angewandter Mathematik an BHS*. Verfügbar unter: <https://www.matura.gv.at/index.php?eID=dumpFile&t=f&f=2629&token=034b82cc429e9882bc6562c37bc8296aa9a45cd4>
- BM BW F (2020a). *Information: Neuerungen bei den abschließenden Prüfungen ab Frühjahr 2021 (Haupttermin 2021)*. Verfügbar unter: <https://www.bmbwf.gv.at/dam/jcr:0fc239b4-e50e-4bba-a28c-ed23e3054f6a/Neuerungen%20abschlie%C3%9Fende%20Pr%C3%Bcfungen.pdf>
- BM BW F (2020b). *Die Mittelschule. Änderungen ab dem Schuljahr 2020/21 im Überblick*. Verfügbar unter: https://www.bmbwf.gv.at/dam/jcr:7b6de1bc-36c1-4b54-88f0-7683120238d0/mittelschule_2020.pdf
- BM BW F (2021). *Zentralmatura 2021*. Verfügbar unter: <https://www.bmbwf.gv.at/Themen/schule/schulpraxis/zentralmatura/zentralmatura2021.html>
- BM BW F (2022a). *Information zu abschließenden Prüfungen für den Haupttermin zum Schuljahr 2022/23*. Verfügbar unter: <https://rundschreiben.bmbwf.gv.at/rundschreiben/?id=1048>
- BM BW F (2022b). *Cluster-Einteilung Angewandte Mathematik (BHS)*. Verfügbar unter: <https://www.matura.gv.at/index.php?eID=dumpFile&t=f&f=4481&token=b48be6820546fcfb61d5ce6402dee02357c9e0c2>
- BM BW F (2023a). *Zentralmatura*. Verfügbar unter: https://www.oesterreich.gv.at/themen/bildung_und_neue_medien/schule/2/1.html

- BMBWF (2023b). *Standardisierte kompetenzorientierte Reifeprüfung an AHS*. Verfügbar unter: https://www.bmbwf.gv.at/Themen/schule/schulpraxis/zentralmatura/srdp_ahs.html
- BMBWF (2023c). *Standardisierte kompetenzorientierte Reifeprüfung an AHS*. Verfügbar unter: https://www.bmbwf.gv.at/Themen/schule/schulpraxis/zentralmatura/srdp_ahs.html
- BMBWF (2023d). *Standardisierte kompetenzorientierte Reifeprüfung an BHS*. Verfügbar unter: https://www.bmbwf.gv.at/Themen/schule/schulpraxis/zentralmatura/srdp_bhs.html
- BMBWF (2023e). *Angewandte Mathematik*. Verfügbar unter: <https://www.matura.gv.at/srdp/angewandte-mathematik>
- BMBWF (2023f). *Korrektur- und Beurteilungsanleitung zur SRP Mathematik (AHS)*. Verfügbar unter: <https://www.matura.gv.at/downloads/download/korrektur-und-beurteilungsanleitung-zur-srp-mathematik-ahs>
- BMBWF (2023g). *Teilstandardisierte kompetenzorientierte Reife- und Diplomprüfung an BHS (Zentralmatura)*. Verfügbar unter: https://www.oesterreich.gv.at/themen/bildung_und_neue_medien/schule/2/1/Seite.110038.html
- BMBWF (2023h). *Beilage A zu Rundschreiben Nr. 13/2023: Vorbereitung und Durchführung, Korrektur und Beurteilung, Ergebniserhebung und wissenschaftliche Auswertung der standardisierten Reifeprüfung bzw. Reife- und Diplomprüfung und der Berufsreifeprüfung für das Schuljahr 2022/23*. Verfügbar unter: <https://www.matura.gv.at/index.php?eID=dumpFile&t=f&f=5810&token=def13831b8cafa7418a6b4557b0736bd0c6ab010>
- Bönisch, M., Ernst, D., Hirt, E. M., Basas, S. & Dentler, E.-K. (2022). *Standardisierte Reifeprüfung und Reife- und Diplomprüfung Abschlussjahrgang 2020/21 – Haupttermin – Tabellenband*. Wien: Statistik Austria. Verfügbar unter: https://www.statistik.at/fileadmin/pages/322/standardisierte_reife-_und_diplompruefung_haupttermin_2021__3_.pdf
- Bönisch, M., Ernst, D., Hirt, E. M., Basas, S. & Zheng, J. (2021). *Standardisierte Reifeprüfung und Reife- und Diplomprüfung Abschlussjahrgang 2019/20 – Haupt- und Nebentermine*. Wien: Statistik Austria. Verfügbar unter: https://www.statistik.at/fileadmin/pages/322/standardisierte_reife-_und_diplompruefung_abschlussjahrgang_201920_-_tabel__17_.pdf
- Bönisch, M., Maldet, A., Mayerweck, E. & Zheng, J. (2019a). *Standardisierte Reife- und Diplomprüfung Abschlussjahrgang 2016/17 – Haupt- und Nebentermine – Tabellenband*. Wien: Statistik Austria. Verfügbar unter: https://www.statistik.at/fileadmin/pages/322/standardisierte_reife-_und_diplompruefung_abschlussjahrgang_201617_-_tabel__7_.pdf
- Bönisch, M., Maldet, A., Mayerweck, E. & Zheng, J. (2019b). *Standardisierte Reife- und Diplomprüfung Abschlussjahrgang 2017/18 – Sommertermin – Tabellenband*. Wien: Statistik Austria. Verfügbar unter: https://www.statistik.at/fileadmin/pages/322/standardisierte_reife-_und_diplompruefung__sommertermin_2018.pdf
- Bönisch, M., Maldet, A. & Zheng, J. (2020). *Standardisierte Reife- und Diplomprüfung Abschlussjahrgang 2018/19 – Sommertermin – Tabellenband*. Wien: Statistik Austria. Verfügbar unter: https://www.statistik.at/fileadmin/pages/322/standardisierte_reife-_und_diplompruefung_abschlussjahrgang_201819_-_tabel__14_.pdf

- Brandmair, K. & Riegler, A. (2023). Geschlechterdifferenzen in Naturwissenschaft. In B. Toferer, B. Lang & S. Salchegger (Hrsg.), *PISA 2022 – Kompetenzen in Mathematik, Lesen und Naturwissenschaft am Ende der Pflichtschulzeit im internationalen Vergleich* (S. 67–70). Salzburg: IQS. doi:10.17888/pisa2022-eb
- Brandmair, K. & Schmoller, K. (2023). Geschlechterdifferenzen in Mathematik. In B. Toferer, B. Lang & S. Salchegger (Hrsg.), *PISA 2022 – Kompetenzen in Mathematik, Lesen und Naturwissenschaft am Ende der Pflichtschulzeit im internationalen Vergleich* (S. 35–38). Salzburg: IQS. doi:10.17888/pisa2022-eb
- Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F. et al. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research*, 86, 803–848. doi:10.3102/0034654316672069
- Bruneforth, M., Oberwimmer, K. & Robitzsch, A. (2016). Reporting und Analysen. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (S. 333–362). Wien: facultas. Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/3981a91ecff730ff465386108eb6f9411827dd49/Large_Scale_Assessment_mit_R_gesamt.pdf
- Duckworth, A. L. & Seligman, M. E. P. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology*, 98, 198–208. doi:10.1037/0022-0663.98.1.198
- Eccles, J. S. (1994). Understanding women's educational and occupational choices. *Psychology of Women Quarterly*, 18, 585–609.
- Else-Quest, N. M., Hyde, J. S. & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, 136, 103–127. doi:10.1037/a0018053
- Gaspard, H., Dicke, A.-L., Flunger, B. & Brisson, B. M. (2015). More value through greater differentiation: Gender differences in value beliefs about math. *Journal of Educational Psychology*, 107, 663–677. doi:10.1037/edu0000003
- Glaeser, A., Rölz, M., Suchaň, B., Schmich, J., Wiesinger, L. & Höller, I. (2019). Kompetenzen der Schüler/innen in Mathematik. In B. Suchaň, I. Höller & C. Wallner-Paschon (Hrsg.), *PISA 2018. Grundkompetenzen am Ende der Pflichtschulzeit im internationalen Vergleich* (S. 49–56). Graz: Leykam. doi:10.17888/pisa2018-eb
- Guo, J., Parker, P. D., Marsh, H. W. & Morin, A. J. S. (2015). Achievement, motivation, and educational choices: A longitudinal study of expectancy and value using a multiplicative perspective. *Developmental Psychology*, 51, 1163–1176. doi:10.1037/a0039440
- Helbig, M. (2012). Die Umkehrung – Geschlechterungleichheiten beim Erwerb des Abiturs im Wandel. In R. Becker & H. Solga (Hrsg.), *Soziologische Bildungsforschung* (S. 374–392). Wiesbaden: Springer.
- Hemmings, B., Grootenboer, P. & Kay, R. (2011). Predicting mathematics achievement: The influence of prior achievement and attitudes. *International Journal of Science and Mathematics Education*, 9 (3), 691–705. doi:10.1007/s10763-010-9224-5
- Hirt, E. M., Basas, S. & Meisterhofer, M. (2023). *Standardisierte Reifeprüfung und Reife- und Diplomprüfung Abschlussjahrgang 2021/22 – Haupttermin – Tabellenband*. Wien: Statistik Austria.

- Höller, I. & Toferer, B. (2019). Die getesteten Kompetenzbereiche. In B. Suchan (Hrsg.), *PISA 2018. Technischer Bericht*. Salzburg: BIFIE. doi:10.17888/pisa2018-tb.2
- HTL Krems (2020). *Voraussetzungen für eine Aufnahme ohne Aufnahmeprüfung*. Verfügbar unter: https://www.htlkrems.ac.at/wp-content/uploads/2020/06/Voraussetzungen-f%C3%BCr-die-Aufnahme_Neu.pdf
- Kampa, N., Krämer, S. & Hannover, B. (2020). Secondary school leaving examinations: The impact of expectancies, values, and comparisons on male and female students' science-oriented choices. *Frontiers in Education*, 5, 1–16. doi:10.3389/educ.2020.545608
- Köszegi, S., Günther, E., Haas, M., Keinert, C. & Zedlacher, E. (2011). *Leaky Pipeline: Eine Studie zur horizontalen und vertikalen Segregation an der TU Wien*. Verfügbar unter: <https://www.tuwien.at/mwbw/im/ao/projekte/abgeschlossene-forschungsprojekte/leaky-pipeline>
- Lang, B. & Glaeser, A. (2023). Geschlechterunterschiede in der Mathematik-Unterrichtszeit. In B. Toferer, B. Lang & S. Salchegger (Hrsg.), *PISA 2022 – Kompetenzen in Mathematik, Lesen und Naturwissenschaft am Ende der Pflichtschulzeit im internationalen Vergleich*. Salzburg: IQS. doi:10.17888/pisa2022-eb
- Lazarides, R. & Lauermaun, F. (2019). Gendered paths into STEM-related and language-related careers: Girls' and boys' motivational beliefs and career plans in math and language arts. *Frontiers in Psychology*, 10 (1243), 1–17. doi:10.3389/fpsyg.2019.01243
- Lindberg, S. M., Hyde, J. S., Petersen, J. L. & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, 136, 1123–1135. doi:10.1037/a0021276
- Maaz, K., Baeriswyl, F. & Trautwein, U. (2013). II. Studie: „Herkunft zensiert?“ Leistungsdiagnostik und soziale Ungleichheiten in der Schule. In D. Deißner (Hrsg.), *Chancen bilden. Wege zu einer gerechteren Bildung – ein internationaler Erfahrungsaustausch* (S. 187–326). Wiesbaden: Springer VS.
- Moser, U., Oostlander, J. & Tomasik, M. J. (2017). Soziale Ungleichheiten im Leistungszuwachs und bei Bildungsgängen. In M. P. Neuenschwander & C. Nägele (Hrsg.), *Bildungsverläufe von der Einschulung bis in den ersten Arbeitsmarkt. Theoretische Ansätze, empirische Befunde und Beispiele* (S. 59–77). Wiesbaden: Springer VS. doi:10.1007/978-3-658-16981-7_4
- Neubacher, M. & Wimmer, C. (2021). Indikatoren D: Output – Ergebnisse des Schulsystems. In BMBWF (Hrsg.), *Nationaler Bildungsbericht Österreich 2021, Teil 2 – Bildungsindikatoren* (S. 286–328). Wien: BMBWF. doi:10.17888/nbb2021
- Neuwirth, E. (2016). *Geschlechterunterschiede bei der Zentralmatura*. Verfügbar unter: <http://www.neuwirth.priv.at/forblogs/Zentralmatura/>
- Nimmervoll, L. (2016, 1. Februar). Mädchen bei Zentralmatura in Englisch unerwartet schlecht. *Der Standard*. Verfügbar unter: <https://www.derstandard.at/story/2000030115353/maedchen-bei-zentralmatura-in-englisch-unerwartet-schlecht>
- Nosek, B. A., Smyth, F., Sriram, N., Lindner, N., Devos, T., Ayala, A. et al. (2009). National differences in genderscience stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, 106, 10593–10597. doi:10.1073/pnas.0809921106

- Oberwimmer, K., Bruneforth, M., Siegle, T., Vogtenhuber, S., Lassnigg, L., Schmich, J. et al. (2016). Indikatoren D: Output – Ergebnisse des Schulsystems. In M. Bruneforth, L. Lassnigg, S. Vogtenhuber, C. Schreiner & S. Breit (Hrsg.), *Nationaler Bildungsbericht Österreich 2015, Band 1: Das Schulsystem im Spiegel von Daten und Indikatoren* (S. 129–194). Graz: Leykam. doi:10.17888/nbb2015-1-D
- OECD (2009). *PISA Data Analysis Manual: SPSS, Second Edition*. Paris: OECD Publishing. doi:10.1787/9789264056275-en
- OECD (2015). *The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence*. Paris: OECD Publishing. doi:10.1787/9789264229945-en
- OECD (2019). *PISA 2018 Assessment and Analytical Framework*. Paris: OECD Publishing. doi:10.1787/b25efab8-en
- OECD (2020a). *PISA 2018 – Technical Report*. Verfügbar unter: <https://www.oecd.org/pisa/data/pisa2018technicalreport/>
- OECD (2020b). *PISA 2018 Database*. Verfügbar unter: <https://www.oecd.org/pisa/data/2018database/>
- Palmer, T.-A., Burke, P. F. & Aubusson, P. (2017). Why school students choose and reject science: A study of the factors that students consider when selecting subjects. *International Journal of Science Education*, 39 (6), 645–662. doi:10.1080/09500693.2017.1299949
- Parker, P., Nagy, G., Trautwein, U. & Lüdtke, O. (2014). Predicting career aspirations and university majors from academic ability and self-concept: A longitudinal application of the internal-external frame of reference model. In I. Schoon & J. S. Eccles (Eds.), *Gender differences in aspirations and attainment: A life course perspective* (pp. 224–246). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139128933.015
- Perez-Felkner, L., Nix, S. & Thomas, K. (2017). Gendered pathways: How mathematics ability beliefs shape secondary and postsecondary course and degree field choices. *Frontiers in Psychology*, 8 (386). doi:10.3389/fpsyg.2017.00386
- Rechnungshof Österreich (2020). *Zentralmatura. Bericht des Rechnungshofes*. Wien: Herausgeber. Verfügbar unter: https://www.rechnungshof.gv.at/rh/home/home/004.743_Zentralmatura.pdf
- Reilly, D. (2012). Gender, culture, and sex-typed cognitive abilities. *PloS ONE*, 7 (7). doi:10.1371/journal.pone.0039904
- Reilly, D., Neumann, D. L. & Andrews, G. (2015). Sex differences in mathematics and science achievement: A meta-analysis of National Assessment of Educational Progress assessments. *Journal of Educational Psychology*, 107, 645–662. doi:10.1037/edu0000012
- Salchegger, S. (2015). Mathematik ≠ weiblich? Leistung, Selbstkonzept und Studienabschlüsse im Geschlechtervergleich. In B. Suchaň, C. Wallner-Paschon & C. Schreiner (Hrsg.), *PIRLS & TIMSS 2011 – Die Kompetenzen in Lesen, Mathematik und Naturwissenschaft am Ende der Volksschule* (S. 39–54). Graz: Leykam. Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/71b6509c59dd3bbddbcb339a4f76c3b524d5e7d/PIRLS_TIMSS_Expertenbericht_2011_web.pdf
- Salchegger, S., Glaeser, A., Widauer, K. & Bitesnich, H. (2017). Warum besuchen Mädchen mit Spitzenleistungen in Mathematik so selten eine höhere technische Lehranstalt? Ursachen und Folgen von Geschlechterunterschieden bei der Schulwahl. In P. Schlögl,

- D. Moser, M. Stock, K. Schmid & F. Gramlinger (Hrsg.), *Berufsbildung, eine Renaissance? Motor für Innovation, Beschäftigung, Teilhabe, Aufstieg, Wohlstand* (S. 172–183). Bielefeld: wbv. doi:10.3278/6004552w172
- Salchegger, S. & Suchań, B. (2018). Was bedeutet es für den Geschlechterunterschied in der Mathematikkompetenz bei PISA, wenn dem Schulsystem leistungsschwache Jungen verloren gehen? *Zeitschrift für Bildungsforschung*, 8, 81–99. doi:10.1007/s35834-017-0190-7
- Schmid, C. (2018). Asymmetrien zwischen Noten und Testleistungen am Ende der 8. Schulstufe? Eine Analyse auf Grundlage der österreichischen Bildungsstandardüberprüfung Mathematik 2012. *Zeitschrift für Bildungsforschung*, 8, 249–268. doi:10.1007/s35834-019-00229-3
- Schreiner, C. & Breit, S. (Hrsg.). (2014). *Standardüberprüfung 2013. Mathematik, 4. Schulstufe. Bundesergebnisbericht*. Salzburg: Bundesinstitut für Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens. Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/7b5c3ea0cdcb6908fa80c10e60adbb6eb2c9ffa3/BiSt-UE_M4_2013_Bundesergebnisbericht.pdf
- Schreiner, C., Breit, S. & Haider, G. (2008). Zur Validität der Mathematiknoten. Ein Vergleich von Lehrerbeurteilung und Leistungsmessung bei PISA. In F. Hofmann, C. Schreiner & J. Thonhauser (Hrsg.), *Qualitative und quantitative Aspekte. Zu ihrer Komplementarität in der erziehungswissenschaftlichen Forschung* (S. 211–223). Münster: Waxmann.
- Schreiner, C., Breit, S., Pointinger, M., Pacher, K., Neubacher, M. & Wiesner, C. (Hrsg.). (2018). *Standardüberprüfung 2017. Mathematik, 8. Schulstufe. Bundesergebnisbericht*. Salzburg: Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens. Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/t4a28609fd6414dbc257274688ffa37e44e4a3cf7/BiSt-UE_M8_2017_Bundesergebnisbericht.pdf
- Schreiner, C. & Wiesner, C. (2019). Die Überprüfung der Bildungsstandards in Österreich: Der erste Zyklus als Meilenstein für die Schul- und Unterrichtsentwicklung – eine gelungene Innovation im österreichischen Schulsystem. In A. C. George, C. Schreiner, C. Wiesner, M. Pointinger & K. Pacher (Hrsg.), *Fünf Jahre flächendeckende Bildungsstandardüberprüfungen in Österreich: Vertiefende Analysen zum Zyklus 2012 bis 2016* (S. 13–53). Münster: Waxmann. Verfügbar unter: <https://www.waxmann.com/buch3925>
- Speer, J. D. (2023). Bye bye Ms. American Sci: Women and the leaky STEM pipeline. *Economics of Education Review*, 93.
- Spiel, C., Schober, B. & Litzenberger, M. (2008). *Evaluation der Eignungstests für das Medizinstudium in Österreich. Zusammenfassung und Empfehlungen. Evaluationsprojekt im Auftrag des Ministeriums für Wissenschaft und Forschung*. Universität Wien – Fakultät für Psychologie.
- Statistik Austria (2023). *Bildung in Zahlen 2021/22. Schlüsselindikatoren und Analysen*. Wien: Herausgeber. Verfügbar unter: <https://www.statistik.at/services/tools/services/publikationen/detail/1562>
- Steiner, M. (2024). Leisten Neue Mittelschulen einen Beitrag zu mehr Chancengerechtigkeit? Diskussion von Evidenzen auf der Grundlage sekundärstatistischer Analysen. In L. Jesacher-Rößler & D. Kemethofer (Hrsg.), *10 Jahre Regelschule – die (Neue) Mittelschule* (S. 346–368). Münster: Waxmann.

- Suchań, B. & Höller, I. (2019). PISA 2018 – eine Einführung in die aktuelle Studie. In B. Suchań, I. Höller & C. Wallner-Paschon (Hrsg.), *PISA 2018. Grundkompetenzen am Ende der Pflichtschulzeit im internationalen Vergleich* (S. 9–37). Graz: Leykam. doi:10.17888/pisa2018-eb
- Suchań, B., Schmich, J., Toferer, B. & Höller, I. (2016). Kompetenzen der Schüler/innen in Mathematik. In B. Suchań & S. Breit (Hrsg.), *PISA 2015. Grundkompetenzen am Ende der Pflichtschulzeit im internationalen Vergleich* (S. 65–75). Graz: Leykam.
- Van den Ham, A. K., Ehmke, T., Hahn, I., Wagner, H. & Schöps, K. (2016). Mathematische und naturwissenschaftliche Kompetenz in PISA, im IQB-Ländervergleich und in der National Educational Panel Study (NEPS) – Vergleich der Rahmenkonzepte und der dimensional Struktur der Testinstrumente. In Bundesministerium für Bildung und Forschung (Hrsg.), *Forschungsvorhaben in Anknüpfung an Large-Scale-Assessments* (S. 140–160). Berlin: Bundesministerium für Bildung und Forschung. doi:10.25656/01:12677
- Voyer, D. & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, 140, 1174–1204. doi:10.1037/a0036620
- Wang, M.-T., Eccles, J. S. & Kenny, S. (2013). Not lack of ability but more choice: Individual and gender differences in choice of careers in science, technology, engineering, and mathematics. *Psychological Science*, 24 (5), 770–775. doi:10.1177/0956797612458937
- Wigfield, A. & Eccles, J. S. (2000). Expectancy—value Theory of achievement motivation. *Contemporary Educational Psychology*, 25, 68–81. doi:10.1006/ceps.1999.1015
- Wimmer, C. & Oberwimmer, K. (2021). Indikatoren C: Prozesse des Schulsystems. In BMBWF (Hrsg.), *Nationaler Bildungsbericht Österreich 2021. Teil 2 – Bildungsindikatoren* (S. 250–285). Wien: BMBWF. doi:10.17888/nbb2021-2-C

Lisa Wiesinger, Silvia Salchegger & Christina Wallner-Paschon

Das Zusammenspiel von Unterricht und Kompetenzen: Sind Volksschulkinder in Mathematik besser geworden und in Naturwissenschaft zurückgefallen, weil sich der Unterricht verändert hat?

Österreich beteiligte sich im Jahr 2019 zum vierten Mal nach 1995, 2007 und 2011 an der internationalen Schülerleistungsstudie TIMSS (Trends in International Mathematics and Science Study). TIMSS findet seit 1995 in einem Vier-Jahres-Rhythmus statt und erhebt die Kompetenzen in Mathematik und Naturwissenschaft auf der vierten und achten Schulstufe. Mit TIMSS ist es demnach möglich, die Veränderungen von Kompetenzen der österreichischen Volksschulkinder (4. Schulstufe) im Zeitvergleich zu beobachten. Die Ergebnisse von TIMSS 2019 zeigen, dass die Mathematikkompetenz von Schülerinnen und Schülern der vierten Schulstufe in Österreich innerhalb von acht Jahren (2011–2019) beträchtlich zugenommen hat, während gleichzeitig die Kompetenz in Naturwissenschaft signifikant gesunken ist (Itzlinger-Bruneforth & Wiesinger, 2020; Itzlinger-Bruneforth & Brandmair, 2020). Auch im Vergleich mit anderen EU-Ländern ist dieses Muster in Österreich außergewöhnlich. Nur in drei weiteren EU-Ländern lässt sich eine ähnliche Tendenz (Kompetenzzuwachs in Mathematik bei gleichzeitiger Abnahme in Naturwissenschaft) feststellen: Italien, Slowakei und Tschechische Republik. Die Auseinanderentwicklung der Kompetenzen war dort jedoch wesentlich geringer als in Österreich und nur teilweise statistisch signifikant (Wiesinger, Wallner-Paschon & Salchegger, 2023).

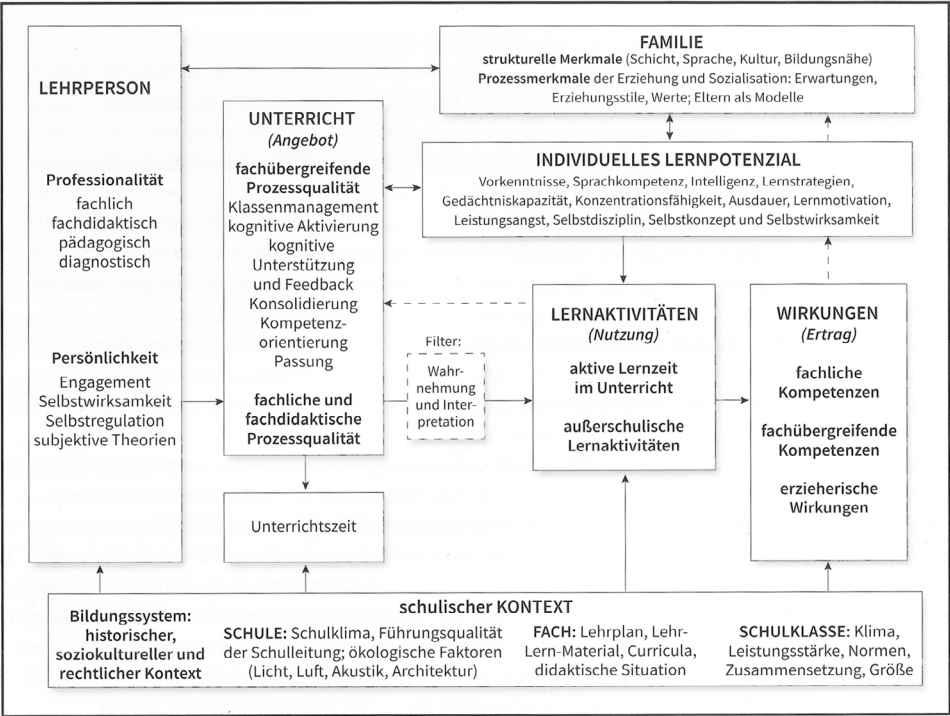
Trenddaten zur Mathematikkompetenz auf der 4. Schulstufe, die einen ähnlichen Zeitraum abdecken (2010–2018), wurden auch bei den nationalen Bildungsstandardüberprüfungen (BIST-Ü) erhoben. Bei den BIST-Ü handelt es sich um eine Vollerhebung aller Schülerinnen und Schüler der 4. bzw. 8. Schulstufe in Österreich. Die Baseline-Testung¹ der Bildungsstandards (4. Schulstufe) fand im Jahr 2010 statt, die BIST-Ü in Mathematik fanden im Jahr 2013 und 2018 statt. Ähnlich wie TIMSS zeigen auch die Ergebnisse der nationalen BIST-Ü eine Zunahme der Mathematikkompetenz von Kindern der 4. Schulstufe: Die Kompetenzen der Viertklässlerinnen und Viertklässler in Österreich sind im Achtjahresvergleich von der Baseline-Testung 2010 auf die BIST-Ü im Jahr 2018 deutlich gestiegen (Schreiner,

1 Die Baseline-Testung wurde etwa zum Zeitpunkt der Einführung der Bildungsstandards anhand einer Stichprobe in Österreich durchgeführt, um eine Bestandsaufnahme des Kompetenzstands für relevante Kompetenzbereiche zu erhalten (Schreiner & Wiesner, 2019).

Breit, Pointinger, Pacher, Neubacher & Wiesner, 2018). Da im Rahmen der BIST-Ü auf der 4. Schulstufe nur Mathematik- und Deutschkompetenzen getestet wurden, gibt es für Naturwissenschaft keine Vergleichswerte von nationalen Überprüfungen.

In diesem Kapitel wird die Größenordnung der gemessenen Kompetenzzunahme in Mathematik zwischen den BIST-Ü und TIMSS verglichen und nach möglichen Erklärungen für die bei TIMSS gefundenen gegenläufigen Entwicklungen gesucht: Kompetenzzuwachs in Mathematik und Kompetenzabnahme in Naturwissenschaft. Zunächst wird der Frage nachgegangen, welche Merkmale generell Einfluss auf Schülerkompetenzen haben, um zu klären, auf welchen Ebenen nach möglichen Ursachen für die Auseinanderentwicklung der Mathematik- und Naturwissenschaftskompetenz gesucht werden kann.

1 Bedingungsfaktoren schulischer Kompetenzen



Um systematisch beschreiben zu können, welche Faktoren den Erwerb schulischer Kompetenzen beeinflussen bzw. bedingen, wird auf das Angebots-Nutzungs-Modell der Wirkungsweise des Unterrichts (Helmke, 2022; Abbildung 1) zurückgegriffen. Als wirksam für den Lernertrag werden in diesem Modell neben der Qualität des Unterrichts sowie Merkmalen der Lehrperson auch Merkmale der individuellen Herkunft der Schülerinnen und Schüler (Familie) sowie die individuellen Lernpotenziale und Lern-

aktivitäten der Schülerinnen und Schüler gesehen. Der Unterricht, die Lernaktivitäten und letztlich auch der Lernertrag werden wiederum durch den schulischen Kontext beeinflusst. Der schulische Kontext wird sehr breit gesehen und umfasst das Bildungssystem mit seiner historischen, soziokulturellen und rechtlichen Einbettung ebenso wie die fachbezogenen Lehrpläne sowie Merkmale der Schule (z. B. Schulklima) und der Schulklasse (z. B. Leistungsstärke, Zusammensetzung). Kompetenzen sind damit multikausal bedingt durch ein Wirkungsgefüge vieler ineinandergreifender und einander beeinflussender Faktoren. Auf der Suche nach Ursachen von Kompetenzveränderungen kann nach diesem Modell vorgegangen werden. Spezifisch für das vorliegende Kapitel wird der Frage nachgegangen, welche Faktoren dieses Modells sich im österreichischen Volksschulbereich so verändert haben, dass sie einen deutlichen Kompetenzanstieg in Mathematik und gleichzeitig einen leichten (aber signifikanten) Kompetenzabfall in den Naturwissenschaften zwischen 2011 und 2019 erklären könnten.

1.1 Familiärer Hintergrund

Bezüglich des familiären Hintergrunds der Schülerinnen und Schüler gab es zwischen 2011 und 2019 einen Anstieg an Schülerinnen und Schülern mit Migrationshintergrund von 19 % auf 28 % (Bergmüller & Herzog-Punzenberger, 2012; Lindemann & Höller, 2020). Ein Kompetenzzuwachs in Mathematik ist durch diese Veränderung allerdings nicht erklärbar, da Schülerinnen und Schüler mit Migrationshintergrund eine wesentlich schwächere (und nicht bessere) Mathematikkompetenz aufweisen als Schülerinnen und Schüler ohne Migrationshintergrund (Lindemann & Höller, 2020).

In Bezug auf den Sozialstatus berichten Salchegger und Höller (2019) von einem Anstieg über die Zeit. Die TIMSS-Ergebnisse zeigen, dass die Naturwissenschaftskompetenz mehr noch als die Mathematikkompetenz mit der sozialen Herkunft der Schülerinnen und Schüler zusammenhängt: Ein höherer Sozialstatus geht mit höheren Kompetenzen einher (Höller & Rothe, 2020; Rothe & Höller, 2020). Ein Kompetenzrückgang in Naturwissenschaft lässt sich durch mehr Schülerinnen und Schüler mit höherem sozialem Status nicht erklären.

1.2 Individuelles Lernpotenzial

Zum individuellen Lernpotenzial zählen im Modell von Helmke (2022) Faktoren wie Intelligenz, Lernmotivation, Selbstdisziplin, Konzentrationsfähigkeit oder Ausdauer. Könnte es sein, dass die österreichischen Schülerinnen und Schüler am Ende der Volksschule im Jahr 2010 bzw. 2011 intelligenter, motivierter, selbstdisziplinierter, konzentrierter und/oder ausdauernder waren als jene im Jahr 2018 bzw. 2019? Die gegenläufige Kompetenzentwicklung in Mathematik und Naturwissenschaft spricht nicht für diese Entwicklung. Denn wenn das allgemeine Lernpotenzial der Kinder gestiegen wäre, wäre davon auszugehen, dass es nicht nur in Mathematik, sondern auch in Naturwissenschaft zu Kompetenzsteigerungen gekommen wäre. Dies ist jedoch nicht der Fall.

1.3 Rechtlicher Kontext des Bildungssystems

Gesetzliche Änderungen, auch wenn sie umgehend umgesetzt werden, schlagen sich in der Regel erst einige Zeit später in den Kompetenzen der Schülerinnen und Schüler nieder. Etwa können mögliche Kompetenzveränderungen nach Einführung eines verpflichtenden Kindergartenjahres erst fünf Jahre später bei Kindern der 4. Schulstufe sichtbar werden und zehn Jahre später bei Pflichtschulabsolventinnen und -absolventen. Im österreichischen Bildungssystem gab es eine Reihe an gesetzlichen Änderungen (Reformmaßnahmen), die bei 2011 getesteten Volksschulkindern noch nicht (oder nicht gänzlich) gewirkt haben können, bei 2019 getesteten Kindern hingegen schon. Diese gesetzlichen Veränderungen beziehen sich auf die Einführung der Bildungsstandards (2008/09), die Einführung der „Schulqualität Allgemeinbildung (SQA)“ (2011/12) bzw. auf Neuerungen in der Elementarbildung, wie die Einführung der frühen sprachlichen Förderung (2008/09) oder des verpflichtenden Kindergartenjahres (2010/11). Einen systematischen Überblick über Bildungsreformen seit 2005 geben Nekula, Bruneforth und Sobanski (2018), eine Bewertung in Bezug auf deren Kompetenzwirksamkeit findet sich bei Salchegger und Höller (2019).

Für das vorliegende Kapitel ist entscheidend, welche dieser Maßnahmen zum oben beschriebenen Muster der Kompetenzentwicklung (Zuwachs in Mathematik, aber Abnahme in Naturwissenschaft) beitragen kann bzw. können. Die Maßnahmen in der Elementarbildung lassen entweder einen allgemeinen Kompetenzzuwachs erwarten oder beziehen sich auf den sprachlichen Bereich. Hingegen erfolgte bei den Bildungsstandards eine Implementierung für Deutsch und Mathematik (im Volksschulbereich), nicht aber für Naturwissenschaft (vgl. Schreiner & Wiesner, 2019), womit sie als Erklärungsmöglichkeit für die gegenläufige Kompetenzentwicklung bei Volksschulkindern in Frage kommen.

Die Bildungsstandards wurden in Österreich im Schuljahr 2008/2009 auf Grundlage des nationalen Lehrplans eingeführt (BGBl. II Nr. 1/2009). Die Bildungsstandards sind konkret formulierte Lernergebnisse. Sie legen die von den Schülerinnen und Schülern erwarteten Kompetenzen in bestimmten Pflichtgegenständen fest. Bildungsstandards beschreiben die erwünschten Lernergebnisse so konkret und detailliert, dass sie in methodisch-didaktische Aufgaben umgesetzt und mithilfe von Testaufgaben überprüft werden können (BGBl. I 117/2008). Die Bildungsstandards in Mathematik (4. Schulstufe) wurden österreichweit zweimal überprüft, 2013 und 2018 (Schreiner & Wiesner, 2019). Dabei wurden alle Schülerinnen und Schüler der 4. Klasse Volksschule in Mathematik getestet und die Ergebnisse wurden ihnen selbst, den Lehrkräften, den Schulleitungen sowie der Schulaufsicht rückgemeldet (BIFIE [Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Bildungswesens], 2019; Schreiner, Suchaň & Salchegger, 2020). Die Bildungsstandards erhöhen die Verbindlichkeit der Lernziele für die Grundkompetenzen und zielen direkt auf den Unterricht ab, indem sie eine Orientierung für die Lehrerinnen und Lehrer bieten, wie gut die Absicherung der grundlegenden Kompetenzen gelungen ist. Die Bildungsstandards wurden zwar bereits im Schuljahr 2008/2009 eingeführt, es ist aber davon auszugehen, dass es einer gewissen Zeitspanne bedurfte, um die erhöhte Kompetenzorientierung im Unterricht zu etablieren (vgl. Abschnitt „Einfluss der Bildungsstandards auf die Unter-

richtsentwicklung“). Mit der ersten flächendeckenden BIST-Ü auf der 4. Schulstufe im Jahr 2013 dürfte diese Umstellung in den Volksschulen in der Breite angekommen sein und zu einem stärker kompetenzorientierten Unterricht in Mathematik (aber möglicherweise nicht in Naturwissenschaft, das nicht in den Bildungsstandards inkludiert war) geführt haben.

Die Implementierung der SQA verstärkt die Bedeutung der BIST-Ü-Ergebnisse, da diese im Rahmen der jährlich stattfindenden Bilanz- und Zielvereinbarungsgespräche zwischen der Schulleitung und der zuständigen Schulqualitätsmanagerin bzw. dem zuständigen Schulqualitätsmanager sowie auch in den Entwicklungsplänen der jeweiligen Schule verpflichtend berücksichtigt werden müssen (Nekula et al., 2018). Demnach dürfte auch in der SQA dem Monitoring und der Entwicklung der Mathematikkompetenz eine wesentlich stärkere Rolle zukommen als der Naturwissenschaftskompetenz, weil für Letztere keine Daten aus BIST-Ü vorhanden sind, die in die SQA einfließen könnten.

1.4 Lehrplan

Die Lehrpläne fokussieren auf zu vermittelnde Lehrstoffinhalte und auf meist allgemein gehaltene Bildungsziele (v. a. Bildungs- und Lehraufgaben, Bildungsbereiche). Der Lehrplan ist laut österreichischem Gesetz der Unterrichtsarbeit zugrunde zu legen (BGBl. I 117/2008). Die Kompetenzbereiche Mathematik und Naturwissenschaft, die bei TIMSS getestet werden, sind in Österreich großteils in den Lehrplänen der Unterrichtsgegenstände Mathematik und Sachunterricht abgedeckt. Der Sachunterrichtslehrplan enthält neben den Erfahrungs- und Lernbereichen Natur, Wirtschaft und Technik auch die Lernbereiche Gemeinschaft, Raum und Zeit. Zum Zeitpunkt der Datenerhebung von TIMSS 2019 war der Volksschullehrplan in der Fassung vom 13. September 2012 gültig (BGBl. Nr. 134/1963 in der Fassung BGBl. II Nr. 303/2012 vom 13. September 2012), wobei der Mathematiklehrplan auf die Version von Juni 2003 zurückgeht und der Sachunterrichtslehrplan auf die Version von Dezember 2010 (BGBl. II Nr. 402/2010, Dezember 2010). Der Lehrplan für technisches Werken beinhaltet ebenfalls einen Teilbereich Technik, in dem sich TIMSS-Inhalte aus Naturwissenschaft widerspiegeln. Im Unterrichtsgegenstand Mathematik gab es von 2003 bis 2019 keine Änderungen im Lehrplan (siehe dazu auch Itzlinger-Bruneforth, Wiesinger, Brandmair & Eibelhuber, 2020). Beim Lehrplan für Sachunterricht erfolgten im Jahr 2010 geringfügige Anpassungen. Diese Änderungen wirkten sich jedoch nicht auf die naturwissenschaftlichen Inhalte aus (diese Lernbereiche wurden weder geändert noch unterschiedlich gewichtet. Generell ist anzumerken, dass es sich beim Lehrplan der österreichischen Volksschule um Rahmenlehrpläne mit breiter inhaltlicher Abdeckung handelt, deren Auswahl und Gewichtung den Lehrkräften überlassen bleibt.

Eine Veränderung in den nationalen Lehrplänen fällt damit als Grund für die gegenläufige Kompetenzentwicklung der Schülerinnen und Schüler in Mathematik und Naturwissenschaft weg. Es wird daher weiter der Frage nachgegangen, ob es systematische Änderungen im implementierten Lehrplan, das heißt in den tatsächlich unterrichteten Inhalten gegeben hat, durch die die gegenläufigen Kompetenzentwicklungen erklärt werden könnten.

1.5 Unterrichtszeit

Laut Stundentafel des österreichischen Volksschullehrplans sind auf der 4. Schulstufe in Mathematik vier Wochenstunden und in Sachunterricht drei Wochenstunden vorgesehen (BGBl. II Nr. 368/2005, November 2005). Durch schulautonome Lehrplanbestimmungen ist es möglich, dass diese Zeit um höchstens eine Wochenstunde erhöht oder verringert wird (BGBl. Nr. 134/1963 in der Fassung BGBl. II Nr. 303/2012 vom 13. September 2012). Da naturwissenschaftliche Inhalte – neben anderen Inhalten – großteils in Sachunterricht (Erfahrungs- und Lernbereich Natur und Technik) und teilweise in technischem Werken (Teilbereich Technik) vermittelt werden, ist hier mehr individueller Spielraum in Bezug auf die Unterrichtszeit gegeben als in Mathematik. Laut Vorgaben im Lehrplan hat es im Betrachtungszeitraum (2010–2019) keine Änderungen im Stundenausmaß gegeben.

1.6 Professionalität der Lehrpersonen

Die Einführung der Bildungsstandards wurde von umfangreichen Schulungsmaßnahmen für Lehrpersonen, Schulleitungen und die Schulaufsicht begleitet, um diese mit dem Thema „Kompetenzorientierung“ vertraut zu machen. Wichtigste Träger dieser Schulungen waren die Pädagogischen Hochschulen (vgl. Müller, Kemethofer, Andreitz, Nachbaur & Soukup-Altrichter, 2019). Entsprechende bundesweite Regelungen wurden ab dem Schuljahr 2014/15 im Rahmen der PädagogInnenausbildung NEU auch in der Lehrerinnen- und Lehrerausbildung implementiert (Nekula et al., 2018). Während die Schulungsmaßnahmen die Professionalität der Lehrkräfte bei den im Jahr 2019 getesteten Kindern positiv beeinflusst haben könnten, haben Lehrkräfte der PädagogInnenausbildung NEU in der Regel noch keine Kinder unterrichtet, die 2019 die 4. Schulstufe besucht haben.

Eine Auswertung von PH-Online-Daten aus dem Studienjahr 2015/16 zeigt, dass es eine hohe Anzahl an Teilnahmen bei Lehrveranstaltungen in der Kategorie „Kompetenzorientierung und Bildungsstandards sowie Leistungsfeststellung“ gegeben hat (Müller, Kemethofer, Andreitz, Nachbaur & Soukup-Altrichter, 2019).

1.7 Unterricht

Wichtig ist die Frage, welche fachlichen Inhalte im Unterricht behandelt (angeboten) wurden und welche der behandelten Inhalte bei den Schülerinnen und Schülern tatsächlich angekommen sind und im Lernertrag (in diesem Fall fachliche Kompetenzen) sichtbar werden. Im *Angebots-Nutzungs-Modell der Wirkungsweise des Unterrichts* (Abbildung 1) stehen zwischen den Lernangeboten im Unterricht und der Wirkung in Form von Kompetenzaeignung zwei weitere Faktoren, die bestimmen, wie viel des Unterrichteten tatsächlich in Schülerkompetenzen umgewandelt wird: Die Wahrnehmung des Unterrichtsinputs durch die Schülerinnen und Schüler sowie dessen Nutzung in ihren Lernaktivitäten. Das heißt, die unterrichteten Inhalte spiegeln sich nicht automatisch in den Kompetenzen der

Schülerinnen und Schüler wider. Bei TIMSS wird in diesem Zusammenhang zwischen dem implementierten Lehrplan (d. h. dem, was die Lehrkräfte als Unterrichtsangebot einbringen) und dem erreichten Lehrplan (d. h. den fachlichen Kompetenzen, die bei den Schülerinnen und Schülern gemessen werden können) unterschieden (Mullis & Martin, 2017; vgl. Abschnitt „Grundlagen der Testung bei TIMSS und BIST-Ü“).

Wie im *Angebots-Nutzungs-Modell der Wirkungsweise des Unterrichts* beschrieben, wird der Unterricht unter anderem durch Systemmerkmale wie den rechtlichen Kontext oder den Lehrplan beeinflusst (siehe oben). Infolge wird genauer darauf eingegangen, wie sich der Unterricht in Mathematik und Naturwissenschaft im Zuge der Einführung der BIST-Ü verändert haben könnte.

2 Einfluss der Bildungsstandards auf die Unterrichtsentwicklung

Die Bildungsstandards dienen vorrangig der nachhaltigen Einführung, Etablierung und Unterstützung der Kompetenzorientierung. Lehrerinnen und Lehrer sind verpflichtet, den systematischen Aufbau der mit den Bildungsstandards definierten grundlegenden Kompetenzen bei der Planung und Gestaltung ihrer Unterrichtsarbeit zu berücksichtigen (BGBl. II Nr. 1/2009). Damit einhergehend steht der Wandel in der Planung, Gestaltung und Durchführung von Unterricht und des unterrichtlichen Handelns im Sinne von „teaching to competencies“, dem langfristigen und planmäßigen Aufbau grundlegender Kompetenzen (Schreiner & Wiesner, 2019). Tatsächlich dürfte es im Zuge der Einführung der Bildungsstandards zu einem stärker kompetenzorientierten Unterricht gekommen sein (Breit & Schreiner, 2021). Diese Veränderung vollzog sich nicht abrupt, sondern über einen längeren Zeitraum hinweg (Fritz & Kulmhofer-Bommer, 2020). Nach Einführung der Bildungsstandards haben sich viele Fortbildungen auf kompetenzorientierten Unterricht fokussiert und somit den Transfer in den Unterricht unterstützt (Müller, Musilek & Wimmer, 2022). Auch wurden Schulbücher auf Basis der Kompetenzorientierung überarbeitet (z. B. Leitner-Jones, 2015). Darüber hinaus wurden unterrichtsnahe Lehr-, Lern- und Begleitmaterialien (z. B. Praxishandbücher² und Aufgabensammlungen³ zu den Fächern Deutsch, Mathematik und Englisch) zur Verfügung gestellt (Altricher & Kanape-Willingshofer, 2012). Diese flankierenden Maßnahmen scheinen eine wesentliche Rolle für die Unterrichtsentwicklung zu spielen. So weist eine qualitative Studie, die 27 Lehrkräfte und Schulleiterinnen/Schulleiter an drei österreichischen Volksschulen einbezog, darauf hin, dass die Lehrkräfte und Schulleitungen im Zuge der Bildungsstandards von einer Entwicklung zu mehr kompetenzorientiertem Unterricht berichten. Diese Veränderung ist meist nicht nur auf die Bildungsstandardreform, sondern auch auf die eigene Weiterbildung, veränderte Rahmenbedingungen, andere Reformmaßnahmen (z. B. Qualitätsmanagementstrategie SQA) und verändertes Unterrichtsmaterial (stärker kompetenzorientierte Schulbücher) zurückzu-

2 <https://www.iqs.gv.at/downloads/nationale-kompetenzerhebung/materialien-zu-ikm-und-bildungsstandards>

3 <https://www.iqs.gv.at/downloads/archiv-des-bifie/unterstuetzende-materialien-zu-bildungsstandards>

führen (Gamsjäger, Altrichter & Steiner, 2019). Gamsjäger et al. (2019) stellten jedoch auch Änderungen im Unterricht fest, die weniger im Sinne einer langfristigen „teaching to competencies“ als vielmehr eines kurzfristigen „teaching to the test“ zu interpretieren sind. Beispielsweise berichteten Lehrkräfte davon, vor einer BIST-Ü-Unterrichtszeit zugunsten des überprüften Gegenstands umzuverteilen oder das Testformat mit den Schülerinnen und Schülern zu üben. Ein derartiges „teaching to the test“ zeigte sich auch für die Sekundarstufe I (Plaimauer, Prammer-Semmler & Altrichter, 2019).

Im Rahmen der BIST-Ü 2018 hatten die Lehrkräfte die Möglichkeit, im Lehrerfragebogen weitere Konsequenzen vor dem Hintergrund der Ergebnisse der BIST-Ü für sich anzugeben. Bei der BIST-Ü 2018 haben 114 der insgesamt 4.925 befragten Lehrkräfte davon Gebrauch gemacht. Auch in diesen Antworten kommt eine Änderung des Unterrichtens zum Ausdruck: einerseits in Bezug auf (langfristig) mehr Kompetenzorientierung und Förderung im Unterricht (z. B. „Einsatz kompetenzorientierter Materialien, kompetenzorientierte Arbeitspläne“, „Jenen Kompetenzbereichen, in welchen die Schülerinnen und Schüler Defizite aufweisen, einen besonderen Stellenwert in der Unterrichtsarbeit einräumen“), andererseits aber auch ein gezieltes Hinarbeiten auf die BIST-Ü (z. B. „Versuch den Kindern vor dem 15. Mai noch viele Mathematikinhalte zu lernen [sic], obwohl wir dafür eigentlich bis Ende Juni Zeit hätten“) (BIFIE, 2020b).

Im Rahmen der BIST-Ü 2013 wurden die Lehrkräfte gefragt, welche anderen Veränderungen (neben einer Veränderung des zeitlichen Aufwands für den Unterricht bzw. dem Austausch von Unterrichtsmaterialien unter Kolleginnen und Kollegen) sich in der Unterrichtsplanung und -gestaltung durch die Umsetzung der Bildungsstandards ergeben. Auf diese offene Frage haben 545 von insgesamt 3.808 Lehrkräften eine Antwort gegeben. In den Antworten wird neben den eben genannten Inhalten unter anderem eine Verschiebung von Unterrichtsinhalten thematisiert. Zum Beispiel: „Vor der Testung wurde verstärkt mit den Bildungsstandards gearbeitet. Daneben wurde der Mathematikstoff schneller durchgenommen (z. B. würden Brüche später kommen)“; „Durch die vermehrten Mathematikeinheiten kommen der Sachunterricht und künstlerisch-kreative Fächer zu kurz“; „Intensivere Arbeit am Verständnis mathematischer Inhalte, teilweise auf Kosten anderer Unterrichtsinhalte (SU, D)“; „Me, Be... kommt zu kurz!!!“ (BIFIE [IQS], 2021).

Als wichtiges Selbstevaluationsinstrument für den Unterricht, das auf den Bildungsstandards bzw. dem Lehrplan beruht, stand ab dem Schuljahr 2009/10 die informelle Kompetenzmessung (IKM) zur Verfügung (Brock, Scherf & Werbowsky, 2011; Friedl-Lucyshyn, 2011). Mithilfe der IKM konnten Lehrkräfte den Kompetenzstand ihrer Schülerinnen und Schüler erheben, um den Förderbedarf zu ermitteln und den Unterricht zu entwickeln. Darüber hinaus sollte die IKM Lehrpersonen dabei unterstützen, Bildungsstandards und Kompetenzorientierung im Unterricht zu verankern (Friedl-Lucyshyn, 2011; Kemethofer, Gugerell, Hafner, Breit & Pacher, 2018). Der Einsatz der IKM war ursprünglich gesetzlich nicht vorgeschrieben, aber dennoch stark genutzt: So wurde die IKM im Schuljahr 2015/16 von

68 % aller Volksschulen österreichweit genutzt, 2017/18 von 82 % (Rechnungshof, 2020). Erst mit dem Jahr 2021 und dem darauffolgenden Ausbau zur iKM^{PLUS} wurde ihr Einsatz verpflichtend.⁴ IKM und iKM^{PLUS} umfass(t)en auch Materialien zu Kompetenzmessung in Naturwissenschaft, jedoch nur für die 7. bzw. 8. Schulstufe (nicht für die Volksschule). In der Folge der IKM berichteten knapp zwei Drittel der von Kemethofer et al. (2018) befragten Lehrkräfte von einer Unterrichtsveränderung in dem Sinn, dass sie sich stärker auf Kompetenzen konzentrieren, die in den Bildungsstandards beschrieben sind. Ob damit eine geringere Konzentration auf die in den Bildungsstandards nicht beschriebenen und bei den BIST-Ü nicht getesteten Naturwissenschaftskompetenzen einhergeht, bleibt offen.

Zusammenfassend konnten im Zuge der Einführung der Bildungsstandards und durch die BIST-Ü-Unterrichtsveränderungen beobachtet werden, die oft nicht unmittelbar durch die Einführung der Bildungsstandards, sondern durch vermittelnde Prozesse wirksam werden. Dabei ist wichtig zu beachten, dass es sich bei den BIST-Ü nicht nur um die ersten Vollerhebungen in Österreich gehandelt hat, sondern auch um die ersten Erhebungen, die mit einer Rechtfertigung (Lehrkräfte gegenüber Schulleitung und Schulleitung gegenüber Schulqualitätsmanagerin/ Schulqualitätsmanager im SQA) verbunden sind. Tatsächlich scheinen die Lehrkräfte ihre Unterrichtsbemühungen nicht nur in langfristige Kompetenzsteigerung investiert zu haben („teaching to competencies“), sondern auch in kurzfristiges „teaching to the test“, um bei den BIST-Ü möglichst gut abzuschneiden und nicht wegen schwacher Ergebnisse zur Verantwortung⁵ gezogen zu werden. Der Unterricht könnte sich im Zuge der Einführung der Bildungsstandards über vielfältige Prozesse verändert haben, die sich sowohl auf ein langfristiges „teaching to competencies“ (1–5) als auch auf ein kurzfristiges „teaching to the test“ beziehen (6–7):

1. Genauere Beschreibung von Lernergebnissen: Die Bildungsstandards spezifizieren die erwarteten Lernergebnisse deutlicher als der Lehrplan (Den Lehrkräften wird klarer, was die Schülerinnen und Schüler können sollen);
2. Lehrkräfte können durch die IKM-Materialien Förderbedarf besser diagnostizieren und Schülerinnen und Schüler gezielter fördern;
3. Dissemination konkreter Aufgaben: Die Aufgabensammlungen zu den Bildungsstandards und die IKM-Aufgaben veranschaulichen, welche konkreten Aufgaben Schülerinnen und Schüler lösen können sollten;
4. Fortbildung der Lehrkräfte in Bezug auf Kompetenzorientierung;
5. Umgestaltung der Schulbücher in Richtung Kompetenzorientierung;
6. Lehrkräfte fokussierten Unterricht auf die in der BIST-Ü getesteten Gegenstände/ Inhalte (und vernachlässigten möglicherweise nicht getestete);
7. Lehrkräfte üben die Testformate gezielt vor einer BIST-Ü.

⁴ <https://www.iqs.gv.at/themen/nationale-kompetenzerhebung/ikm-plus/von-der-ikm-zur-ikm-plus>

⁵ Eine Schulentwicklungsberatung wegen schwacher Ergebnisse bei der BIST-Ü wird von den Schulen (zumindest zu Beginn der Beratung) häufig negativ konnotiert (Hoffmann, Janko, Carmignola & Reindl, 2022).

3 Grundlagen der Testung bei TIMSS und BIST-Ü

Die Testinhalte bei TIMSS orientieren sich an den Lehrplänen der Teilnehmerländer. Eine Herausforderung ist die Lehrplanabdeckung von Lehrplaninhalten unterschiedlichster Teilnehmerländer. Bei TIMSS wird von einem dreistufigen Lehrplanmodell ausgegangen (TIMSS Curriculum Model; Mullis & Martin, 2017; siehe Abbildung 2), das auch auf die österreichische BIST-Ü übertragen werden kann: An dessen oberster Ebene steht der *intendierte Lehrplan* – also die Erfassung jener mathematischen und naturwissenschaftlichen Inhalte, die die Schülerinnen und Schüler laut Vorgaben des jeweiligen Landes in der Schule lernen sollten. Dies ist in Österreich in Lehrplänen und Verordnungen (insbesondere der Bildungsstandardverordnung; BGBl. II Nr. 1/2009 i. d. F. BGBl. Nr. 262/2023) festgehalten. Bei TIMSS sind nicht alle Inhalte, die getestet werden, in den nationalen Lehrplänen aller Teilnehmerländer abgedeckt. Umgekehrt ist es auch möglich, dass in den nationalen Lehrplänen Themen vorgesehen sind, die sich nicht im länderübergreifenden TIMSS-Framework finden. Die zweite Ebene des Curriculum-Modells bildet der *implementierte Lehrplan*. Dieser zeigt, welche Lehrinhalte im Unterricht tatsächlich umgesetzt werden und auf welche Art und Weise diese Inhalte unterrichtet werden. Diese Information wird bei TIMSS mithilfe des Lehrer- und Schulleiterfragebogens erhoben. Zu guter Letzt wird mit den Kompetenzerhebungen festgestellt, welche Inhalte des Lehrplans die Schülerinnen und Schüler tatsächlich gelernt haben und beherrschen – diese dritte Ebene wird bei TIMSS als *erreichter Lehrplan* bezeichnet (Wallner-Paschon, Grafendorfer, Widauer & Suchaň, 2012).

Das TIMSS-Curriculum-Modell geht demnach davon aus, dass die Leistungen von Schülerinnen und Schülern durch den intendierten sowie den implementierten Lehrplan beeinflusst werden. Gegenläufige Kompetenzentwicklungen in Mathematik und Naturwissenschaft (wie in der Einleitung für die österreichischen Volksschulkinder beschrieben) könnten demnach durch Änderungen im intendierten Lehrplan (= nationaler Lehrplan) und/oder durch das, was tatsächlich im Unterricht vermittelt wird (= implementierter Lehrplan), erklärt werden.

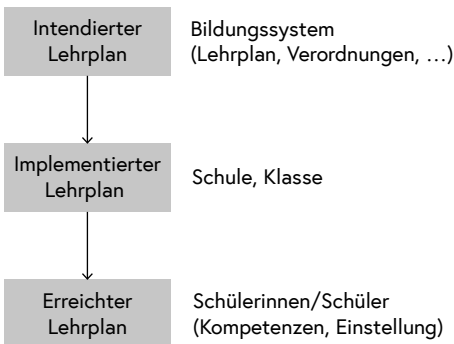


Abbildung 2: Das TIMSS-Curriculum-Modell (Abbildung in Anlehnung an Mullis & Martin, 2017)

Der Baseline-Testung und der BIST-Ü liegen die Bildungsstandards zugrunde, welche aus dem österreichischen Lehrplan abgeleitet wurden. Die Bildungsstandards wurden in Österreich im Schuljahr 2008/2009 in Mathematik und Deutsch (4. und 8. Schulstufe) sowie

Englisch (8. Schulstufe) eingeführt. Diese wurden zwischen 2012 und 2019 abwechselnd (je nach Schulstufe und Unterrichtsfach) flächendeckend überprüft. In den Jahren 2013 und 2018 wurde die BIST-Ü in Mathematik auf der 4. Schulstufe durchgeführt und im Jahr 2010 erfolgte die Baseline-Testung auf der 4. Schulstufe (Schreiner & Wiesner, 2019). Die Kompetenzmodelle im jeweiligen Unterrichtsfach strukturieren die Bildungsstandards unter Berücksichtigung von fachdidaktischen Gesichtspunkten (BGBl. II Nr. 1/2009, § 2, Abs. 4).

Vergleicht man die Bildungsstandards (4. Schulstufe) in Mathematik (BIFIE, 2011), die definieren, welche Inhalte bei den entsprechenden BIST-Ü überprüft wurden, mit dem TIMSS-2019-Framework, das definiert, welche Inhalte bei TIMSS getestet werden (Mullis & Martin, 2017), stehen grundsätzlich vier Inhaltsbereiche der BIST-Ü drei Inhaltsbereichen bei TIMSS gegenüber. Der Bereich *Zahlen* deckt sich in beiden Fällen sehr gut, wobei beispielsweise die Grundrechnungsarten oder einfache Gleichungen bei TIMSS dem Bereich Zahlen zugeordnet werden und bei den Bildungsstandards dem Bereich *Arbeiten mit Operationen*. Die Themen aus *Maße und Geometrie* (TIMSS) teilen sich bei den BIST-Ü auf die Bereiche *Arbeiten mit Größen* sowie *Arbeiten mit Ebene und Raum* auf. Erwähnenswert ist der Bereich Daten, der bei TIMSS einen eigenen Inhaltsbereich darstellt. Bei den Bildungsstandards werden Kompetenzen, wie das *Entnehmen von Informationen aus Tabellen und Grafiken* sowie das *Erstellen von Tabellen, Grafiken, Zeichnungen und Diagrammen*, keinem eigenen Inhaltsbereich zugerechnet, sondern als allgemeine mathematische Kompetenzen aufgefasst, die über unterschiedliche Inhaltsbereiche hinweg zur Anwendung kommen und ebenfalls getestet werden (BIFIE, 2011). Da bei TIMSS die Lehrpläne unterschiedlichster Teilnehmerländer einfließen, kommen vereinzelt Themen vor, die in den österreichischen Bildungsstandards nicht verankert sind (z. B. Dezimalzahlen).

4 Die vorliegende Studie

Der Vergleich von Baseline-Testung/BIST-Ü (2010–2018) sowie von TIMSS (2011–2019) bietet die Gelegenheit, die Entwicklung der Mathematikkompetenz österreichischer Kinder der 4. Klasse Volksschule im Acht-Jahres-Abstand zu vergleichen. In der vorliegenden Studie soll zunächst der Frage nachgegangen werden, ob die positive Entwicklung der Mathematikkompetenz bei der BIST-Ü in ähnlicher Größenordnung wie bei TIMSS ausgefallen ist. Wie oben beschrieben, sind die erhobenen Inhaltsbereiche bei TIMSS und den BIST-Ü zwar ähnlich, aber nicht deckungsgleich. Insbesondere zählt das *Erstellen und Interpretieren von Diagrammen und Grafiken* bei TIMSS als eigener Inhaltsbereich (Daten), bei den Bildungsstandards jedoch als allgemeine Kompetenz, die über unterschiedliche inhaltliche Kompetenzbereiche hinweg zur Anwendung kommt. Wenn sich auch die Kompetenzmodelle nicht decken, so waren doch die erhobenen Kompetenzen ähnlich. Die Autorinnen gehen daher davon aus, dass die Kompetenzen in allen mathematischen Inhaltsbereichen sowohl bei TIMSS als auch bei der BIST-Ü um mindestens ein halbes Lernjahr zugenommen haben, was auf der 4. Schulstufe etwa einem Cohen's d von 0.30 entspricht (Hill, Bloom, Black & Lipsey, 2008; Hypothese 1). Für den bei den Bildungsstandards nicht inkludierten Kompetenzbereich Naturwissenschaft belegen die TIMSS-Daten bereits eine leichte, aber

statistisch signifikante Abnahme. Im vorliegenden Kapitel werden die Veränderungen in den naturwissenschaftlichen Inhaltsbereichen ebenfalls standardisiert, um sie besser mit den Veränderungen in Mathematik vergleichen zu können.

Wie oben beschrieben, dürfte es eine Reihe an Unterrichtsveränderungen im Zuge der Einführung der Bildungsstandards gegeben haben, die die Domäne Mathematik betreffen, nicht aber Naturwissenschaft. Im TIMSS-Lehrerfragebogen wurde detailliert abgefragt, welche Inhalte die Lehrkräfte in Mathematik und Naturwissenschaft bereits unterrichtet haben. Wir gehen davon aus, dass mathematische Inhaltsbereiche im Jahr 2019 zum Testzeitpunkt bereits häufiger unterrichtet worden sind als im Jahr 2011 (Hypothese 2).

Die naturwissenschaftlichen Inhaltsbereiche sind nicht bildungsstandardrelevant. Wir vermuten, dass die naturwissenschaftlichen Inhalte 2019 entweder gleich häufig unterrichtet wurden wie 2011 oder sogar seltener (Hypothese 3).

Um auszuschließen, dass die österreichischen Ergebnisse auf eine allgemeine (etwa gesamteuropäische) Entwicklung zurückgeführt werden können (und nicht auf Veränderungen, die spezifisch für Österreich sind), wird für die TIMSS-Ergebnisse auch der EU-Referenzwert hinzugefügt. Wir gehen davon aus, dass gegenläufige Muster mit a) einem Zuwachs der Mathematikkompetenz bei gleichzeitiger Abnahme der Naturwissenschaftskompetenz und b) mehr abgedeckten mathematischen Inhalten bei gleichzeitig weniger abgedeckten naturwissenschaftlichen Inhalten zwischen 2011 und 2019 spezifisch für Österreich sind und sich im EU-Schnitt nicht zeigen (Hypothese 4).

5 Methodische Vorgehensweise

Das vorliegende Kapitel basiert auf Daten von TIMSS 2011 und TIMSS 2019 sowie auf Daten der nationalen Baseline-Testung 2010 und der BIST-Ü 2018. Bei TIMSS werden jeweils rund 4.500 Schülerinnen und Schüler in Österreich getestet und befragt (siehe Tabelle 1). Darüber hinaus erhalten die Schülerinnen und Schüler selbst sowie ihre Eltern, Lehrkräfte und die Schulleitung einen Fragebogen, um Hintergrundinformationen zu erfassen (IQS [Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen], 2020). Im vorliegenden Kapitel sind die Kontextinformationen auf Ebene der Lehrkräfte relevant, die von 296 Lehrkräften bei TIMSS 2011 bzw. 337 Lehrkräften bei TIMSS 2019 bearbeitet wurden.

Bei der BIST-Ü handelt es sich um eine Vollerhebung und es wurden 2018 in Mathematik rund 75.000 Schülerinnen und Schüler auf der 4. Schulstufe getestet. Die Baseline-Testung im Jahr 2010 wurde mit einer Stichprobe von rund 9.500 Schülerinnen und Schüler durchgeführt und liefert für Österreich repräsentative Daten. Gleich wie bei TIMSS, kamen auch bei der BIST-Ü-Fragebögen für Schülerinnen und Schüler selbst sowie deren Lehrerinnen und Lehrer, die Schulleitung und den Eltern zum Einsatz, um Kontextdaten zu erfassen.

Tabelle 1: Übersicht über die Anzahl einbezogener Schülerinnen und Schüler, Schulen und Klassen bei TIMSS 2011 und 2019 sowie der Baseline-Testung 2010 und der BIST-Ü 2018 (Wintersteller, 2012; Perner, 2020; Breit & Schreiner, 2011; BIFIE, 2019)

Studie/Testung	TIMSS 2011	TIMSS 2019	Baseline-Testung 2010	BIST-Ü 2018
Schulen	158	197	267	2.963
Klassen	276	303	500	4.806
Schülerinnen/Schüler	4.668	4.464	9.478	73.780

5.1 Variablen

In der Folge werden jene Merkmale genauer beschrieben, die in den eben genannten Studien erhoben und im vorliegenden Kapitel analysiert werden.

5.1.1 Kompetenzen in Mathematik und Naturwissenschaft

Die genaue Umsetzung der Kompetenzfassung wird für TIMSS bei Martin und Mullis (2012) sowie Martin, von Davier und Mullis (2020) beschrieben und für die Baseline-Testung und BIST-Ü bei Breit und Schreiner (2016). Die Leistungsskalierung bei TIMSS basiert auf der Erhebung von 1995. Hier wurden die Kompetenzskalen von Mathematik und Naturwissenschaft so verankert, dass sie über alle bei dieser Erhebung auf der 4. Schulstufe teilnehmenden Länder einen Mittelwert von 500 und eine Standardabweichung von 100 aufweisen. Auch die Skalierung nachfolgender Erhebungen basiert auf dieser Verankerung von 1995, um Trendvergleiche anstellen zu können (Foy, Fishbein, von Davier & Yin, 2020). Bei den BIST-Ü erfolgte die Verankerung der Skalen bei der Baseline-Testung im Jahr 2010. Hier wurde der Österreich-Mittelwert auf 500 gesetzt und die Standardabweichung auf 100. Die Skalen von TIMSS und der BIST-Ü weisen damit ähnliche Wertebereiche auf, sind aber dennoch nicht miteinander vergleichbar.

Neben der Mathematik- und Naturwissenschaftsgesamtskala liegen auch Kompetenzwerte für Subdomänen vor. Bei TIMSS werden in Naturwissenschaft die drei Inhaltsbereiche *Biologie*, *Physik* und *Erdkunde* unterschieden. In Mathematik sind es bei TIMSS *Zahlen*, *Maße & Geometrie* sowie *Daten*⁶. Bei der BIST-Ü werden die inhaltlichen mathematischen Kompetenzen *Arbeiten mit Zahlen*, *Arbeiten mit Operationen*, *Arbeiten mit Größen* sowie *Arbeiten mit Ebene und Raum* unterschieden.

5.1.2 Unterrichtete Inhalte in Mathematik und Naturwissenschaft

Die Informationen zu den unterrichteten Inhalten in Mathematik und Naturwissenschaft stammen aus den Lehrerfragebögen zu TIMSS 2011 und TIMSS 2019. Es wurden jeweils eine Reihe an konkreten Unterrichtsinhalten in Bezug auf Mathematik (z. B. Daten aus

6 Wording von TIMSS 2019; bei TIMSS 2011: Zahlen, Geometrische Formen und Maße, Darstellen von Daten.

Tabellen, Piktogrammen, Balken-, Linien- und Kreisdiagrammen ablesen und interpretieren) und Naturwissenschaft (z. B. Aggregatzustände [fest, flüssig, gasförmig] und deren Eigenschaften [Volumen, Form]) vorgegeben (vgl. Tabelle 5 und 7 für eine umfassende Liste). Die Lehrkräfte sollten beantworten, ob diese Inhalte „größtenteils bereits vor diesem Schuljahr unterrichtet“ oder „größtenteils in diesem Schuljahr unterrichtet“ oder ob sie „noch nicht unterrichtet oder gerade erst eingeführt“ wurden. Für alle nachfolgenden Analysen wurden die ersten beiden Kategorien zur Kategorie „größtenteils bereits unterrichtet“ zusammengefasst und es wurden nur jene Items einbezogen, die bei TIMSS 2011 und TIMSS 2019 in identer oder ähnlicher Form erhoben wurden. Der Anteil fehlender Werte liegt bei diesen Einzelitems in Österreich bei TIMSS 2011 in Mathematik zwischen 0,8 % und 3,5 % und in Naturwissenschaft zwischen 3,2 % und 6,2 %. Bei TIMSS 2019 liegt der Anteil fehlender Werte in Mathematik zwischen 0,4 % und 1,6 % und in Naturwissenschaft zwischen 1,8 % und 3,2 %.

Jedes Einzelitem bei TIMSS kann einem der drei mathematischen Inhaltsbereiche (Zahlen, Maße & Geometrie, Daten) oder naturwissenschaftlichen Inhaltsbereiche (Biologie, Physik, Erdkunde) zugeordnet werden. Um einen besseren Gesamtüberblick darüber zu erhalten, wie häufig diese Inhaltsbereiche im Durchschnitt schon unterrichtet wurden, wurde für jeden Inhaltsbereich ein Durchschnitt über die Einzelitems berechnet. So entstanden sechs Skalen, die wiedergeben, ob die Schülerinnen und Schüler schon in Themen dieses Bereichs unterrichtet wurden (siehe Anhang 3 in Foy, Arora & Stanco, 2013; Fishbein, Foy & Yin, 2021). Diese Skalen wurden nicht als Trendskaalen konzipiert und es wurden einige Items zwischen Erhebungen geändert oder gestrichen oder es kamen neue dazu. In die vorliegende Untersuchung werden nur jene Items einbezogen, die zu den Erhebungszeitpunkten 2011 und 2019 gleich oder annähernd gleich lauteten. Ausgehend von diesen wurden die Skalen neu berechnet. Die fehlenden Werte bei diesen Skalen zu den unterrichteten Inhalten schwanken in Österreich in Mathematik bei TIMSS 2011 zwischen 2,8 % und 5,7 % und bei TIMSS 2019 zwischen 0,4 % und 1,6 %. In Naturwissenschaft bewegen sich die fehlenden Werte bei TIMSS 2011 zwischen 6,6 % und 12,3 % bzw. zwischen 2,2 % und 3,9 % bei TIMSS 2019.

5.1.3 EU-Schnitt

Der EU-Schnitt beinhaltet alle Länder, die aktuell EU-Länder sind und sowohl an TIMSS 2011 als auch an TIMSS 2019 teilgenommen haben: Belgien (flämisch), Dänemark, Deutschland, Finnland, Irland, Italien, Kroatien, Litauen, Malta, Niederlande, Österreich, Polen, Portugal, Schweden, Slowakei, Spanien, Tschechische Republik und Ungarn. Damit umfasst der EU-Schnitt die Werte von 80.318 (TIMSS 2011) bzw. 79.559 (TIMSS 2019) Schülerinnen und Schülern.

5.2 Datengewichtung und -analyse

Alle Analysen erfolgen auf Schülerebene, was dem internationalen Standard entspricht (Fishbein et al., 2021). Das heißt, auch die Aussagen der Lehrkräfte müssen auf Schüler-

ebene interpretiert werden, wie z. B. „52 % der Schülerinnen und Schüler wurden im Jahr 2019 bereits im Bruchrechnen unterrichtet“ (und nicht: „52 % der Lehrkräfte haben dies bereits unterrichtet“). Dies wurde bei Unterrichtsanalysen durch die Verwendung des Lehrergewichts (MATWGT bzw. SCIWGT) erreicht.

Für die Analyse der TIMSS-Daten wurde der von der IEA (International Association for the Evaluation of Educational Achievement) zur Verfügung gestellte IDB Analyzer verwendet (IEA, 2022). Dieses Programm berücksichtigt das komplexe Stichprobendesign (Schülerinnen und Schüler innerhalb von Schulen). Für die Analyse der Daten der Baseline-Testung 2010 und der BIST-Ü 2018 wurden die BIFIE-SPSS-Makros – v1.8 verwendet (BIFIE, 2020a), die alle Besonderheiten der BIST-Daten berücksichtigen (vgl. Breit & Schreiner, 2016). Inferenzstatistische Ergebnisse werden dann als statistisch signifikant erachtet, wenn der Alpha-Fehler unter 5 % liegt.

Zur Analyse der vorliegenden Hypothesen ist es notwendig, Veränderungen von Variablen unterschiedlicher Metriken zu vergleichen (z. B. Veränderungen in Kompetenzen mit Veränderungen in der Unterrichtszeit). Um hier eine Vergleichbarkeit herstellen zu können, wurden die für unterschiedliche Skalen gefundenen Differenzwerte in Effektstärkenmaße umgerechnet. Für Mittelwertdifferenzen wurde das Effektstärkenmaß Cohen's d berechnet (vgl. Cohen, 1988). Als Untergrenze für einen kleinen Effekt nennt Cohen (1988) einen Unterschied von $d \geq 0,20$.

6 Ergebnisse

6.1 Veränderungen in der Mathematik- und Naturwissenschaftskompetenz im Acht-Jahres-Vergleich

Sowohl die Ergebnisse der nationalen Testungen (Baseline-Testung 2010 im Vergleich zu BIST-Ü 2018) als auch die Ergebnisse von TIMSS (2011 im Vergleich zu 2019) zeigen wesentliche Zuwächse in der Mathematikkompetenz österreichischer Schülerinnen und Schüler von etwa $d = 0.50$ auf Ebene der Mathematik-Gesamtskala (Tabellen 2 und 3). Betrachtet man die Subskalen der BIST-Ü, fallen die Zuwächse in den inhaltlichen Kompetenzbereichen *Arbeiten mit Zahlen* ($d = 0.75$) und *Arbeiten mit Operationen* ($d = 0.57$) am höchsten aus. Bei TIMSS zeigt sich ganz ähnlich der größte Zuwachs für den Inhaltsbereich Zahlen ($d = 0.59$), gefolgt vom Bereich *Maße und Geometrie* ($d = 0.41$). Beim Inhaltsbereich *Daten* lässt sich nur ein geringer Zuwachs ($d = 0.17$) feststellen. Hypothese 1 kann somit teilweise bestätigt werden: Abgesehen vom Inhaltsbereich *Daten bei TIMSS* haben die Kompetenzen in den mathematischen Inhaltsbereichen sowohl bei TIMSS als auch bei der BIST-Ü um mindestens $d = 0.30$ zugenommen. Für den TIMSS-Mathematikinhaltsbereich *Daten* zeigt sich entgegen Hypothese 1 keine bedeutsame Verbesserung der Kompetenzen: Die Effektstärke von $d = 0.17$ liegt unter der Schwelle von 0.20, die einen kleinen Effekt markiert.

Tabelle 2: Zeitvergleich der Mathematikkompetenz von Schülerinnen und Schülern der 4. Schulstufe in Österreich (Baseline-Testung 2010 und BIST-Ü 2018)

Inhaltlicher Kompetenzbereich	Differenz 2018–2010		
	<i>M</i>	<i>SE</i>	<i>d</i>
Mathematik gesamt	51	2,52	0,52
Arbeiten mit Zahlen	72	2,51	0,75
Arbeiten mit Operationen	57	2,21	0,57
Arbeiten mit Größen	39	2,41	0,39
Arbeiten mit Ebene und Raum	32	2,66	0,32

Anmerkung: *M* = Mittelwert; *SE* = Standardfehler; *d* = Cohen's *d*; statistisch signifikante Differenzen ($p < .05$) sind fett gedruckt.

Tabelle 3: Zeitvergleich der Mathematik- und Naturwissenschaftskompetenz von Viertklässlerinnen und Viertklässlern in Österreich und im EU-Schnitt (TIMSS 2011 und TIMSS 2019)

(Inhalts-)Bereich	AUT			EU		
	Differenz 2019–2011			Differenz 2019–2011		
	<i>M</i>	<i>SE</i>	<i>d</i>	<i>M</i>	<i>SE</i>	<i>d</i>
Mathematik	31	3,29	0,49	8	0,82	0,12
Zahlen	36	3,07	0,59	8	0,84	0,11
Maße & Geometrie	30	4,14	0,41	10	0,96	0,14
Daten	13	4,21	0,17	2	1,01	0,03
Naturwissenschaft	–9	3,85	–0,13	1	0,91	0,01
Biologie	–3	3,63	–0,04	–1	0,96	–0,01
Physik	–16	3,81	–0,21	2	0,98	0,02
Erdkunde	–15	4,94	–0,17	3	1,11	0,04

Anmerkung: *M* = Mittelwert; *SE* = Standardfehler; *d* = Cohen's *d*; statistisch signifikante Differenzen ($p < .05$) sind fett gedruckt.

Die Kompetenzen in Naturwissenschaft wurden nur bei TIMSS erhoben (und nicht bei den BIST-Ü). Laut Tabelle 3 lässt sich bei der Gesamtskala von Österreich ein signifikanter, aber praktisch nicht bedeutsamer Kompetenzrückgang feststellen ($d = -0.13$). Bei der Betrachtung nach Inhaltsbereichen erreicht der Rückgang nur in *Physik* die Schwelle zur praktischen Bedeutsamkeit ($d = -0.21$), in *Erdkunde* liegt er etwas darunter ($d = -0.17$) und in *Biologie* sind die Kompetenzen von 2011 auf 2019 ungefähr gleich geblieben.

Tabelle 3 stellt neben Österreich auch den Zeitvergleich von TIMSS 2011 und 2019 für die 18 EU-Länder dar, die sowohl 2011 als auch 2019 an TIMSS teilgenommen haben. Hier zeigt sich in Mathematik für alle Inhaltsbereiche ein signifikanter, aber praktisch nicht bedeutsamer Kompetenzzuwachs (d schwankt zwischen 0.03 im Inhaltsbereich *Daten* und 0.14 in *Maße & Geometrie*). In Naturwissenschaft sind die Kompetenzen in allen Inhaltsbereichen im EU-Schnitt annähernd gleich geblieben (Cohen's d liegt stets unter 0.05). Hypothese 4a kann angenommen werden: Das Muster der Kompetenzentwicklung in Österreich hebt sich vom EU-Schnitt ab. In Österreich ist die Mathematikkompetenz wesentlich stärker gestiegen als im EU-Schnitt und die Naturwissenschaftskompetenz ist in Österreich geringfügig zurückgegangen, während sie im EU-Schnitt annähernd gleich geblieben ist.

6.2 Veränderungen in den unterrichteten Inhalten (TIMSS 2011–TIMSS 2019)

Tabelle 4 fasst zusammen, wie viele Schülerinnen und Schüler die drei mathematischen Inhaltsbereiche laut Angaben der Lehrpersonen bereits im Unterricht behandelt haben und wie groß die Unterschiede zwischen 2019 und 2011 in Österreich sowie im EU-Schnitt sind.

Tabelle 4: Zeitvergleich unterrichteter Inhalte in Mathematik bei TIMSS 2011 und 2019 in Österreich und im EU-Schnitt

		TIMSS 2011		TIMSS 2019		Differenz 2019–2011	
		%	SE	%	SE	PPT	SE
AUT	Zahlen	66	1,15	76	1,10	9	1,59
	Maße & Geometrie	71	1,73	75	1,38	4	2,21
	Daten	47	2,73	84	1,99	37	3,37
EU-Schnitt	Zahlen	82	0,24	83	0,25	1	0,34
	Maße & Geometrie	65	0,38	70	0,37	5	0,53
	Daten	72	0,59	75	0,59	3	0,83
Differenz AUT–EU-Schnitt	Zahlen	–16	1,17	–7	1,13	9	1,63
	Maße & Geometrie	6	1,77	4	1,43	–2	2,27
	Daten	–25	2,79	9	2,07	34	3,48

Anmerkung: % = Prozent; SE = Standardfehler; PPT = Prozentpunkte; statistisch signifikante Differenzen ($p < .05$ sind fett gedruckt).

Für Österreich zeigt sich hypotesenkonform, dass Inhalte aus dem Bereich *Zahlen* sowie *Daten* im Jahr 2019 signifikant häufiger bereits unterrichtet worden sind. Bei den unter-

richteten Inhalten aus dem Bereich *Maße & Geometrie* ist der Zuwachs hingegen nicht signifikant. Insgesamt haben Inhalte aus dem Bereich Daten den stärksten Zuwachs erfahren: In diesen sind Viertklässlerinnen und Viertklässler im Jahr 2019 wesentlich häufiger bereits unterrichtet worden als jene im Jahr 2011. Hypothese 2 hat sich somit größtenteils bewährt.

Betrachtet man den EU-Schnitt, so zeigt sich, dass es bei den mathematischen Inhalten nur minimale Veränderungen in Bezug auf den Unterricht gibt. Während es in Österreich signifikante Zuwächse in den Inhaltsbereichen *Zahlen* und *Daten* gibt, gibt es diese im EU-Schnitt in den Bereichen *Maße & Geometrie* sowie *Daten*. Der Zuwachs in den Bereichen *Daten* und *Zahlen* fällt in Österreich deutlich höher aus als im EU-Schnitt. Hypothese 4b hat sich somit teilweise bewährt: Die Veränderung unterrichteter Inhalte weicht in Österreich im Bereich *Daten* und *Zahlen* signifikant vom EU-Schnitt ab, nicht jedoch im Bereich *Maße & Geometrie*.

Tabelle 5 fokussiert im Detail auf die Inhalte in den jeweiligen Bereichen und zeigt, wie viele Schülerinnen und Schüler in bestimmten mathematischen Inhalten 2011 bzw. 2019 bereits unterrichtet worden sind. Es werden dabei nur jene Inhalte dargestellt, die sowohl 2011 als auch 2019 gleich oder ähnlich erhoben wurden. Besonders große Zuwächse lassen sich beim Unterrichten von *Daten ordnen und darstellen, um Fragen zu beantworten* aus dem Inhaltsbereich *Daten* feststellen. Während im Jahr 2011 nur 33 % der Schülerinnen und Schüler darin bereits unterrichtet worden sind (laut den Angaben ihrer Lehrkräfte), sind es im Jahr 2019 mit 81 % wesentlich mehr. Im Bereich *Zahlen* zeigt sich insbesondere bei *Bruchrechnung, einschließlich Darstellung, Vergleichen und Ordnen, Addieren und Subtrahieren einfacher Brüche* eine Zunahme: Während im Jahr 2011 nur 19 % der Schülerinnen und Schüler darin bereits unterrichtet worden sind, liegt ihr Anteil im Jahr 2019 bei 52 %.

Tabelle 6 zeigt, wie viele Schülerinnen und Schüler die drei naturwissenschaftlichen Inhaltsbereiche laut Angaben der Lehrpersonen bereits im Unterricht behandelt haben und wie groß die Unterschiede zwischen 2011 und 2019 in Österreich sowie im EU-Schnitt sind. Hypothese 3 geht davon aus, dass die naturwissenschaftlichen Inhalte in Österreich 2019 entweder gleich häufig unterrichtet worden sind wie 2011 oder sogar seltener. Die österreichischen Daten zeigen, dass die Inhalte aus dem Bereich *Physik* 2011 etwa gleich häufig unterrichtet worden sind wie 2019 (keine statistisch signifikante Veränderung). Inhalte aus den Bereichen *Biologie* und *Erdkunde* sind 2019 signifikant seltener unterrichtet worden als 2011. Hypothese 3 hat sich somit empirisch bewährt.

Hypothese 4b geht davon aus, dass die Veränderung in den unterrichteten Inhalten im EU-Schnitt von Österreich abweicht. In allen naturwissenschaftlichen Unterrichtsbereichen fällt der Rückgang in Österreich signifikant stärker aus als im EU-Schnitt. Hypothese 4b hat sich somit bewährt.

Tabelle 5: Bereits unterrichtete Inhalte in Mathematik in Österreich: TIMSS 2011 im Vergleich zu TIMSS 2019

Bereich	Wortlaut	2011		2019		Differenz 2019–2011		Anmerkung
		%	SE	%	SE	%	SE	
Zahlen	Ganze Zahlen, einschließlich Stellenwert und Reihung	100	0,00	98	0,52	-2	0,52	
	Addieren, Subtrahieren, Multiplizieren und Dividieren mit ganzen Zahlen	100	0,00	98	0,62	-2	0,62	
	Gleichungen (die fehlende Zahl finden, Aufgabenstellungen mit Gleichungen darstellen)	46	3,38	57	3,30	11	4,73	a
	Zahlenfolgen (Zahlenfolgen ergänzen und fehlende Zahlen finden)	80	2,83	99	0,43	19	2,86	a
	Bruchrechnung, einschließlich Darstellung, Vergleichen und Ordnen, Addieren und Subtrahieren einfacher Brüche	19	2,95	52	3,62	33	4,67	b
Zahlen MW	Dezimalzahlen, einschließlich Stellenwert und Reihung, Addieren und Subtrahieren mit Dezimalzahlen	55	3,24	49	3,09	-6	4,48	b
	Aufgaben lösen, bei denen es um Längen geht, einschließlich Messung und Schätzung; parallele und senkrechte Geraden	66	1,15	76	1,10	9	1,59	
	Umfang, Fläche und Volumen berechnen und schätzen	97	1,18	99	0,72	1	1,38	b
Maße und Geometrie	Winkel vergleichen und zeichnen	54	3,64	63	3,38	9	4,97	a
	Grundlegende Eigenschaften einfacher geometrischer Figuren	61	3,58	60	2,96	-1	4,64	
	Dreidimensionale Figuren und ihr Zusammenhang mit zugrunde liegenden zweidimensionalen Formen	98	0,94	95	1,56	-3	1,82	a
		45	3,60	56	3,36	11	4,92	a
Maße und Geometrie MW		71	1,73	75	1,38	4	2,21	
Daten	Daten aus Tabellen, Piktogrammen, Balken-, Linien- und Kreisdiagrammen ablesen und interpretieren	53	3,41	89	1,97	36	3,94	b
	Daten ordnen und darstellen, um Fragen zu beantworten	33	3,08	81	2,54	48	3,99	b
	Schlussfolgerungen aus dargestellten Daten ziehen	56	2,90	82	2,34	25	3,73	
Daten MW		47	2,73	84	1,99	37	3,37	

Anmerkung: % = Prozent; SE = Standardfehler; statistisch signifikante Differenzen ($p < .05$ sind fett gedruckt).

a geringfügige Änderung in der Formulierung gegenüber 2011, aber gleicher Inhalt. b geringfügige Änderung in Formulierung und Inhalt gegenüber 2011.

Tabelle 6: Zeitvergleich unterrichteter Inhalte in Naturwissenschaft bei TIMSS 2011 und 2019 in Österreich und im EU-Schnitt

		TIMSS 2011		TIMSS 2019		Differenz 2019–2011	
		%	SE	%	SE	PPT	SE
AUT	Biologie	80	1,31	75	1,75	–6	2,18
	Physik	57	1,97	52	1,61	–5	2,54
	Erdkunde	76	1,33	69	1,87	–8	2,30
EU-Schnitt	Biologie	76	0,39	76	0,40	0	0,56
	Physik	50	0,47	53	0,43	3	0,64
	Erdkunde	67	0,42	66	0,44	–1	0,61
Differenz AUT–EU-Schnitt	Biologie	4	1,37	–2	1,79	–6	2,25
	Physik	7	2,02	–1	1,66	–8	2,62
	Erdkunde	9	1,39	3	1,92	–7	2,37

Anmerkung: % = Prozent; SE = Standardfehler; PPT = Prozentpunkte; statistisch signifikante Differenzen ($p < .05$ sind fett gedruckt).

In Tabelle 7 sind die einzelnen Inhalte der drei naturwissenschaftlichen Bereiche dargestellt. Es werden dabei nur jene Inhalte dargestellt, die sowohl 2011 als auch 2019 gleich oder ähnlich erhoben wurden. In Naturwissenschaft zeigen sich bei keinem Item signifikante Zuwächse, jedoch bei vielen signifikante Rückgänge. Die größten Rückgänge von mindestens 30 Prozentpunkten lassen sich für den Unterricht zum *physikalischen Aufbau der Erdoberfläche* aus dem Inhaltsbereich *Erdkunde* sowie zu *Licht und Geräuschen im Alltag* aus dem Inhaltsbereich *Physik* feststellen.

7 Diskussion und Ausblick

Die Mathematikkompetenz von Volksschulkindern hat sich nach Einführung der Bildungsstandards und der BIST-Ü enorm verbessert. Das zeigen nicht nur die Ergebnisse der BIST-Ü, sondern auch jene von TIMSS. Das vorliegende Kapitel suchte mögliche Gründe dafür. Auf eine allgemeine Verbesserung des individuellen Lernpotenzials bei Volksschulkindern kann die steigende Mathematikkompetenz jedenfalls nicht zurückgeführt werden, da sich ihre Naturwissenschaftskompetenz im gleichen Zeitraum (zwischen 2011 und 2019) leicht verschlechtert hat. Wie lässt sich dieser Zuwachs bei der Mathematikkompetenz, zugleich aber der Rückgang bei der Naturwissenschaftskompetenz nach Einführung der Bildungsstandards erklären? Ganz einfach dadurch, dass die Mathematikkompetenz bei der BIST-Ü überprüft wurde, Naturwissenschaft aber nicht? Dass allein die Überprüfung von Kompetenzen nicht zu höheren Kompetenzen führt, wird in Bildungsdiskussionen gerne durch das Sprichwort

Tabelle 7: Bereits unterrichtete Inhalte in Naturwissenschaft in Österreich: TIMSS 2011 im Vergleich zu TIMSS 2019

Bereich	Wortlaut	2011		2019		Differenz 2019–2011		Anmerkung
		%	SE	%	SE	%	SE	
Biologie	Körperliche Merkmale und Verhaltensweisen von Lebewesen und die Hauptgruppen von Lebewesen (z. B. Säugetiere, Vögel, Insekten, Blütenpflanzen)	92	1,94	88	2,05	-4	2,82	b
	Lebenszyklen von gewöhnlichen Pflanzen und Tieren (z. B. Blütenpflanzen, Schmetterlinge, Frösche)	75	3,20	83	2,78	8	4,24	a
	Wechselwirkungen zwischen Organismen und ihrer Umwelt (z. B. körperliche Merkmale und Verhaltensweisen, die Lebewesen dabei helfen, in ihrer Umgebung zu überleben)	64	3,22	54	3,23	-10	4,56	a
	Zusammenhänge in Ökosystemen (z. B. einfache Nahrungsketten, Räuber-Beute-Beziehungen, Konkurrenzkampf)	72	2,81	56	3,47	-16	4,46	a
Biologie MW	Menschliche Gesundheit (Übertragung und Vorbeugung von Krankheiten, gesundheitsfördernde Verhaltensweisen im Alltag)	99	0,67	91	2,15	-7	2,25	a
	Aggregatzustände (fest, flüssig, gasförmig) und deren Eigenschaften (Volumen, Form)	80	1,31	75	1,75	-6	2,18	
	Klassifizierung von Stoffen nach physikalischen Eigenschaften (z. B. Gewicht/Masse, Volumen, Aggregatzustand, Leitfähigkeit von Wärme oder Elektrizität)	64	3,50	63	2,90	0	4,55	b
	Stoffgemische, einschließlich Methoden zur Trennung eines Stoffgemisches in seine Bestandteile (z. B. Sieben, Filtern, Verdunstung, Verwendung eines Magneten)	56	4,10	39	3,58	-17	5,44	a
Physik	Elektrizität und einfache elektrische Stromkreise (z. B. ein Stromkreis muss geschlossen sein, um ordnungsgemäß zu funktionieren) und Eigenschaften von Magneten (z. B. gleiche Pole stoßen sich ab und ungleiche Pole ziehen sich an, Magnete können bestimmte Gegenstände anziehen)	22	2,98	40	2,78	18	4,08	b
	Chemische Veränderungen im Alltag (z. B. Zerfall, Verbrennen, Rosten, Kochen)	72	4,00	81	2,27	9	4,60	b
	Gewöhnliche Energiequellen (z. B. Sonne, Wind, Erdöl) und Energienutzung (Heizen und Kühlen von Wohnräumen, Erzeugung von Licht)	37	3,37	28	3,20	-9	4,65	b
	Licht und Geräusche im Alltag (z. B. Schatten und Spiegelungen, vibrierende Objekte machen Geräusche)	86	2,57	81	2,73	-5	3,75	a
Physik MW	Licht und Geräusche im Alltag (z. B. Schatten und Spiegelungen, vibrierende Objekte machen Geräusche)	72	2,82	41	3,57	-30	4,55	b
	Kräfte, durch die sich Gegenstände bewegen (z. B. Erdanziehung, Anziehungs-/Abstoßungskräfte) oder ihre Bewegung ändern (z. B. Reibung)	51	3,40	40	3,08	-10	4,59	a
		57	1,97	52	1,61	-5	2,54	

„Vom Wiegen wird die Sau nicht fett“⁷ verbildlicht. Nicht das Messen ändert Kompetenzen, sondern vorgelagerte Maßnahmen bzw. strukturelle Änderungen sind es, die zu verbesserten Testwerten führen (können), vgl. Helmke (2022) – wenn auch festgehalten werden muss, dass die hier vorliegenden Zusammenhänge keine Ursachenerklärungen im Nachhinein zulassen.

Im Zuge der Einführung der Bildungsstandards dürfte es eine Reihe an Unterrichtsveränderungen gegeben haben, einerseits in Bezug auf erhöhte Kompetenzorientierung („teaching to competencies“), andererseits in Bezug auf ein gezieltes Hinlernen auf die Überprüfungen („teaching to the test“). Die vorliegenden Daten von TIMSS zeigen, dass es langfristige Unterrichtsveränderungen gegeben hat. Ein Zeitvergleich von Angaben durch die Lehrkräfte zwischen 2011 und 2019 macht sichtbar, dass es zu einer Zunahme bei der Abdeckung mathematischer Unterrichtsinhalte kam, aber zu einer Abnahme bei der Abdeckung naturwissenschaftlicher Inhalte. Beispielsweise fällt auf, dass Bruchrechnen, das früher oft erst am Ende der 4. Volksschulklasse gelehrt wurde (und zum Zeitpunkt der TIMSS-2011-Testung im April/Mai 2011 oft noch nicht unterrichtet worden war) nach Etablierung der BIST-Ü dauerhaft (auch in Nichtüberprüfungsjahren) früher unterrichtet wurde und die meisten Schülerinnen und Schüler zum Zeitpunkt der TIMSS-2019-Testung darin bereits unterrichtet worden sind. Einen wesentlichen Zuwachs an Unterrichtsaktivität gab es beim Vergleich der Lehrkräfteangaben zwischen 2011 und 2019 auch beim mathematischen Inhaltsbereich Daten (z. B. *Daten ordnen und darstellen, um Fragen zu beantworten*). Aufseiten naturwissenschaftlicher Inhalte war ein Rückgang an Unterrichtsaktivität insbesondere in Bezug auf den *physikalischen Aufbau der Erdoberfläche* sowie auf *Licht und Geräusche im Alltag* zu verzeichnen. Aber auch über *Zusammenhänge in Ökosystemen* (z. B. *Nahrungsketten*) oder hoch gesellschaftlich relevanten Themen wie *natürliche Ressourcen* (z. B. *Wasser, Wind, Boden, Wälder, Erdöl, Erdgas ...*) sowie *Wetter und Klima* wurden im Jahr 2019 bereits weniger Viertklässlerinnen und Viertklässler unterrichtet als im Jahr 2011. Dies führt zur Frage, ob der Fokus auf bildungsstandardrelevante Inhalte den Blick auf naturwissenschaftliche Themen verstellt hat, die hoch zukunftsrelevant sind. Hoch zukunftsrelevant nicht nur deshalb, weil Klimakrise und Ressourcenprobleme immer mehr das alltägliche Leben und die Gesellschaft prägen, sondern auch, weil ein Fachkräftemangel insbesondere in naturwissenschaftlichen Bereichen die gesellschaftliche Entwicklung beeinträchtigt (z. B. Industriellenvereinigung, 2021). Um diese aufgeworfene Frage datengestützt zu beantworten, sollte sie in zukünftige Studien einbezogen werden.

Der Rückgang der Naturwissenschaftskompetenz in der Volksschule war zwischen 2011 und 2019 nur im Inhaltsbereich *Physik* praktisch bedeutsam, fällt in *Erdkunde* etwas geringer aus und lässt sich für *Biologie* nicht feststellen. Einem weiteren Rückgang sollte möglichst gut entgegengewirkt werden bzw. sollte ein Kompetenzanstieg forciert werden. In Bezug auf Maßnahmen, die die Kompetenzentwicklung unterstützen, könnte bei Mathematik, wo ein enormer Fortschritt über die 2010er-Jahre gemessen wurde, Anleihe genommen werden. Möglicherweise waren es in Mathematik eine Reihe an Faktoren bzw. deren Zu-

7 Z. B. Konrad Paul Liessmann (2014): <https://www.furche.at/bildung/liessmann-ueber-bildung-vom-wiegen-wird-die-sau-nicht-fett-1230654>

sammenspiel, die diesen großen Kompetenzzuwachs ermöglicht haben. In der Diskussion stehen eine verstärkte Fortbildung der Lehrkräfte in Bezug auf kompetenzorientiertes Unterrichten, eine stärkere Kompetenzorientierung von Schulbüchern, größere Klarheit, was die Schülerinnen und Schüler durch die IKM können sollen, sowie die Handbücher zu den Bildungsstandards, gezieltere Förderung durch bessere Diagnose bzw. die Verschiebung hin zu bildungsstandardrelevanten Unterrichtsinhalten. Bezogen auf Naturwissenschaft könnten folgende analoge Maßnahmen ergriffen werden:

- Verstärkte Fortbildung der Lehrkräfte in Bezug auf kompetenzorientiertes Unterrichten in Naturwissenschaft;
- stärkere Kompetenzorientierung in Schulbüchern zum Sachunterricht, insbesondere in Bezug auf naturwissenschaftliche Inhalte;
- Aufgabensammlungen, die zeigen, welche Naturwissenschaftsaufgaben Kinder am Ende der Volksschule lösen können sollten bzw. klarere Vorgaben als in den Rahmenlehrplänen;
- Materialien, durch die sich Förderbedarf besser diagnostizieren lässt und mithilfe derer eine gezielte Förderung möglich ist;
- mehr Fokus auf Naturwissenschaftskompetenz im Zuge der Schul(system)entwicklung (u. a. beim Qualitätsmanagementsystem für Schulen).

Es muss betont werden, dass in Bezug auf einige dieser Punkte seit 2019 bereits Maßnahmen ergriffen wurden. So werden seit dem Schuljahr 2023/24⁸ neue Lehrpläne implementiert (Pädagogik-Paket). In diesen Lehrplänen erfolgt ein Paradigmenwechsel hin zu einer stärkeren Kompetenzorientierung. Es geht vorrangig nicht mehr darum, festzulegen, was unterrichtet wird, sondern vielmehr darum, was bei den Schülerinnen und Schülern nachhaltig ankommt (BMBWF [Bundesministerium für Bildung, Wissenschaft und Forschung], 2023f). Mit den neuen Lehrplänen gehen die Kompetenzraster und Lernaufgaben einher. Während die Lehrpläne präzisieren, WAS in einem bestimmten Zeitraum gelernt werden soll, geben die Kompetenzraster darüber hinaus Hinweise, WIE GUT etwas gekonnt werden soll und konkretisieren somit die zu erreichenden Kompetenzen einer Schulstufe (BMBWF, 2023e). Die entwickelten beispielhaften Lernaufgaben illustrieren, wie die in den Kompetenzrastern formulierten Kompetenzen (Teilkompetenzen und Kompetenzstufen) durch Aufgabenstellungen im Unterricht erworben werden können (siehe BMBWF, 2023c; BMBWF, 2023d).

In Bezug auf die oben angeführten Punkte wäre es möglich, dass es im Zuge der stärkeren Kompetenzorientierung des Naturwissenschaftslehrplans auch zu mehr Kompetenzorientierung bei der Aus- und Fortbildung der Lehrkräfte im Bereich Naturwissenschaft kommt sowie zur Gestaltung stärker kompetenzorientierter Schulbücher für den Sachunterricht. Darüber hinaus könnten Kompetenzraster und Lernaufgaben für mehr Klarheit bei den Lehrpersonen sorgen, was konkret gekonnt werden muss.

8 Die neuen Lehrpläne treten aufsteigend ab dem Schuljahr 2023/24 in den ersten Schulstufen der Volksschule, der Mittelschule und der AHS-Unterstufe in Kraft (<https://www.paedagogikpaket.at/images/Kernbotschaften-LP.pdf>).

Weiters gab es in jüngster Zeit (Schuljahr 2023/24) Maßnahmen, um das Interesse an MINT-Berufen [Mathematik, Informatik, Naturwissenschaft und Technik] zu attraktivieren. Digi.case (BMBWF, 2023b) sollte dazu beitragen, problemorientiertes Denken zu etablieren und MINT zu fördern sowie Inspiration für MINT-Berufe zu schaffen. Darüber hinaus wurde der Aktionsplan „MI(N)Tmachen“ erarbeitet. Dieser richtet sich an die gesamte Bildungslandschaft in Österreich, vom Kindergarten bis zur Hochschule und zielt darauf ab, die Anzahl der Absolventinnen und Absolventen im Bereich MINT zu erhöhen. Das Ziel ist, bis zum Jahr 2030 diese Anzahl im Bereich MINT um 20 % zu erhöhen und damit die wirtschaftliche Wettbewerbsfähigkeit in Österreich sicherzustellen (BMBWF, 2023a). Um diese Maßnahme zu erreichen, ist eine frühe Förderung und Auseinandersetzung mit naturwissenschaftlichen Themen unerlässlich.

Einer der oben genannten Punkte, der von den neueren Entwicklungen noch unberührt bleibt, ist ein stärkerer Fokus auf die Naturwissenschaftskompetenz bei der Schul(system)entwicklung. Das SQA wurde seit dem Schuljahr 2021/22 vom Qualitätsmanagementsystem für Schulen (QMS) abgelöst. In diesem spielen nicht mehr die Ergebnisse der BIST-Ü eine wesentliche Rolle beim standortbezogenen Qualitätsmanagement, sondern die Ergebnisse der iKM^{PLUS} (BMBWF, 2022). Durch die iKM^{PLUS} werden IKM und BIST-Ü abgelöst und deren Funktionen zusammengeführt. Die iKM^{PLUS} dient sowohl der Feststellung des Förderbedarfs individueller Schülerinnen und Schüler und der Unterrichtsentwicklung von Klassen (bisher hauptsächlich bei der IKM verortet) sowie der Bereitstellung von Daten für eine evidenzbasierte Schul- und Qualitätsentwicklung (bisher hauptsächlich bei der BIST-Ü verortet). Die Grundlage der Erhebungen der iKM^{PLUS} bleiben weiterhin die seit 2009 festgelegten Bildungsstandards (IQS, 2024).

Die iKM^{PLUS} wurde mit dem Schuljahr 2021/22 stufenweise eingeführt und testet alle Schülerinnen und Schüler der 3. und 4. sowie 7. und 8. Schulstufe in Österreich. Wie schon bei den BIST-Ü und bei der IKM, wird auch bei der iKM^{PLUS} Naturwissenschaft in der Volksschule nicht überprüft. Für die Sekundarstufe I gibt es nichtverpflichtende Bonusmodule in Biologie, Chemie und Physik. In das Qualitätsmanagementsystem für Schulen fließen allerdings nur die verpflichtenden Fokusmodule ein (und damit Naturwissenschaft auch auf der Sekundarstufe I nicht). Darüber hinaus nimmt Österreich bei TIMSS 2023 nicht mehr auf der 4. Schulstufe teil. Damit gibt es seit TIMSS 2019 für die Volksschule in Österreich keine Erhebung, die Rückschlüsse auf die Entwicklung der Naturwissenschaftskompetenz von Volksschulkindern auf Systemebene erlaubt. Dies bedeutet, dass die Wirksamkeit der seit dem Jahr 2020 im Volksschulbereich gesetzten Maßnahmen (vgl. oben) nicht beurteilt werden kann. Für die Zukunft wäre es wünschenswert, die Naturwissenschaftskompetenz systematisch in die schul(system)bezogene Qualitätsentwicklung sowohl für die Primarstufe als auch für die Sekundarstufe aufzunehmen. Dies ist nur möglich, wenn regelmäßige Daten zum Kompetenzbereich Naturwissenschaft erhoben werden. Diese Erhebungen würden neben der Bereitstellung von Daten für die Qualitätsentwicklung auch die Bedeutung von Naturwissenschaft als relevantem Kompetenzbereich in den Schulen stärken und somit der stets steigenden gesellschaftlichen Bedeutung von Naturwissenschaft Rechnung tragen.

Literatur

- Altrichter, H. & Kanape-Willingshofer, A. (2012). Bildungsstandards und externe Überprüfung von Schülerkompetenzen: Mögliche Beiträge externer Messungen zur Erreichung der Qualitätsziele der Schule. In B. Herzog-Punzenberger (Hrsg.), *Nationaler Bildungsbericht Österreich 2012, Band 2* (S. 355–394). Graz: Leykam. doi:10.17888/nbb2012-2
- Bergmüller, S. & Herzog-Punzenberger, B. (2012). Kompetenzen und Charakteristika von Kindern mit und ohne Migrationshintergrund im Zeitvergleich. In B. Suchan, C. Wallner-Paschon, S. Bergmüller & C. Schreiner (Hrsg.), *PIRLS & TIMSS 2011. Schülerleistungen in Lesen, Mathematik und Naturwissenschaft in der Grundschule. Erste Ergebnisse* (S. 52–53). Graz: Leykam.
- BIFIE (Hrsg.). (2011). *Praxishandbuch für „Mathematik“ 4. Schulstufe* (2. Auflage). Graz: Leykam. Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/db40c54aa85bb8adb5839f9e2aee9db5cf0c3203/bist_m_vs_praxishandbuch_mathematik_4_2011-08-22.pdf
- BIFIE (Hrsg.). (2019). *Bundesergebnisbericht – Standardüberprüfung 2018 Mathematik, 4. Schulstufe*. Salzburg: BIFIE.
- BIFIE. (2020a). *SPSS-Makros. Version v1.8* [Software]. Verfügbar unter: <https://www.iqs.gv.at/themen/bildungsforschung/forschungsdatenbibliothek/daten-der-bildungsstandardueberpruefungen>.
- BIFIE. (2020b). *Standardüberprüfung Mathematik 2018, 4. Schulstufe, Gruppenebene (M418K)*, DOI:10.14472/IQSPX69EQ, IQS, V3, UNF:6:9zMJHL1MX5MtbUUKlhLbgQ==
- BIFIE (IQS). (2021). *Standardüberprüfung Mathematik, 4. Schulstufe (2013), Gruppenebene (M413K)*, DOI:10.14472/IQSBP7MG0, IQS, V2, UNF:6:y0Q9fBGCBIFmhp9P53qzEg==
- BM BWf. (2022). *QMS im Überblick*. Verfügbar unter: <https://www.qms.at/images/QMS-im-Ueberblick.pdf>
- BM BWf. (2023a). *Aktionsplan MI(N)Tmachen*. Verfügbar unter: [https://www.bmbwf.gv.at/Themen/HS-Uni/Aktuelles/MI\(N\)T-machen.html](https://www.bmbwf.gv.at/Themen/HS-Uni/Aktuelles/MI(N)T-machen.html)
- BM BWf. (2023b). *Denken lernen, Probleme lösen mit digi.case*. Verfügbar unter: <https://www.bmbwf.gv.at/Themen/schule/zrp/dibi/dgb/dlpl/digicase.html>
- BM BWf. (2023c). *Kompetenzraster: „Mathematik“*. Verfügbar unter <https://paedagogikpaket.at/component/edocman/674-kompetenzraster-1-4-schulstufe-2/download.html?Itemid=0>.
- BM BWf. (2023d). *Kompetenzraster: „Sachunterricht“*. Verfügbar unter <https://paedagogikpaket.at/component/edocman/679-kompetenzraster-1-4-schulstufe-2/download.html?Itemid=0>
- BM BWf. (2023e). *Kompetenzraster und Lernaufgaben*. Verfügbar unter <https://paedagogikpaket.at/massnahmen/lehrplaene-neu/materialien-zu-den-unterrichtsgegenst%C3%A4nden.html>
- BM BWf. (2023f). *Lehrpläne NEU*. Verfügbar unter <https://paedagogikpaket.at/massnahmen/lehrplaene-neu.html>
- Breit, S. & Schreiner, C. (Hrsg.). (2011). *Bildungsstandards: Baseline 2010 (4. Schulstufe). Technischer Bericht*. Salzburg: BIFIE.

- Breit, S. & Schreiner, C. (Hrsg.). (2016). *Large-Scale Assessment mit R*. Wien: Facultas.
- Breit, S. & Schreiner, C. (2021). Leseunterricht vor und nach der Einführung von Bildungsstandards. Ein Zehn-Jahres-Vergleich. In A. Holzinger, S. Kopp-Sixt, S. Luttenberger & D. Wohlhart (Hrsg.), *Fokus Grundschule, Band 2. Qualität von Schule und Unterricht* (S. 163–174). Münster: Waxmann
- Brock, R., Scherf, S. & Werbowsky, I. (2011). IKM – Informelle Kompetenzmessung. In BIFIE (Hrsg.), *Kompetenzorientierter Unterricht in Theorie und Praxis* (S. 139–143). Graz: Leykam. Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/8d36f307bacc072b4f335e58f087e61fa973e843/bist_vs_sek1_kompetenzorientierter_unterricht_2011-03-23.pdf
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.
- Fishbein, B., Foy, P. & Yin, L. (2021). *TIMSS 2019 User Guide for the International Database* (2nd ed.). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Foy, P., Arora, A. & Stanco, G. M. (2013). *TIMSS 2011 User Guide for the International Database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Foy, P., Fishbein, B., von Davier, M. & Yin, L. (2020). Implementing the TIMSS 2019 Scaling Methodology. In M. O. Martin, M. von Davier & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 12.1–12.146). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Friedl-Lucyshyn, G. (2011). Instrumente der Informellen Kompetenzmessung – präzise Diagnosen des Lernstands als Voraussetzung für individualisierte Lernbegleitung. In F. Hofmann, D. Martinek & U. Schwantner (Hrsg.), *Binnendifferenzierter Unterricht und Bildungsstandards – (k)ein Widerspruch?* (S. 63–80). Wien: Lit.
- Fritz, A. & Kulmhofer-Bommer, A. (2020). Evidenzbasierte Unterrichtsentwicklung durch Bildungsstandards und Kompetenzmessungen in Österreich. *Erziehung und Unterricht*, 7–8, 573–581.
- Gamsjäger, M., Altrichter, H. & Steiner, R. (2019). Wirkungen und Wirkungswege einer Bildungsstandard-Reform: Die Sichtweise von Lehrpersonen und Schulleitungen in österreichischen Primarschulen. *Zeitschrift für Bildungsforschung*, 9 (2), 139–158. doi:10.1007/s35834-019-00239-1
- Helmke, A. (2022). *Unterrichtsqualität und Professionalisierung: Diagnostik von Lehr-Lern-Prozessen und evidenzbasierte Unterrichtsentwicklung*. Hannover: Klett/Kallmeyer.
- Hill, C. J., Bloom, H. S., Black, A. R. & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child development perspectives*, 2, 172–177.
- Hoffmann, F., Janko, D., Carmignola, M. & Reindl, C. (2022). *Schlussbericht zur Begleit-evaluation Grundkompetenzen absichern – GruKo*. doi:10.17605/OSF.IO/Y6VZK
- Höller, I. & Rothe, E. (2020). Bildung der Eltern und Schülerleistung. In U. Itzlinger-Bruneforth (Hrsg.), *TIMSS 2019. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Österreich im internationalen Vergleich. Erstbericht* (S. 40–41). Salzburg: IQS. doi:10.17888/timss2019-eb
- Industriellenvereinigung (2021). *MINT-Factsheet*. Verfügbar unter: <https://www.iv.at/-Dokumente-/Publikationen/399-20-pm-mint-factsheet-v8.pdf>

- International Association for the Evaluation of Educational Achievement (IEA). (2022). *Help Manual for the IEA IDB Analyzer* (Version 5.0). Verfügbar unter: <https://www.iea.nl/sites/default/files/2022-06/IDB-Analyzer-Manual-%28Version-5-0%29.pdf>
- Itzlinger-Bruneforth, U. & Brandmair, K. (2020). Naturwissenschaftskompetenz im Trend. In U. Itzlinger-Bruneforth (Hrsg.), *TIMSS 2019. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Österreich im internationalen Vergleich. Erstbericht* (S. 26–27). Salzburg: IQS. doi:10.17888/timss2019-eb
- Itzlinger-Bruneforth, U. & Wiesinger, L. (2020). Mathematikkompetenz im Trend. In U. Itzlinger-Bruneforth (Hrsg.), *TIMSS 2019. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Österreich im internationalen Vergleich. Erstbericht* (S. 14–15). Salzburg: IQS. doi:10.17888/timss2019-eb
- Itzlinger-Bruneforth, U., Wiesinger, L., Brandmair, K. & Eibelhuber, S. (2020). Austria. In D. L. Kelly, V. A. S. Centurino, M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS 2019 Encyclopedia: Education Policy and Curriculum in Mathematics and Science* (pp. 1–11). Verfügbar unter <https://timssandpirls.bc.edu/timss2019/encyclopedia/>
- IQS. (2020). *TIMSS 2019 – Materialien*. Verfügbar unter: <https://iqs.gv.at/downloads/internationale-studien/timss/timss-2019>
- IQS. (2024). *Weiterentwicklung der nationalen Kompetenzmessungen – Von der IKM zur iKM^{PLUS}*. Verfügbar unter: <https://www.iqs.gv.at/ikm>
- Kemethofer, D., Gugerell, S., Hafner, J., Breit, S. & Pacher, K. (2018). Die informelle Kompetenzmessung (IKM) als Ausgangspunkt für Unterrichtsentwicklung. *Erziehung und Unterricht*, 5–6, 495–508.
- Leitner-Jones, C. (2015). *Speaking Tasks in Austrian Lower-School English Coursebooks: developments since the introduction of the Bildungsstandards* ÖFEB-Kongress 2015. Klagenfurt 1.9.2015.
- Lindemann, R. & Höller, I. (2020). Migrationshintergrund und Schülerleistung. In U. Itzlinger-Bruneforth (Hrsg.), *TIMSS 2019. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Österreich im internationalen Vergleich. Erstbericht* (S. 38–39). Salzburg: IQS. doi:10.17888/timss2019-eb
- Martin, M. O. & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Martin, M. O., von Davier, M. & Mullis, I. V. S. (Eds.). (2020). *Methods and Procedures: TIMSS 2019 Technical Report*. Retrieved from <https://timssandpirls.bc.edu/timss2019/methods/index.html>
- Müller, F. H., Kemethofer, D., Andreitz, I., Nachbaur, G. & Soukup-Altrichter, K. (2019). Lehrerfortbildung und Lehrerweiterbildung. In S. Breit, F. Eder, K. Krainer, C. Schreiner, A. Seel & C. Spiel (Hrsg.), *Nationaler Bildungsbericht Österreich 2018, Band 2* (S. 99–142). Graz: Leykam. doi:10.17888/nbb2012-2
- Müller, M., Musilek, M. & Wimmer, C. (2022). Bildungsstandardüberprüfungen in Mathematik auf der Sekundarstufe 1 über den Zeitverlauf 2009–2012–2017. In A. C. George, S. Götz, M. Illetschko & E. Süss-Stepancik (Hrsg.), *Empirische Befunde zu Kompetenzen im Mathematikunterricht der Sekundarstufe 1 und Folgerungen für die Praxis* (S. 55–78). Münster/New York: Waxmann.

- Mullis, I. V. S. & Martin, M. O. (Eds.). (2017). *TIMSS 2019 Assessment Frameworks*. Retrieved from <http://timssandpirls.bc.edu/timss2019/frameworks/>
- Nekula, K., Bruneforth, M. & Sobanski, F. (2018). Entfaltung versus Reformstau: Die Entwicklung des Schulsystems in der ministeriellen Wahrnehmung. In P. Posch, F. Rauch & S. Zehetmeier (Hrsg.), *Das Lernen von Lehrerinnen und Lehrern, Organisationen und Systemen* (S. 215–230). Münster/New York: Waxmann.
- Perner, N. (2020). Stichprobe und Teilnahmequoten. In U. Itzlinger-Bruneforth & N. Perner (Hrsg.), *TIMSS 2019 – Technischer Bericht* (S. 6.1–6.6). Salzburg: IQS. doi:10.17888/timss2019-tb
- Plaimauer, C., Prammer-Semmler, E. & Altrichter, H. (2019). Unterrichts- und Schulentwicklung in der Sekundarschule durch eine Politik der Bildungsstandards? In J. Zuber, H. Altrichter & M. Heinrich (Hrsg.), *Bildungsstandards zwischen Politik und schulischem Alltag* (S. 177–203). Wiesbaden: Springer. doi:10.1007/978-3-658-22241-3_8
- Rechnungshof (2020). *Leseförderung an Schulen. Bericht des Rechnungshofs*. Verfügbar unter: https://www.parlament.gv.at/PAKT/VHG/XXVII/III/III_00091/imfname_780138.pdf
- Rothe, E. & Höller, I. (2020). Beruf der Eltern und Schülerleistung. In U. Itzlinger-Bruneforth (Hrsg.), *TIMSS 2019. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Österreich im internationalen Vergleich. Erstbericht* (S. 42–43). Salzburg: IQS. doi:10.17888/timss2019-eb
- Salchegger, S. & Höller, I. (2019). Haben Reformmaßnahmen im österreichischen Bildungssystem Effekte auf die Lesekompetenz von Kindern mit und ohne Migrationshintergrund? In C. Wallner-Paschon & U. Itzlinger-Bruneforth (Hrsg.), *Lesekompetenz der 10-Jährigen im Trend. Vertiefende Analysen zu PIRLS* (S. 157–176). Graz: Leykam.
- Schreiner, C., Breit, S., Pointinger, M., Pacher, K., Neubacher, M. & Wiesner, C. (2018). *Standardüberprüfung 2017 Mathematik, 8. Schulstufe. Bundesergebnisbericht*. Salzburg: BIFIE.
- Schreiner, C., Suchań, B. & Salchegger, S. (2020). Monitoring student achievement in Austria: Implementation, results and political reactions. In H. Harju-Luukkainen, N. McElvany & J. Stang (Eds.), *Monitoring student achievement in the 21st century* (pp. 65–77). Cham: Springer. doi:10.1007/978-3-030-38969-7_6
- Schreiner, C. & Wiesner, C. (2019). Die Überprüfung der Bildungsstandards in Österreich: Der erste Zyklus als Meilenstein für die Schul- und Unterrichtsentwicklung – eine gelungene Innovation im österreichischen Schulsystem. In A. C. George, C. Schreiner, C. Wiesner, M. Pointinger & K. Pacher (Hrsg.), *Fünf Jahre flächendeckende Bildungsstandardüberprüfungen in Österreich. Vertiefende Analysen zum Zyklus 2012 bis 2016* (S. 13–54). Münster: Waxmann.
- Wallner-Paschon, C., Grafendorfer, A., Widauer, K. & Suchań, B. (2012). Die getesteten Kompetenzbereiche. In B. Suchań & C. Schreiner (Hrsg.), *PIRLS & TIMSS 2011 Technischer Bericht* (S. 11–25). Salzburg: BIFIE.
- Wiesinger, L., Wallner-Paschon, C. & Salchegger, S. (2023, June). *Increase in math but decrease in science. An explanation for Austria's puzzling trends in TIMSS*. Paper presented at the 10th International Research Conference (IEA IRC), Dublin.
- Wintersteller, A. (2012). Stichprobe und Teilnahmequoten. In B. Suchań & C. Schreiner (Hrsg.), *PIRLS & TIMSS 2011 Technischer Bericht* (S. 71–82). Salzburg: BIFIE.

Anhang

A1 (zu Tabelle 2): Zeitvergleich der Mathematikkompetenz von Schülerinnen und Schülern der 4. Schulstufe (Baseline-Testung 2010 und BIST-Ü 2018)

Inhaltlicher Kompetenzbereich	Baseline-Testung 2010 M4			BIST-Ü M4 2018			Differenz 2018–2010		
	<i>M</i>	<i>SE</i>	<i>SD</i>	<i>M</i>	<i>SE</i>	<i>SD</i>	<i>M</i>	<i>SE</i>	<i>d</i>
Mathematik gesamt	500	2,52	100	551	0,06	99	51	2,52	0,52
Arbeiten mit Zahlen	500	2,50	100	572	0,09	90	72	2,51	0,75
Arbeiten mit Operationen	500	2,21	100	557	0,07	99	57	2,21	0,57
Arbeiten mit Größen	500	2,41	100	539	0,08	102	39	2,41	0,39
Arbeiten mit Ebene und Raum	500	2,66	100	532	0,16	101	32	2,66	0,32

Anmerkung: *M* = Mittelwert; *SE* = Standardfehler; *SD* = Standardabweichung; *d* = Cohen's *d*; statistisch signifikante Differenzen ($p < .05$ sind fett gedruckt).

A2 (zu Tabelle 3): Zeitvergleich der Mathematik- und Naturwissenschaftskompetenz österreichischer Viertklässlerinnen und Viertklässler bei TIMSS

(Inhalts-)Bereich	TIMSS 2011			TIMSS 2019			Differenz 2019–2011		
	<i>M</i>	<i>SE</i>	<i>SD</i>	<i>M</i>	<i>SE</i>	<i>SD</i>	<i>M</i>	<i>SE</i>	<i>d</i>
Mathematik	508	2,60	63	539	2,02	65	31	3,29	0,49
Zahlen	506	2,41	59	542	1,91	62	36	3,07	0,59
Maße & Geometrie	512	3,38	75	542	2,39	71	30	4,14	0,41
Daten	515	3,22	77	528	2,72	81	13	4,21	0,17
Naturwissenschaft	532	2,87	70	522	2,58	74	–9	3,85	–0,13
Biologie	526	2,79	69	523	2,32	75	–3	3,63	–0,04
Physik	535	2,80	70	519	2,59	78	–16	3,81	–0,21
Erdkunde	539	3,52	87	524	3,47	87	–15	4,94	–0,17

Anmerkung: *M* = Mittelwert; *SE* = Standardfehler; *SD* = Standardabweichung; *d* = Cohen's *d*; statistisch signifikante Differenzen ($p < .05$ sind fett gedruckt).

A3 (zu Tabelle 3): Zeitvergleich der Mathematik- und Naturwissenschaftskompetenz von Viertklässlerinnen und Viertklässlern im EU-Schnitt bei TIMSS

(Inhalts-)Bereich	TIMSS 2011			TIMSS 2019			Differenz 2019–2011		
	<i>M</i>	<i>SE</i>	<i>SD</i>	<i>M</i>	<i>SE</i>	<i>SD</i>	<i>M</i>	<i>SE</i>	<i>d</i>
Mathematik	516	0,60	70	525	0,57	72	8	0,82	0,12
Zahlen	516	0,61	71	525	0,58	73	8	0,84	0,11
Maße & Geometrie	516	0,71	76	526	0,65	76	10	0,96	0,14
Daten	519	0,74	82	521	0,68	85	2	1,01	0,03
Naturwissenschaft	521	0,67	73	522	0,62	71	1	0,91	0,01
Biologie	525	0,69	74	524	0,66	73	–1	0,96	–0,01
Physik	516	0,71	73	518	0,68	76	2	0,98	0,02
Erdkunde	520	0,80	84	523	0,77	82	3	1,11	0,04

Anmerkung: *M* = Mittelwert; *SE* = Standardfehler; *SD* = Standardabweichung; *d* = Cohen's *d*; statistisch signifikante Differenzen ($p < .05$ sind fett gedruckt).

System- und Schulebene

Einfluss von Schuleffektivitätsfaktoren über den fairen Vergleich hinaus

1 Einleitung

Eine der großen Fragen in der Bildungsforschung ist, was eine gute Schule ausmacht und welche Einflussgrößen Einfluss auf das System Schule haben (Beutel, Höhmann, Schratz & Pant, 2016). Mit dieser Frage beschäftigen sich die Schulforschung, die Schulqualitätsforschung sowie die Schuleffektivitätsforschung und bieten jeweils unterschiedliche Perspektiven auf das Thema, was gute Schulen auszeichnet (Bischof, 2017).

Gute Schulen können unter anderem dadurch charakterisiert werden, dass deren Schülerinnen und Schüler gute Schulleistungen erbringen (Beutel et al., 2016). In der Literatur werden Faktoren genannt, die Einfluss auf die Leistung von Schülerinnen und Schülern haben können (für einen detaillierten Überblick siehe Javornik & Klemenčič Mirazchiyski, 2023; Reynolds et al., 2014), z. B. positives Schulklima (Khan, 2019) und die Schulkultur (Gülse & Çelik, 2021; Schmid, 2018), die Wirksamkeit der Lehrpersonen (Aslantas, 2020; Schmid, 2018) und Führungskräfte (Özgenel & Karsantik, 2020) und sowohl die technologischen Ressourcen (Makharova, Nurzhanova, Adilbayeva, Dossanova & Aimagambetova, 2022) als auch die digitale Kompetenz (Zou, Huang, Ma & Qiu, 2021). Um Schulleistung zu erheben, werden Lernstandserhebungen durchgeführt, die bereits gut erforscht sind (Groß Ophoff, 2012). Schulen mit Berücksichtigung u. a. der Kompositionseffekte der Schülerinnen und Schüler vergleichbar zu machen, ist durch den fairen Vergleich möglich (George & Robitzsch, 2013). Somit können sich Schulen, die einander ähnlich sind, miteinander vergleichen und es kann untersucht werden, welche Einflussfaktoren auf Schulen wirken, die veränderbar sind, ohne dass Disparitäten aufgrund der Komposition der Schülerinnen- und Schülerzusammensetzung reproduziert werden.

Dieses Thema, dass aus theoretisch abgeleiteten Einflussfaktoren über den fairen Vergleich hinaus die Leistung von Schulen vorhergesagt werden kann, ist bisher wenig erforscht. Dies stellt allerdings eine wichtige Information für Maßnahmen der Bildungspolitik dar, da dadurch Verbesserungspotenziale für Schulen offengelegt werden können. In vielen Studien (George, Schreiner, Wiesner, Pointinger & Pacher, 2019; Liu, Peng, Zhao & Lui, 2022; Wang & Wu, 2023) hat sich gezeigt, dass schulische Leistung sehr stark von Hintergrundvariablen der Population der Schülerinnen und Schüler abhängig ist. Um theoretisch

relevante Einflussgrößen von den so gezeigten sozialen Disparitäten in den Schulleistungen analytisch trennen zu können, ist der faire Vergleich wichtig. Das Ziel des vorliegenden Beitrags ist es, die folgende Frage näher zu untersuchen: Welche Schulen sind besser oder schlechter als der vom fairen Vergleich vorhergesagte Wert, und können Einflussfaktoren auf Schulebene diese Über- oder Unterperformanz erklären?

2 Theoretischer Rahmen

Seit den 2000er-Jahren und der Einführung internationaler Vergleichsstudien wie z. B. PISA (Programme für International Student Assessment) ist die evidenzbasierte Schul-, Unterrichts- und Systementwicklung ein integrativer Bestandteil der kontinuierlichen Verbesserungsprozesse des Bildungswesens (Maag Merki, 2016). Ergebnisse aus Large-Scale Assessments (LSA) an Schulen werden für evidenzbasierte Entscheidungsprozesse in Schule und Unterricht genutzt (Bos, Postlethwaite & Gebauer, 2010; Groß Ophoff, 2012). LSA spielen eine zentrale Rolle in der bildungswissenschaftlichen Forschung und Praxis. Diese groß angelegten Schulleistungsuntersuchungen erheben Kompetenzen von Schülerinnen und Schülern in unterschiedlichen Unterrichtsgegenständen und lassen detaillierte Aussagen über den Leistungsstand der gesamten Schülerinnen- und Schülerpopulation treffen. Die gewonnenen Daten bilden die Grundlage für die Steuerung und Planung des Schulwesens sowie für die Weiterentwicklung von Unterricht und Schule. Durch die flächendeckende Durchführung von LSA können unterschiedliche Schülerinnen- und Schülergruppen, Unterrichtspraktiken und Schulstandorte verglichen werden.

Dieser internationalen Entwicklung folgend wurden in Österreich auf nationaler Ebene die Bildungsstandardüberprüfungen (BIST-Ü) eingeführt und zwischen 2009 und 2019 auf der vierten bzw. achten Schulstufe im jeweiligen Fach (Mathematik, Deutsch, Englisch) periodisch durch das Bundesinstitut für Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens (BIFIE) durchgeführt. Weiterführende Informationen zur BIST-Ü können George et al. (2019) entnommen werden. Die BIST-Ü sollen u. a. folgende Punkte ermöglichen (BIFIE, 2019; Schreiner & Breit, 2014):

- Die Ziel- und Ergebnisorientierung des Unterrichts,
- die kontinuierliche Qualitätsentwicklung von Schule und Unterricht sowie
- ein faktenbasiertes System-Monitoring und bildungspolitische Entscheidungen.

2.1 Fairer Vergleich: Berücksichtigung von Kovariaten

Darüber hinaus ermöglichen Bildungsstandards den Vergleich von Schulen untereinander. Dies kann dazu beitragen, Schulen mit Förderbedarf zu identifizieren und gezielt zu unterstützen. Kontextfaktoren haben Einfluss auf Schülerinnen- und Schülerleistungen, so hat z. B. Hattie (2009) auf Grundlage von mehr als 9000 Metaanalysen eine Zusammenführung über den Effekt verschiedener Faktoren erstellt. Es zeigte sich u. a., dass der sozioöko-

nomische Status von Schülerinnen und Schülern durchschnittlich eine Effektgröße von $d = 0.57$ hat, die der Schulgröße im Durchschnitt von $d = 0.43$. Wenn Leistungsvergleiche von Interesse sind, ist es deswegen notwendig, Kontextfaktoren zu berücksichtigen, vor allem dann, wenn Lehrpersonen keinen Einfluss auf diese Kontextfaktoren haben. In Deutschland wird der faire Vergleich verwendet und findet Verbreitung durch die Vergleichsarbeiten (Isaac & Hosenfeld, 2008), die Bestandteil des Bildungsmonitorings auf Systemebene sind (KMK, 2006). Auch in Österreich in den BIST-Ü ist der sogenannte „faire Vergleich“ ein zentrales Element (George & Robitzsch, 2013; Pham, Freunberger, Robitzsch & Itzlinger-Bruneforth, 2016; Pham & Robitzsch, 2014). Dieser Vergleich bewertet die Leistungen von Schülerinnen und Schülern und somit auch von Schulen unter besonderer Berücksichtigung von Faktoren, die außerhalb der Kontrolle der Schulen liegen, jedoch erheblichen Einfluss auf die Leistungen von Schülerinnen und Schülern haben können (Fiege, Reuther & Nachtigall, 2011). Dieser Ansatz hat den Vorteil, gerechtere und objektivere Bewertungen und Vergleiche der schulischen Leistungen zu gewährleisten. Der faire Vergleich berücksichtigt eine Reihe von Kovariaten, die sich auf verschiedene Aspekte beziehen (George & Robitzsch, 2013). Diese Kovariaten umfassen:

1. Standort: Hier wird der Urbanisierungsgrad am Schulstandort berücksichtigt.
2. Schule: Dieser Bereich beinhaltet Faktoren wie die Schulgröße, gemessen an der Anzahl der Schülerinnen und Schüler, und ob eine Schule eine Privatschule oder öffentlich ist.
3. Gruppen: Die Gruppengröße, definiert durch die Anzahl der Schülerinnen und Schüler in einer Unterrichtsgruppe auf Schulebene, wird ebenfalls in Betracht gezogen.
4. Schülerinnen und Schüler: Verschiedene schülerinnen- und schülerbezogene Faktoren werden hier einbezogen. Dazu gehören der Migrationshintergrund, ob Deutsch die Erstsprache der Schülerin/des Schülers ist, das Bildungsniveau der Eltern, der berufliche Status der Eltern (HISEI), die Anzahl der Bücher im Haushalt, das Geschlecht der Schülerin/des Schülers, und ob Schülerinnen und Schüler vom Test ausgeschlossen wurden.

Nähere Informationen zur Zusammenstellung der Kovariaten und Berechnung des fairen Vergleichs können der technischen Dokumentation zum fairen Vergleich entnommen werden (George & Robitzsch, 2013). Die Informationen zu den oben genannten Kovariaten entstammen Kontextbefragungen der Schulleitungen, Lehrpersonen, Schülerinnen und Schüler und Eltern. Diese wurden begleitend zu den BIST-Ü durchgeführt. Durch die Berücksichtigung dieser Kovariaten (fairer Vergleich) wird eine genauere Beurteilung der Schulleistungen ermöglicht, indem sichergestellt wird, dass Unterschiede in den Ergebnissen nicht auf unterschiedliche Hintergründe oder Umgebungen der Schülerinnen und Schüler zurückgeführt werden (Timmermans & Thomas, 2015).

2.2 Schuleffektivitätsforschung: Prädiktoren von Schulleistung

Forschungsergebnisse zeigen, dass eine gute Schule durch mehrere Merkmale charakterisiert werden kann: Gute Lernergebnisse (Hattie, 2009); Gerechtigkeit (Kazemi, 2016); positive Schulentwicklung (Bischof, 2017); gute Schulleitung (Schatz, Hartmann & Schley, 2010);

kollegiales Miteinander (Massenkeil & Rothland, 2016); förderndes Lernklima (Klem & Connell, 2014). Leistungsunterschiede zwischen Schulen, Klassen und Schülerinnen und Schüler lassen sich durch eine Vielzahl von Faktoren erklären (Brühwiler, Helmke & Schrader, 2017).

Die Schuleffektivitätsforschung untersucht die Einflussfaktoren auf den Lernerfolg der Schülerinnen und Schüler, um herauszufinden, welche Faktoren die Leistung der Schülerinnen und Schüler beeinflussen und wie diese Faktoren optimiert werden können, um die Qualität der Schule zu verbessern. Die Schuleffektivitätsforschung fokussiert dabei auf folgende Ebenen (Lindorff, Sammons & Hall, 2020): Schulsysteme, Schulstrukturen, Unterrichtsmerkmale, Merkmale von Lehrpersonen und Merkmale von Schülerinnen und Schüler. Die Schuleffektivitätsforschung liefert wichtige Erkenntnisse für die Schulentwicklung und die Verbesserung der Schulqualität (Scheerens, 1990). Durch die gezielte Optimierung der Einflussfaktoren auf den Lernerfolg kann die Schule allen Schülerinnen und Schülern die Möglichkeit eröffnen, ihre Potenziale zu entfalten und ihre Bildungsziele zu erreichen (Teddlie & Reynolds, 2000). Der Begriff „Schuleffektivität“ ist in der Literatur nicht eindeutig definiert. Eine häufig verwendete Definition ist die, die Schuleffektivität als die Erreichung eines gewünschten Ergebnisses oder Ziels, z. B. die Beherrschung der Lehrplaninhalte durch die Schülerinnen und Schüler (Levine & Lezotte, 1990). Eine weitere Definition fokussiert auf den Organisationsgedanken und sieht Schuleffektivität als die Erfüllung der Ziele einer Schule unter Berücksichtigung der vorhandenen Ressourcen und ohne Überlastung der Mitglieder einer Schule (Bollen, 1996). Häufiger als explizite Definitionen von Schuleffektivität finden sich Definitionen effektiver Schulen. Eine Schule gilt als effektiv, wenn sie den Leistungszuwachs ihrer Schülerinnen und Schüler über die Erwartungen hinaus steigert (Creemers, Peters & Reynolds, 1989; Sammons, Thomas & Mortimore, 1997; Stoll & Fink, 1996). Dieser Ansatz, auch Value-Added-Ansatz genannt, berücksichtigt die unterschiedlichen Ausgangsbedingungen der Schülerinnen und Schüler und misst den Mehrwert, den eine Schule zum Lernerfolg beiträgt. Ziel des Value-Added-Ansatzes ist es, herauszufinden, inwieweit eine Schule über die von ihr nicht beeinflussbaren Faktoren hinaus zur Verbesserung der Lernergebnisse ihrer Schülerinnen und Schüler beiträgt. Dieser Ansatz verwendet meist Längsschnittdaten und ist nach allgemeiner Auffassung die valideste und geeignetste Messgröße für Schuleffektivität (Timmermans & Thomas, 2015). Die Schuleffektivitätsforschung untersucht den Einfluss von Faktoren auf die Lernergebnisse auf verschiedenen Ebenen, z. B. auf der Schulebene, der Ebene der Lehrperson und der Ebene der Schülerinnen und Schüler. Als Kriterien der Schuleffektivität werden die Ergebnisse auf der Ebene der Schülerinnen und Schüler herangezogen (Bischof, 2017). Diese lassen sich in kognitive und nichtkognitive Kriterien einteilen. Kognitive Kriterien beziehen sich auf die Leistung oder Kompetenz der Schülerinnen und Schüler, während nichtkognitive Kriterien Aspekte wie Wohlbefinden, Einstellungen, Motivation und Selbstkonzept umfassen. Die Schuleffektivitätsforschung (Educational Effectiveness Research, EER) ist eng mit der Frage nach den Merkmalen guter Schulen verbunden (Reynolds, Chapman, Kelly, Muijs & Sammons, 2011). Schuleffektivität wird in der Regel über die durchschnittliche Schulleistung operationalisiert und im vorliegenden Beitrag als abhängige Variable verwendet. Dieser Ansatz (Day et al., 2009; Reynolds

et al., 2014) geht davon aus, dass die Gesamtwirkung einer Schule den Einfluss dieser Schule auf eine „durchschnittlichen“ Schülerin/einen „durchschnittliche“ Schüler widerspiegelt.

2.3 Dynamisches Modell der Bildungswirksamkeit

Für die Auswahl der Prädiktoren wird als Grundlage das „dynamic model of educational effectiveness“ (DMEE) verwendet. Das von Creemers und Kyriakides entwickelte Modell stellt ein dynamisches, integratives Modell der Bildungswirksamkeit dar (s. Abbildung 1). Das dynamische Modell der Bildungswirksamkeit ist ein komplexes Theoriemodell, das die Zusammenhänge zwischen verschiedenen Faktoren auf der Schul- und Klassenebene und den Lernergebnissen der Schülerinnen und Schüler beschreibt (Creemers, Kyriakides & Antoniou, 2012). Es basiert auf der Annahme, dass Schuleffektivität ein Mehrebenenmodell mit multiplen Wirkungszusammenhängen ist (Kyriakides & Creemers, 2008). Das DMEE gehört zu den bestuntersuchten Theoriemodellen in der Schulwirksamkeitsforschung. Der vorliegende Artikel fokussiert auf die Ebene der Schule.

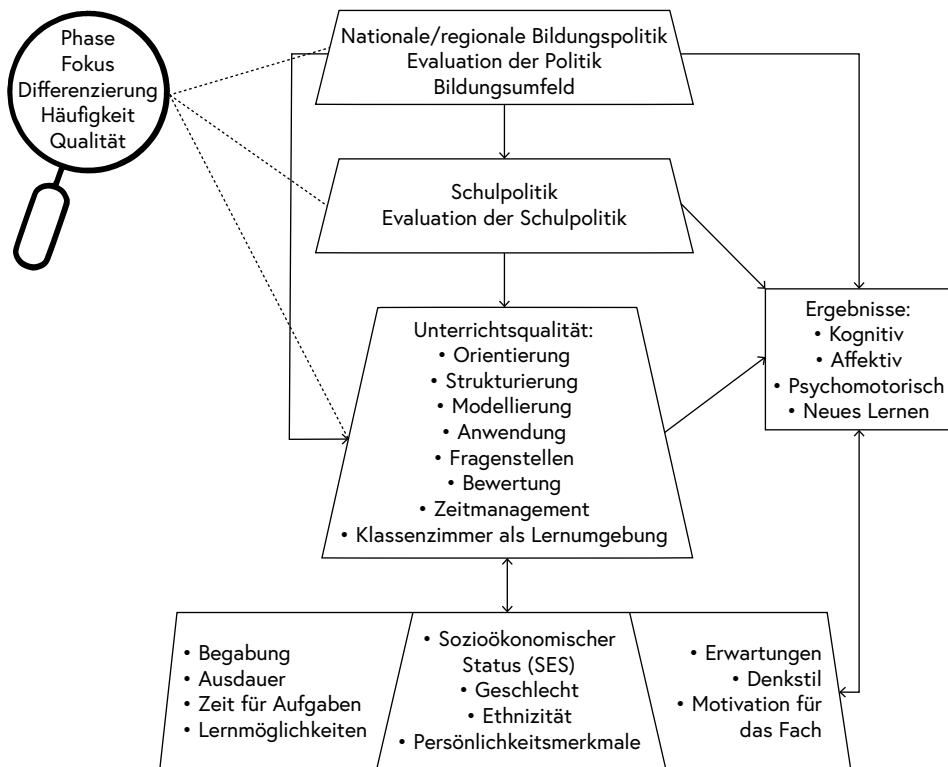


Abbildung 1: Das dynamische Modell der Bildungswirksamkeit (entnommen aus Creemers & Kyriakides, 2010)

Laut Modell kann die Ebene der Schule in folgende Faktoren aufgeteilt werden (Kyriakides, Creemers & Charalambous, 2018):

1. Schulpolitik hinsichtlich des Unterrichts und der Handlungen, die die Unterrichtspraxis verbessern:
 - a. Qualität des Unterrichts (z. B. Unterstützung von Lehrpersonen beim Einsatz von effektiven Lehrmethoden).
 - b. Ausmaß des Unterrichts (z. B. Beginn und Ende der Unterrichtszeit).
 - c. Bereitstellung von Lernmöglichkeiten (z. B. Bereitstellung von Unterstützung für Schülerinnen und Schüler mit besonderem Förderbedarf).
2. Evaluation der Schulpolitik hinsichtlich des Unterrichts und der Handlungen, die die Unterrichtspraxis verbessern (z. B. Überprüfung der Umsetzung der festgelegten Regeln).
3. Richtlinien hinsichtlich der Lernumwelt der Schule und der Handlungen, um die Lernumwelt zu verbessern:
 - a. Verhalten der Schülerinnen und Schüler außerhalb des Klassenzimmers (z. B. Regeln für das Verhalten außerhalb des Klassenzimmers).
 - b. Zusammenarbeit und Interaktion der Lehrpersonen (z. B. Förderung des Austauschs von Lehrpersonen hinsichtlich deren professioneller Weiterentwicklung).
 - c. Politik der Schulpartnerschaft (z. B. Verhältnis zur Gemeinde, den Eltern usw.).
 - d. Bereitstellung von ausreichend Lernressourcen für die Schülerinnen und Schüler und Lehrpersonen (z. B. Bereitstellung von Unterstützungspersonal).
4. Evaluation der Richtlinien hinsichtlich der Lernumwelt der Schule und der Handlungen, um die Lernumwelt zu verbessern (z. B. Kontrolle der in Punkt 3 festgelegten Richtlinien).

Das DMEE bietet einen theoretischen Rahmen, um Zusammenhänge zwischen verschiedenen Faktoren und Schulleistung zu verstehen. Ressourcen und ihre Verfügbarkeit spielen im DMEE eine zentrale Rolle. Die Bereitstellung von Lernressourcen und Lernmöglichkeiten umfasst sowohl materielle Ressourcen (z. B. Lehrbücher, Computer) als auch personelle Ressourcen (z. B. Lehrpersonen, Unterstützungspersonal). Förderangebote, Ganztags- und/oder Nachmittagsbetreuung und zusätzliches Unterstützungspersonal können als Ressourcen gesehen werden und somit als Prädiktoren für Schulleistung betrachtet werden. Die Einschränkung auf diese Faktoren musste vorgenommen werden, weil keine weiteren Prädiktoren auf Schulebene für beide Erhebungszeitpunkte konsistent vorhanden waren. Die Förderung lernschwacher Schülerinnen und Schüler ist nicht nur aus Gründen der Chancengleichheit und Chancengerechtigkeit wichtig, sondern kann auch die gesamte Schulleistung positiv beeinflussen (Kiel, 2017; Semper, Mende & Berkemeyer, 2017): Gezielte Förderprogramme für lernschwache Schülerinnen und Schüler können dazu beitragen, Wissenslücken zu schließen und Kompetenzen zu verbessern (Bruneforth, Weber & Bacher, 2012). Dies kann wiederum zu einer Steigerung der gesamten Schulleistung führen. Die Teilnahme an einer Ganztags- oder Nachmittagsbetreuung kann Schülerinnen und Schüler zusätzliche Möglichkeiten bieten, um zu lernen und Hausaufgaben zu erledigen (Börner, 2011). Darüber hinaus können diese Angebote die soziale und emotionale Entwicklung der

Schülerinnen und Schüler fördern und somit auch ihre Lernbereitschaft und -motivation erhöhen (Greenberg et al., 2003). Zusätzliches Unterstützungspersonal wie Sozialarbeiterinnen und Sozialarbeiter oder Schulpsychologinnen und Schulpsychologen können die Lehrpersonen entlasten und sie sowohl bei pädagogischen als auch bei nichtpädagogischen Aufgaben unterstützen (Vogtenhuber, Lassnigg, Bruneforth, Edelhofer-Lielacher & Siegle, 2015). In Zeiten des Mangels an Lehrpersonen, des hohen administrativen Aufwands für die Lehrpersonen und der psychischen Belastung von Schülerinnen und Schüler könnte das Vorhandensein von Unterstützungspersonal eine zentrale Variable darstellen. Ein Schnittpunkt zwischen der Bereitstellung von Lernressourcen und Lernmöglichkeiten kann die Abfrage der Verwendung von Instrumenten zur Qualitätssicherung und -entwicklung sein. Die in diesem Prädiktor verwendeten Aspekte umfassen Feedback vonseiten der Schülerinnen und Schüler hinsichtlich der vorhandenen Ressourcen, das Vorhandensein von Mentoring und interne/externe Evaluation. Generell soll der Einsatz solcher Instrumente zur Qualitätssteigerung an Schulen beitragen (Altrichter, Kemethofer & George, 2019), im vorliegenden Beitrag soll jedoch der Status quo erhoben und untersucht werden, ob das Vorhandensein dieser Instrumente ein Prädiktor für die Schulleistung und die Abweichung vom Erwartungswert sein kann.

2.4 Zielsetzung des Beitrags: Vorhersage von Schulleistung über den fairen Vergleich hinaus

Folglich wird in diesem Beitrag untersucht, inwiefern eine Auswahl an Schuleffektivitätsfaktoren zur Bereitstellung von Lernmöglichkeiten und Lernressourcen die schulische Leistung in der BIST-Ü Mathematik der 4. Schulstufe in zwei Erhebungszeitpunkten (2013 und 2018) erklären kann und wie diese Einflüsse über den fairen Vergleich hinaus analysiert werden können. Als Prädiktoren wurden fünf Items aus dem Kontextfragebogen der Schulleitungen ausgewählt (siehe Methodenteil für weiterführende Informationen zu den Items) und der Ebene der Bereitstellung von Lernressourcen und Lernmöglichkeiten zugeordnet. Folgende Hypothesen wurden für den vorliegenden Beitrag formuliert:

- Hypothese 1: Die fünf betrachteten Aspekte der Bereitstellung von Lernressourcen und Lernmöglichkeiten sagen die durchschnittliche Schulleistung für beide Erhebungsjahre vorher.
- Hypothese 2: Die fünf betrachteten Aspekte der Bereitstellung von Lernressourcen und Lernmöglichkeiten sagen die Abweichung vom Erwartungswert für beide Erhebungsjahre vorher.

Explorativ wird überprüft, ob Schulen nach Standortmerkmalen (Urbanisierungsgrad, Anzahl der Schülerinnen und Schüler, Anzahl der Schülerinnen und Schüler mit anderer Erstsprache als Deutsch, Schulerhalter) in Gruppen eingeteilt werden können (McCracken & Barcinas, 1991; Opdenakker & Van Damme, 2000; Stewart, 2009) und ob diese einen Zusammenhang mit der durchschnittlichen Schulleistung und der Abweichung vom Erwartungswert aufweisen. Die vier genannten Standortmerkmale bzw. Hintergrundmerkmale

auf Schulebene gingen in die Bestimmung des Erwartungswerts des fairen Vergleichs allein wechselseitig kontrolliert ein. Diese Hintergrundmerkmale sind jedoch nicht unkorreliert, daher ist anzunehmen, dass diese in Clustern auftreten. Diese relevanten Gruppen von Hintergrundvariablen können über den Erwartungswert hinaus weitere Erklärungskraft besitzen. Auf schulischer Ebene zeigen sich in Ergebnissen aus der Hannoverschen Grundschulstudie z. B. Leistungsunterschiede, die mit der Migrationsquote zusammenhängen (Tiedemann & Billmann-Mahecha, 2004), wohingegen andere Studien andere Effekte auf die Schulleistung finden (Dumont, Neumann, Maaz & Trautwein, 2013). Daher sollen die oben genannten Standortmerkmale zusätzlich explorativ in die Analysen miteinbezogen werden.

Der Beitrag geht nicht allein durch die Verknüpfung des DMEE aus der Schuleffektivitätsforschung mit den Daten aus den Bildungsstandardüberprüfungen über den bestehenden Forschungsstand hinaus, sondern auch dadurch, dass die Abweichung vom Erwartungswert aus dem fairen Vergleich vorhergesagt werden soll. Hierbei kann dieser „added value“ darin bestehen, dass nicht nur Ausgangsbedingungen beleuchtet werden, sondern auch Konstellationen von Bedingungen durch den Einbezug der Clusteranalysen. Diese Betrachtung birgt das Potenzial, inhaltsvalider zu sein als die reine Betrachtung von Einzelbedingungen.

3 Methode

3.1 Datengrundlage

Grundlage der Analysen bilden die Daten der Bildungsstandardüberprüfungen aus den Jahren 2013 und 2018 auf der vierten Schulstufe in Mathematik. Die Datenbasis umfasst die Befragungen der Schulleitungen sowie die Testergebnisse der Schülerinnen und Schüler aggregiert als Schulleistungsmittelwert. Bei beiden Erhebungen handelt es sich um Vollerhebungen. In die Analysen flossen nur Schulen ein, die Leistungsdaten zu beiden Testzeitpunkten vorwiesen, also sowohl an den Bildungsstandardüberprüfungen 2013 als auch 2018 teilgenommen hatten. Schulen, die fehlende Werte in einem der Prädiktoren aufwiesen, wurden aus den Analysen ausgeschlossen. Im Datensatz des Erhebungsjahres 2018 wurden vier weitere Schulen ausgeschlossen, da mehr als 1.000 Schülerinnen und Schüler diese vier Schulen besuchten. Nachdem die maximale Schulgröße im Erhebungsjahr 2013 bei 510 Schülerinnen und Schüler lag, wäre mit Einschluss der oben genannten vier Schulen die Vergleichbarkeit der Schulen zwischen den Erhebungszeitpunkten eingeschränkt gewesen. Eine Übersicht über die Anzahl der in die Analysen eingeflossenen Datenpunkte ist Tabelle 1 zu entnehmen. Die Analysen umfassten Korrelationen, multiple Regressionen und Clusteranalysen für beide Erhebungszeitpunkte.

Tabelle 1: Übersicht über die Datenbasis der BIST-Ü 2013 und 2018

Jahr	Schulen	
	Leistungswerte für beide Erhebungszeitpunkte	Nach Ausschluss von fehlenden Werten in den Prädiktoren
2013	2.922	2.756
2018	2.922	2.482

Für die Datenaufbereitung und alle Analysen wurde die Software R (R Core Team, 2023) verwendet.

3.2 Instrumente

Für die Analysen wurden einerseits die Plausible Values ($n = 10$ pro Schule) aus den Bildungsstandardüberprüfungen in Mathematik (2013 und 2018) verwendet und andererseits Items aus den Kontextbefragungen auf Ebene der Schulleiterinnen/Schulleiter (Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen, 2023). Der vorliegende Beitrag fokussiert auf die Faktoren der Bereitstellung von Lernressourcen und Lernmöglichkeiten aus dem DMEE und verwendet folgende Prädiktoren:

- Förderunterricht für lernschwache Schülerinnen und Schüler in Mathematik
- Vorhandensein von ausreichend Unterstützungspersonal
- Vorhandensein eines Ganztagsangebots
- Vorhandensein einer Nachmittags-/Tagesbetreuung
- Verwendung von Instrumenten zur Qualitätssicherung und -entwicklung

Die genaue Formulierung der fünf verwendeten Items ist Tabelle 8 im Anhang zu entnehmen. Zusätzlich wurde die Abweichung vom Erwartungswert wie folgt berechnet: Plausible Value des Schulleistungsmittelwerts – Mittelwert zwischen oberem und unterem Erwartungsbereich.

Die Aufgaben der Bildungsstandardüberprüfung auf der vierten Schulstufe in Mathematik fußen auf den Bildungsstandards, die sich aus den Lehrplänen ableiten. Die Lehrpläne bieten einen Rahmen für den Unterricht und die Lernziele und legen fest, welche Inhalte und Kompetenzen in einem Fach oder einer Domäne vermittelt werden sollten. Um für ein Fach Aufgaben entwickeln zu können, wurden Kompetenzmodelle entwickelt, die verschiedene Fähigkeiten und Fertigkeiten strukturieren. Somit bilden diese Kompetenzmodelle die Grundlage der Entwicklung von Testinstrumenten. In den Bildungsstandardüberprüfungen der vierten Schulstufe in Mathematik ist das Kompetenzmodell in „Inhaltliche mathematische Kompetenzen“ und „Allgemeine mathematische Kompetenzen“ unter-

teilt. Inhaltliche mathematische Kompetenzen untergliedern sich in „Arbeiten mit Zahlen“, „Operationen“, „Größen“ und „Ebenen und Raum“. Die „Allgemeinen mathematischen Kompetenzen“ beinhalten „Modellieren“, „Operieren“, „Kommunizieren“ und „Problemlösen“. Für eine detailliertere Beschreibung des Kompetenzbereichs siehe die Bundesergebnisberichte für die Erhebungsjahre 2013 und 2018 (BIFIE, 2019; Schreiner & Breit, 2014).

4 Ergebnisse

4.1 Gruppierung der Schulen in Cluster

Ein Ziel der Untersuchung war es, zu überprüfen, ob die Hintergrundmerkmale einer Schule, die zum Teil bereits in den fairen Vergleich eingingen, die Schulleistung oder die potenzielle Abweichung von dem zu erwartenden Wert (fairer Vergleich) vorhersagen können. Falls sich Volksschulen in den Erhebungsjahren 2013 und 2018 in Gruppen einteilen lassen könnten, könnten diese Gruppierungen ebenfalls prädiktiv für die Schulleistungen sein. Deswegen wurde in Voranalysen eine K-Means-Clusteranalyse durchgeführt, um Muster in den Merkmalen von Schulen für beide Erhebungsjahre zu identifizieren. Folgende Variablen flossen in die Clusteranalysen ein:

- Urbanisierungsgrad: dicht besiedelt, mittel besiedelt, dünn besiedelt
- Schulerhalter: privat, öffentlich
- Größe der Schule: Anzahl an Schülerinnen und Schülern
- Erstsprache: Anzahl an Schülerinnen und Schülern mit anderer Erstsprache als Deutsch

Für das Erhebungsjahr 2018 wurden vier Schulen mit einer abweichend großen Anzahl ($n > 1000$) an Schülerinnen und Schülern ausgeschlossen. Die Analyse ergab für beide Jahre jeweils drei Cluster mit ähnlichen Eigenschaften. Jeweils erwies sich eine Drei-Cluster-Lösung für alle 10 Datensätze als passend (siehe Abbildungen 2 und 3 im Anhang).

Cluster 1 (kleine Landschulen) repräsentierte Schulen, die in dünn besiedelten/ländlichen Gebieten lagen und von wenigen Schülerinnen und Schülern besucht wurden und einen geringen Anteil an Schülerinnen und Schülern mit anderer Erstsprache als Deutsch aufwiesen. Im Vergleich dazu wies Cluster 2 (mittlere Schulen) leicht höhere Werte bei den zuvor beschriebenen Merkmalen auf. Cluster 3 (große Stadtschulen) beinhaltete Schulen, die im städtischen Bereich lagen, einen größeren Anteil an Schülerinnen und Schülern mit anderer Erstsprache als Deutsch hatten und von einer im Vergleich zu den anderen Clustern größeren Anzahl an Schülerinnen und Schülern besucht wurden. Es wurde festgestellt, dass die Variable über den Status der Schule (öffentlich/privat), die in die Clusteranalyse einbezogen wurde, keine signifikanten Unterschiede zwischen den drei Clustern aufwies.

Tabelle 2: Übersicht über die Anzahl an Schulen nach Erhebungsjahr und Cluster

Erhebungsjahr	Cluster 1: kleine Landschulen	Cluster 2: mittlere Schulen	Cluster 3: große Stadtschulen
2013	354 Schulen	781 Schulen	1.621 Schulen
2018	302 Schulen	734 Schulen	1.446 Schulen

Für die Aufnahme der Cluster in die multiplen Regressionen wurde eine Dummy-Codierung vorgenommen. Schulen im Cluster 1 (kleine, ländliche Schulen mit geringstem Anteil an Schülerinnen und Schülern mit anderer Erstsprache als Deutsch) wurden als Referenzgruppe für die notwendige Dummy-Codierung ausgewählt. Diese Variablen wurden in die multiplen Regressionen als Prädiktoren aufgenommen, um herauszufinden, ob die gefundenen Gruppierungen der Schulen einen Vorhersagewert für die Schulleistung und die Abweichung vom Erwartungswert lieferten.

4.2 Zusammenhänge zwischen den Prädiktoren

Die im Folgenden dargestellten Korrelationen der untersuchten Variablen sollen dazu dienen, multiple Zusammenhänge zu identifizieren, Multikollinearität zu vermeiden und – falls notwendig – relevante Prädiktoren anzupassen. Hierdurch sollen potenzielle zusätzliche Zusammenhänge entdeckt, hohe Korrelationen zwischen Variablen überprüft (Möglichkeit der Beeinträchtigung der Stabilität und Genauigkeit des Regressionskoeffizienten) und die Stärke und Richtung der linearen Beziehung zwischen Variablen untersucht werden.

Für das Erhebungsjahr 2013 fanden sich zwischen den Prädiktoren kaum Zusammenhänge. Ausnahme bildeten drei Korrelationspaare, die geringe bis sehr hohe Zusammenhänge aufwiesen. Das Vorhandensein von Unterstützungspersonal korrelierte sowohl mit dem Vorhandensein von Nachmittagsbetreuung ($r(9) = -.11, p < .001$) als auch mit dem Mittelwert der Schulleistung ($r(9) = .14, p < .001$) gering. Höhere Effekte lassen sich erwartungsgemäß für den Zusammenhang zwischen dem Mittelwert der Schulleistung und der Abweichung vom Erwartungswert (Mittelwert der Schulleistung [PV] und Abweichung vom Erwartungswert ($r(9) = .81, p < .001$) berichten. Zusätzlich korrelierten die Dummy-Variablen aus der Clusteranalyse mit einigen der Prädiktoren und den abhängigen Variablen überwiegend gering. Anders verhielten sich die beiden Korrelationspaare, die mittlere bis starke Zusammenhänge aufwiesen. Große Stadtschulen korrelierten mit dem Vorhandensein von Unterstützungspersonal moderat ($r(9) = .33, p < .001$) und mit der Abweichung vom Erwartungswert hoch ($r(9) = -.75, p < .001$).

Für das Erhebungsjahr 2018 zeigten sich zwischen den Variablen überwiegend zu vernachlässigende Zusammenhänge. Die fünf folgenden Korrelationspaare bildeten hier

jedoch eine Ausnahme, die einen geringen bis sehr hohen Zusammenhang aufwiesen. Das Vorhandensein von Qualitätsinstrumenten korrelierte gering mit dem Vorhandensein einer Nachmittagsbetreuung ($r(9) = -.11, p < .001$), ebenso wie mit dem Vorhandensein der Förderung von schwachen Schülerinnen und Schülern in Mathematik ($r(9) = .11, p < .001$) und dem Vorhandensein von Unterstützungspersonal ($r(9) = -.13, p < .001$). Der Mittelwert der Schulleistung hing ebenfalls gering mit dem Vorhandensein von Unterstützungspersonal zusammen ($r(9) = .15, p < .001$), aber hoch mit der Abweichung vom Erwartungswert ($r(9) = .80, p < .001$). Auch im Erhebungsjahr 2018 wiesen die Dummy-Variablen überwiegend keine bis niedrige Zusammenhänge auf. Ausnahme bildeten die beiden folgenden Korrelationspaare. Große Stadtschulen korrelierten moderat mit dem Vorhandensein von Unterstützungspersonal ($r(9) = .30, p < .001$) und hoch mit der Abweichung vom Erwartungswert ($r(9) = -.77, p < .001$).

Nahezu alle Korrelationen waren höchst signifikant ($p < .001$), was jedoch der hohen Anzahl an Schulen geschuldet ist. Die Korrelationskoeffizienten inkl. Konfidenzintervalle können Tabelle 3 entnommen werden, in der unteren Hälfte sind die Werte von 2013 abgebildet, in der oberen Hälfte die Werte für das Erhebungsjahr 2018.

4.3 Vorhersage der Schulleistung über Schuleffektivitätsfaktoren

Ziel der Regressionsanalysen war es, den Schulmittelwert der Leistung in der Mathematik-BIST-Ü (PV) basierend auf verschiedenen Prädiktoren vorherzusagen. Hierzu wurde für beide Erhebungsjahre (2013, 2018) ein identisches Modell spezifiziert (s. Tabellen 4 und 5). Im Erhebungsjahr 2013 werden 3.8 % der Varianz erklärt, wohingegen im Jahr 2018 6.6 % der Varianz erklärt werden. Die höchst signifikanten Ergebnisse sind wiederum auf die hohe Anzahl an Schulen zurückzuführen und in Populationsanalysen nicht aussagekräftig. Daher werden nur Prädiktoren berichtet, deren $\beta > .10$ beträgt. Nur drei standardisierte Regressionskoeffizienten erwiesen sich für das Erhebungsjahr 2013 als bedeutsam (s. Tabelle 5, Vorhandensein von Unterstützungspersonal; Dummy-Variablen). Wenn mehr Unterstützungspersonal vorhanden war, stieg der Schulmittelwert an ($\beta = .10$). Mitttelgroße Schulen in durchschnittlich dicht besiedeltem Gebiet hatten mit einer mittleren Anzahl an Schülerinnen und Schülern, deren Erstsprache nicht Deutsch ist, einen höheren Schulmittelwert (Dummy-Variable 1) als kleine Schulen in ländlichem Gebiet und mit einer geringeren Anzahl an Schülerinnen und Schülern, deren Erstsprache nicht Deutsch ist ($\beta = .11$). Schulen, die von einer hohen Anzahl an Schülerinnen und Schülern besucht wurden, in städtischem Gebiet lagen und die meisten Schülerinnen und Schüler mit anderer Erstsprache als Deutsch aufwiesen, hatten einen höheren Schulmittelwert (Dummy-Variable 2) als kleine Schulen in ländlichem Gebiet und mit einer geringeren Anzahl an Schülerinnen und Schülern deren Erstsprache nicht Deutsch war ($\beta = .19$).

Tabelle 3: Korrelationstabelle für die Erhebungsjahre 2013 und 2018

Variable	2013		2018		2013		2018		2013		2018	
	M	SD	1	2	3	4	5	6	7	8	9	M SD
1. Vorhandensein von Ganztagsbetreuung	0.02	0.14		.00 [-.01, .01]	.05** [.06, .06]	-.09** [-.01, -.09]	.11** [.12, .12]	-.07** [-.08, -.07]	-.01** [-.02, -.01]	.04** [.04, .04]	-.22** [-.23, -.22]	0.05 0.21
2. Vorhandensein von Nachmittagsbetreuung	0.42	0.49	.04** [.04, .04]		.07** [.07, .08]	-.08** [-.09, -.08]	.04** [.04, .05]	.08** [.07, .09]	.01** [.01, .03]	.10** [.10, .11]	-.07** [-.08, -.07]	0.65 0.48
3. Vorhandensein von Förderung schwächer Schülerinnen und Schüler in Mathematik	0.90	0.30	.04** [.03, .05]	.05** [.05, .06]		-.08** [-.08, -.07]	.11** [.11, .12]	-.06** [-.07, -.06]	-.01 [-.02, -.01]	.09** [.09, .09]	-.13** [-.14, -.13]	0.89 0.32
4. Vorhandensein von Unterstützungspersonal	6.13	3.41	-.03** [-.03, -.03]	-.11** [-.12, -.12]	-.08** [-.1, -.08]		-.13** [-.14, -.13]	.15** [.15, .16]	.01** [.01, .02]	-.19** [-.20, -.19]	.30** [.30, .31]	6.05 3.41
5. Vorhandensein von Qualitätsinstrumenten	2.48	1.36	.05** [.06, .06]	.09** [.09, .09]	.06** [.06, .07]	-.07** [-.08, -.08]		-.03** [-.04, -.03]	.02** [.02, .04]	.11** [.11, .12]	-.20** [-.21, -.20]	3.06 1.40
6. Mittelwert der Schulleistung (PV)	537.33	43.91	-.05** [-.05, -.05]	-.01** [-.02, -.01]	-.03** [-.05, -.03]	.14** [.14, .15]	-.05** [-.06, -.05]		.80** [.81, .81]	-.07** [-.08, -.07]	.19** [.19, .19]	557.23 46.54
7. Abweichung vom Erwartungswert	-24.82	35.37	-.03** [-.03, -.03]	-.05** [-.06, -.05]	.03** [.02, .04]	.01** [.01, .02]	-.06** [-.07, -.05]	.81** [.82, .82]		.02** [.02, .03]	-.02** [-.03, -.02]	-25.44 37.30
8. Dummy-Variable 1 (mittlere Schulen)	0.28	0.45	.01** [.01, .01]	.13** [.13, .14]	.10** [.10, .11]	-.20** [-.21, -.21]	.04** [.04, .04]	-.05** [-.06, -.05]	.02** [.02, .02]		-.77** [-.77, -.77]	0.30 0.46
9. Dummy-Variable 2 (große Schulen)	0.59	0.49	-.16** [-.16, -.16]	-.17** [-.18, -.18]	-.14** [-.15, -.14]	.33** [.33, .33]	-.10** [-.11, -.11]	.14** [.14, .15]	-.01** [-.02, -.01]	-.75** [-.76, -.75]		0.58 0.49

Anmerkungen: M = Mittelwert, SD = Standardabweichung. In eckigen Klammern 95-%-Konfidenzintervall [unterer Wert, oberer Wert]. * p < .05. ** p < .01. Über Diagonale: Werte für 2018; unter Diagonale: Werte für 2013.

Tabelle 4: Ergebnisse der multiplen linearen Regression für das Erhebungsjahr 2013: Vorhersage von Schulleistung

Prädiktoren	Schulmittelwert PV				
	<i>B</i>	<i>95CI</i>	<i>Beta</i>	<i>standardisierter CI</i>	<i>p</i>
(Intercept)	518.54	[517.22, 519.86]	.00	[.00, .00]	< .001
Vorhandensein von Ganztagsbetreuung	-3.91	[-463, -3.20]	-.01	[-.02, .02]	< .001
Vorhandensein von Nachmittagsbetreuung	2.02	[1.65, 2.40]	.02	[.02, .03]	< .001
Vorhandensein von Förderung schwacher Schülerinnen und Schüler in Mathematik	-1.18	[-2.21, -.16]	-.00	[-.02, -.01]	.049
Vorhandensein von Unterstützungspersonal	1.34	[1.29, 1.41]	.10	[-.02, -.02]	< .001
Vorhandensein von Qualitätsinstrumenten	-.83	[-1.04, -.63]	-.03	[-.04, -.02]	< .001
Dummy-Variable 1 (mittlere Schulen)	10.73	[10.28, 11.18]	.11	[.11, .12]	< .001
Dummy-Variable 2 (große Stadtschulen)	16.77	[16.19, 17.35]	.19	[.19, .20]	< .001
Beobachtungen	2.756				
<i>R</i> ²	.038				

Tabelle 5: Ergebnisse der multiplen linearen Regression für das Erhebungsjahr 2018: Vorhersage von Schulleistung

Prädiktoren	Schulmittelwert PV				
	<i>B</i>	<i>95CI</i>	<i>Beta</i>	<i>standardisierter CI</i>	<i>p</i>
(Intercept)	524.95	[523.82, 526.08]	.00	[.00, .00]	< .001
Vorhandensein von Ganztagsbetreuung	-2.39	[-3.68, -1.11]	-.01	[-.02, -.01]	.005
Vorhandensein von Nachmittagsbetreuung	8.94	[8.19, 9.69]	.09	[.09, .10]	< .001
Vorhandensein von Förderung schwacher Schülerinnen und Schüler in Mathematik	-4.88	[-5.80, -3.97]	-.03	[-.04, -.03]	< .001
Vorhandensein von Unterstützungspersonal	1.45	[1.38, 1.53]	.11	[.11, .12]	< .001
Vorhandensein von Qualitätsinstrumenten	.76	[.49, 1.04]	.02	[.02, .04]	< .001
Dummy-Variable 1 (mittlere Schulen)	15.74	[15.27, 16.21]	.15	[.15, .16]	< .001
Dummy-Variable 2 (große Stadtschulen)	26.08	[25.66, 26.51]	.28	[.28, .29]	< .001
Beobachtungen	2.482				
<i>R</i> ²	.066				

Für das Erhebungsjahr 2018 (s. Tabelle 5) zeigten sich sehr ähnliche Ergebnisse, jedoch waren die standardisierten Regressionskoeffizienten für die für das Erhebungsjahr 2013 oben beschriebenen Prädiktoren etwas höher. Zusammenfassend kann festgehalten werden, dass der Schulmittelwert für keines der beiden Erhebungsjahre durch die fünf Prädiktoren betreffend die Lernressourcen und Lernmöglichkeiten vorhergesagt wurde.

4.4 Vorhersage der Abweichung vom Erwartungswert über Schuleffektivitätsfaktoren

Ziel der Regressionsanalysen war es, die Abweichung vom Erwartungswert des Schulmittelwerts der Leistung in der Mathematik-BIST-Ü (PV) basierend auf fünf Prädiktoren zur Bereitstellung von Lernressourcen vorherzusagen. Hierzu wurde für beide Erhebungsjahre (2013, 2018) ein identisches Modell spezifiziert (s. Tabelle 6 und 7). Im Erhebungsjahr 2013 wurden 0.8 % der Varianz erklärt, wohingegen im Jahr 2018 0.2 % der Varianz erklärt wurden. Die höchstsignifikanten Ergebnisse waren wiederum auf die hohe Anzahl an Schulen zurückzuführen.

Tabelle 6: Ergebnisse der multiplen linearen Regression für das Erhebungsjahr 2013: Vorhersage der Abweichung vom Erwartungswert

Prädiktoren	Differenz vom Erwartungswert				
	<i>B</i>	<i>95CI</i>	<i>Beta</i>	<i>standardisierter CI</i>	<i>p</i>
(Intercept)	-23.05	[-24.49, -21.62]	.00	[.00, .00]	< .001
Vorhandensein von Ganztagsbetreuung	-6.19	[-6.91, -5.49]	-.03	[-.03, -.03]	< .001
Vorhandensein von Nachmittagsbetreuung	-3.83	[-4.80, -3.47]	-.05	[-.06, -.05]	< .001
Vorhandensein von Förderung schwacher Schülerinnen und Schüler in Mathematik	3.72	[2.65, 4.81]	.03	[.03, .05]	< .001
Vorhandensein von Unterstützungspersonal	.11	[.06, .18]	.01	[.01, .02]	< .001
Vorhandensein von Qualitätsinstrumenten	-1.38	[-1.59, -1.19]	-.05	[-.07, -.05]	< .001
Dummy-Variable 1 (mittlere Schulen)	.72	[.34, 1.11]	.01	[.01, .02]	< .001
Dummy-Variable 2 (große Stadtschulen)	-1.37	[-1.93, -.82]	-.02	[-.03, -.02]	< .001
Beobachtungen	2.756				
<i>R</i> ²	.008				

Die oben beschriebenen Effekte konnten sich weder für das Erhebungsjahr 2013 (s. Tabelle 6) noch für das Erhebungsjahr 2018 (s. Tabelle 7) replizieren. Die kleinen Effekte für das Vorhandensein von Unterstützungspersonal und beide Gruppierungen aus der Clusterana-

lyse verschwanden in den beiden folgend dargestellten Regressionsergebnissen nahezu vollständig. Die Abweichung vom Erwartungswert konnte weder für das Erhebungsjahr 2013 noch für das Erhebungsjahr 2018 durch die fünf Prädiktoren zur Bereitstellung von Lernressourcen vorhergesagt werden.

Tabelle 7: Ergebnisse der multiplen linearen Regression für das Erhebungsjahr 2018: Vorhersage der Abweichung vom Erwartungswert

Prädiktoren	Differenz vom Erwartungswert				
	<i>B</i>	<i>95CI</i>	<i>Beta</i>	<i>standardisierter CI</i>	<i>p</i>
(Intercept)	–28.03	[–29.23, –26.85]	.00	[.00, .00]	< .001
Vorhandensein von Ganztagsbetreuung	–2.49	[–3.31, –1.67]	–.01	[–.02, –.01]	< .001
Vorhandensein von Nachmittagsbetreuung	.94	[.34, 1.56]	.01	[.01, .02]	.015
Vorhandensein von Förderung schwacher Schülerinnen und Schüler in Mathematik	–1.14	[–1.88, –.41]	–.01	[–.02, –.01]	.014
Vorhandensein von Unterstützungspersonal	.21	[.15, .28]	.02	[.02, .03]	< .001
Vorhandensein von Qualitätsinstrumenten	.67	[.45, .89]	.02	[.02, .04]	< .001
Dummy-Variable 1 (mittlere Schulen)	1.17	[.73, 1.63]	.01	[.01, .02]	< .001
Dummy-Variable 2 (große Stadtschulen)	–.93	[–1.33, –.53]	–.01	[–.02, –.01]	< .001
Beobachtungen	2.482				
<i>R</i> ²	.002				

5 Diskussion

Um die Frage „Was macht gute Schulen aus?“ beantworten zu können, untersuchte dieser Beitrag fünf mögliche Einflussfaktoren auf die Effektivität von Schulen, insbesondere ob das Vorhandensein einer Ganztagsbetreuung, Nachmittagsbetreuung, die Förderung von schwachen Schülerinnen und Schülern in Mathematik, das ausreichende Vorhandensein von Unterstützungspersonal und die Verwendung von Qualitätsinstrumenten die Schulleistung oder die Abweichung vom Erwartungswert der Schulleistung vorhersagen. Explorativ wurde untersucht, ob sich diese Effekte für Gruppierungen von Schulen, eingeteilt nach Urbanisierungsgrad, Anzahl an Schülerinnen und Schülern und Anzahl an Schülerinnen und Schülern mit anderer Erstsprache als Deutsch unterscheiden. Zusätzlich wurden die Ergebnisse zwischen den beiden Erhebungszeitpunkten 2013 und 2018 deskriptiv verglichen.

Ein zentrales Ergebnis des Beitrags war, dass der Effekt der sehr geringen Varianzaufklärung der multiplen Regressionen zur Vorhersage der mittleren Schulleistung noch mehr

verschwand, als die Abweichung vom Erwartungswert vorhergesagt wurde. Dies unterstützt, dass der faire Vergleich dazu beiträgt, dass Schulen, wenn für Kompositionseffekte der Schülerinnen und Schüler kontrolliert wird, vergleichbarer werden, und deckt sich mit den in der Literatur gängigen Befunden (Sciffer, Perry & McConney, 2022). Indirekt verdeutlicht dieser Befund die starken sozialen Ungleichheiten, die kaum über Faktoren der Schuleffektivität auf Schulebene aufgefangen werden können. Befunde zu den längsschnittlichen Effekten hinsichtlich des Einflusses von Kompositionseffekten auf die Schulleistung sind jedoch nicht konsistent (Marks, 2024; Marsh, Pekrun & Dicke, 2023). In einer australischen Studie wurde bspw. gezeigt, dass die sozioökonomische Zusammensetzung der Schule einen zu vernachlässigenden Effekt auf die Leistung der Schülerinnen und Schüler hatte (Marks, 2024). Wohingegen in einer Studie in den USA die kompositorischen Effekte des durchschnittlichen sozioökonomischen Status der Schule signifikant positiv für das Bildungsniveau und die Bildungs- und Berufserwartungen waren (Marsh et al., 2023). Weitere Studien, die die Ebenen Schule und Schülerinnen und Schüler berücksichtigen und auf das österreichische Bildungssystem fokussieren, sind hier notwendig, um Klarheit für die genaue Aussagekraft der Prädiktoren zu erlangen.

Für die fünf oben genannten Prädiktoren wurden für keinen der beiden Erhebungszeitpunkte weder für die Vorhersage der mittleren Schulleistung noch für die Abweichung vom Erwartungsbereich nennenswerte Effekte gefunden. Einschränkend kann jedoch angemerkt werden, dass keiner der Prädiktoren tiefgehend durch eine Skala, sondern nur durch ein Item erhoben wurde. Zusätzlich ergeben sich Einschränkungen hinsichtlich der Kontextbefragung. Weitere in das Modell und die Ebene der Schule passende Items konnten nicht als Prädiktoren aufgenommen werden, da diese nur zu einem Erhebungszeitpunkt in den Kontextbefragungen verwendet wurden. Die Studie verdeutlicht die Herausforderungen bei der Messung von Schuleffektivität, insbesondere die Problematik einer inhaltsvaliden und über die Erhebungswellen stabile Erfassung der Kontextvariablen. Wenn Bildungsmonitoring das Ziel hat, Unterschiede in den Schulleistungen nicht nur deskriptiv zu erfassen und den Ist-Stand zu beschreiben, sondern auch – besonders mit dem Fokus auf die Verbesserung der Bildungspraxis – zu erklären sowie die Gründe, Ursachen und Wirkzusammenhänge des erhobenen Ist-Stands aufzuzeigen, sind Instrumente notwendig, die

1. sich an der internationalen Forschungslage zu den betreffenden Konstrukten orientieren, bspw. dem dynamischen Modell der Bildungseffektivität (Creemers & Kyriakides, 2010), und diese in angemessener Breite inhaltsvalide abdecken, und
2. möglichst stabil über alle Erhebungswellen angewendet werden.

Eine alternative Interpretation für die geringe Vorhersagekraft der Prädiktoren könnte in der Heterogenität der Schulen liegen. In Analogie zur personenzentrierten Methodologie (Magnusson & Stattin, 2007; Woo, Jebb, Tay & Parrigon, 2018) könnte angenommen werden, dass innerhalb der Heterogenität der Schulen in Österreich homogene Subgruppen von Schulen mit gegenläufigen Prädiktionsmustern existieren. Die theoretische und empirische Aufklärung dieser differenziellen Effekte, die sich in Summe zu einem Nulleffekt aggregieren würden, wäre eine wichtige Aufgabe der Bildungsforschung.

Für beide Erhebungszeitpunkte fanden sich Zusammenhänge für die Cluster der Hintergrundvariablen. Ohne Berücksichtigung der Erwartungswerte zeigte sich in beiden Erhebungsjahren, dass Schulen mit einer höheren Anzahl an Schülerinnen und Schülern, in städtischem Gebiet und mit einer höheren Anzahl an Schülerinnen und Schülern, deren Erstsprache nicht Deutsch ist, bessere mittlere Schulleistungen aufwiesen als Schulen mit der niedrigsten Anzahl an Schülerinnen und Schülern, in ländlichem Gebiet und mit der niedrigsten Anzahl an Schülerinnen und Schülern deren Erstsprache nicht Deutsch ist. Derselbe Effekt zeigte sich hinsichtlich des Vergleichs der Schulen, die sich aus einer mittleren Anzahl an Schülerinnen und Schülern, in durchschnittlich besiedeltem Gebiet und einer mittleren Anzahl an Schülerinnen und Schülern mit anderer Erstsprache als Deutsch zusammensetzen, im Vergleich zur Referenzgruppe der Schulen mit geringster Schülerinnen- und Schüleranzahl, in ländlichem Gebiet und der geringsten Anzahl an Schülerinnen und Schülern mit anderer Erstsprache als Deutsch. Unter Kontrolle des Erwartungswerts erwiesen sich diese Prädiktionen zu beiden Erhebungszeitpunkten nicht mehr als bedeutsam. Die Erwartung, dass Schulen mit bestimmten Merkmalszusammensetzungen „besser oder schlechter als erwartet“ abschneiden, zeigte sich somit nicht. Somit konnte ein indirekter Hinweis auf die Validität des Modells zur Berechnung der fairen Vergleiche geliefert werden, da auch die Kombinationen der einzelnen Faktoren keine zusätzliche Erklärung über den Erwartungswert hinaus liefern konnten.

Eine Limitation des vorliegenden Beitrags besteht darin, dass sich die Analyse ausschließlich auf die 4. Schulstufe als Abschlussklasse der Primarstufe und die Domäne „Mathematik“ bezog. Jedoch werden auch in internationalen Erhebungen, wie z. B. in PISA, nur Abschlussklassen berücksichtigt, um die Leistungsfähigkeit des Schulsystems zu untersuchen (Toferer, Lang & Salchegger, 2023). Eine weitere Limitation liegt der sehr kleinen Auswahl an Schuleffektivitätsfaktoren zugrunde. Zentrale Prädiktoren für gute Schulen, die auch im DMEE enthalten sind, wie z. B. die Lehrpersonen-Kooperation oder die Einbindung der Eltern (Kyriakides et al., 2018), konnten aufgrund der Datenlage nicht in die Analysen miteinbezogen werden.

Eine weitere Einschränkung dieses Beitrags ergibt sich aus dem Untersuchungsmodell, das die mittlere Schulleistung als abhängige Variable interpretiert. Im DMEE ist das Outcome einer Schule jedoch nicht nur auf die Leistung beschränkt, sondern bezieht auch emotionale und motivationale Faktoren mit ein, die neben der Schulleistung als kognitive Faktoren zum Schulerfolg beitragen (Hascher & Hagenauer, 2011). Zukünftige Untersuchungen sollten folglich Faktoren wie z. B. das Wohlbefinden der Schülerinnen und Schüler an einer Schule integrieren. Es wäre auch eine Anregung für die Weiterentwicklung der großen Schulleistungsstudien, stärker auf theoretisch abgesicherte Kontextfragebögen zu setzen.

Erkenntnisse dieses Beitrags basieren zudem auf Analysen von Sekundärdaten, die aus der BIST-Ü 2013 und BIST-Ü 2018 verfügbar und für die Untersuchung zugänglich waren. Da diese Untersuchung einer Fragestellung nachging, die nicht primär in der BIST-Ü angelegt war, und keine Reproduktion oder Kontrolle der Primärdaten darstellt, können die Erhebungen der Sekundäranalysen hinsichtlich des Studiendesigns und der Erhebungs-

instrumente von der Fragestellung abweichen und es können nur die Daten verwendet werden, die im Nachhinein verfügbar sind (Jansen, Kocaj & Stanat, 2022).

Dieser Beitrag ergänzt die Forschung durch die Berücksichtigung der zeitlichen Dynamik der Prädiktionsmuster von Schuleffektivitätsfaktoren und plädiert für eine ganzheitliche Validierung von Schuleffektivitätsmodellen. Es bleibt die offene Frage bezüglich der Generalisierbarkeit der Ergebnisse auf andere Kontexte, Domänen und Schulstufen. Eine weitere Forschungsfrage wäre unter anderem, ob sich die Ergebnisse für andere Datensätze der BIST-Ü replizieren lassen würden oder ob sich für die Sekundarstufe (v. a. unter Berücksichtigung der Unterscheidung zwischen allgemeinbildenden höheren Schulen und Neuen Mittelschulen) andere Ergebnisse finden würden. Zukünftige Forschungsarbeiten sollten sich auf die Entwicklung robusterer Messinstrumente und Methoden zur Erfassung der Schuleffektivität konzentrieren.

6 Fazit

Der vorliegende Beitrag verdeutlicht die Komplexität und die Herausforderungen bei der Untersuchung von Schuleffektivität und der Frage, was gute Schulen ausmacht. Die Analysen unterstreichen die Wichtigkeit des „fairen Vergleichs“ bei der Bewertung von Schulleistungen, da dieser es ermöglicht, Leistungsunterschiede zu identifizieren, die über die Erwartungen hinausgehen oder hinter diesen zurückbleiben. Die Ergebnisse weisen darauf hin, dass Schuleffektivität ein multidimensionales Phänomen ist, das eine ganzheitliche und gegenstandsadäquate Betrachtungsweise erfordert. Die in dieser Studie untersuchten Schuleffektivitätsfaktoren hatten wenig Einfluss auf die Vorhersage von Schulleistungen, was die Notwendigkeit weiterer Forschung – insbesondere in Österreich – unterstreicht, um diese Befunde zu verstehen und zu kontextualisieren. Zukünftige Studien sollten darauf abzielen, diese Beziehungen weiter zu erforschen, insbesondere durch die Verwendung robusterer und vielfältigerer Messinstrumente, die es ermöglichen, die unterschiedlichen Facetten der Schuleffektivität an den internationalen theoretischen wie empirischen Forschungsstand anknüpfbar genauer zu erfassen und zu analysieren. Darüber hinaus könnte der Einbezug weiterer Bildungsstufen und fachlicher Domänen dazu beitragen, die Generalisierbarkeit der Ergebnisse zu erhöhen und ein vollständigeres Bild der Faktoren zu erhalten, die zur Schuleffektivität beitragen.

Literatur

- Altrichter, H., Kemethofer, D. & George, A. C. (2019). Schulleitungen und evidenzbasierte Bildungsreform im Schulwesen. *Zeitschrift für Bildungsforschung*, 9 (1), 17–35. doi:10.1007/s35834-018-0228-5
- Aslantas, I. (2020). Impact of contextual predictors on value-added teacher effectiveness estimates. *Education Sciences*, 10 (12), 1–16. doi:10.3390/educsci10120390
- Beutel, S.-I., Höhmann, K., Schratz, M. & Pant, H. A. (2016). *Handbuch Gute Schule: Sechs Qualitätsbereiche für eine zukunftsweisende Praxis*. Seelze: Friedrich Verlag.
- BIFIE. (2019). *Bundesergebnisbericht. Standardüberprüfung 2018. Mathematik, 4. Schulstufe*. Salzburg: BIFIE.
- Bischof, L. M. (2017). Schulentwicklung und Schuleffektivität. Ihre theoretische und empirische Verknüpfung. Wiesbaden: Springer Fachmedien. doi:10.1007/978-3-658-14628-3
- Bollen, R. (1996). School effectiveness and school improvement. The intellectual and policy context. In D. Reynolds, R. Bollen, B. P. M. Creemers, D. Hopkins, N. Stoll & N. Lagerweij (Hrsg.), *Making good schools: Linking school effectiveness and school improvement* (S. 1–20). London: Routledge. doi:10.4324/9780203975916
- Börner, N. (2011). Ganztagschule und ihre Auswirkungen auf Familie. *Zeitschrift für Erziehungswissenschaft*, 14 (3), 221–236. doi:10.1007/s11618-011-0237-9
- Bos, W., Postlethwaite, T. N. & Gebauer, M. M. (2010). Potenziale, Grenzen und Perspektiven internationaler Schulleistungsforschung. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch Bildungsforschung* (S. 275–295). Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/978-3-531-92015-3
- Brühwiler, C., Helmke, A. & Schrader, F. W. (2017). Determinanten der Schulleistung. In M. K. W. Schweer (Hrsg.), *Lehrer-Schüler-Interaktion: Inhaltsfelder, Forschungsperspektiven und methodische Zugänge* (S. 291–314). Wiesbaden: Springer Fachmedien. doi:10.1007/978-3-658-15083-9
- Bruneforth, M., Weber, C. & Bacher, J. (2012). Chancengerechtigkeit und Mehrsprachigkeit. Chancengleichheit und garantiertes Bildungsminimum in Österreich. In B. Herzog-Punzenberger (Hrsg.), *Nationaler Bildungsbericht 2012. Band 2. Fokussierte Analysen bildungspolitischer Schwerpunktthemen* (S. 206–218). Graz: Leykam.
- Creemers, B. P. M. & Kyriakides, L. (2010). Using the Dynamic Model to develop an evidence-based and theory-driven approach to school improvement. *Irish Educational Studies*, 29 (1), 5–23. doi:10.1080/03323310903522669
- Creemers, B. P. M., Kyriakides, L. & Antoniou, P. (2012). Towards the development of a dynamic approach to teacher professional development. In B. P. M. Creemers, L. Kyriakides & P. Antoniou (Hrsg.), *Teacher Professional Development for Improving Quality of Teaching* (S. 3–11). New York: Dodrecht. doi:10.1007/978-94-007-5207-8
- Creemers, B. P. M., Peters, T. & Reynolds, D. (1989). *School effectiveness and school improvement*. Amsterdam: Swets & Zeitlinger.
- Day, C., Sammons, P., Hopkins, D., Harris, A., Leithwood, K., Qing, G. et al. (2009). *The impact of school leadership on pupil outcomes. Final report*. University of Nottingham.

- Dumont, H., Neumann, M., Maaz, K. & Trautwein, U. (2013). Die Zusammensetzung der Schülerschaft als Einflussfaktor für Schulleistungen. *Psychologie in Erziehung und Unterricht*, 60 (3), 163–183. doi:10.2378/peu2013.art14d
- Fiege, C., Reuther, F. & Nachtigall, C. (2011). Faire Vergleiche? – Berücksichtigung von Kontextbedingungen des Lernens beim Vergleich von Testergebnissen aus deutschen Vergleichsarbeiten. *Zeitschrift für Bildungsforschung*, 2, 133–149. doi:10.1007/s35834-011-0009-x
- George, A. C. & Robitzsch, A. (2013). „Fairer Vergleich“. *Technische Dokumentation – BIST-Ü Mathematik, 4. Schulstufe, 2013*. Salzburg: BIFIE.
- George, A. C., Schreiner, C., Wiesner, C., Pointinger, M. & Pacher, K. (2019). *Fünf Jahre flächendeckende Bildungsstandardüberprüfungen in Österreich. Vertiefende Analysen zum Zyklus 2012 bis 2016*. Münster: Waxmann.
- Greenberg, M. T., Weissberg, R. P., O'Brien, M. U., Zins, J. E., Fredericks, L. & Resnik, H. (2003). Enhancing school-based prevention and youth development through coordinated social, emotional, and academic learning. *American Psychologist*, 58 (6/7), 466–474. doi:10.1037/0003-066X.58.6-7.466
- Groß Ophoff, J. (2012). *Lernstandserhebungen: Reflexionen und Nutzung* (Bd. 85). Münster: Waxmann.
- Gülens, F. U. & Çelik, Ö. (2021). Secondary School Teachers' Effective School Perception: The Role of School Culture and Teacher Empowerment. *International Journal of Progressive Education*, 17 (5), 332–344.
- Hascher, T. & Hagenauer, G. (2011). Wohlbefinden und Emotionen in der Schule als zentrale Elemente des Schulerfolgs unter der Perspektive geschlechtsspezifischer Ungleichheiten. In A. Hadjar (Hrsg.), *Geschlechtsspezifische Bildungsungleichheiten. Systematischer Überblick zur Frage der Bildungsungleichheit zwischen den Geschlechtern* (S. 285–308). Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/978-3-531-92779-4
- Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge. doi:10.4324/9780203887332
- Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen. (2023). *Archiv des BIFIE – Materialien*. Verfügbar unter: <https://iqs.gv.at/downloads/archiv-des-bifie>
- Isaac, K. & Hosenfeld, I. (2008). Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften. In R. Arnold & C. Grieser (Hrsg.), *Schulleitung und Schulentwicklung* (S. 119–144). Baltmannsweiler: Schneider.
- Jansen, M., Kocaj, A. & Stanat, P. (2022). Sekundäranalysen in der Schulforschung. In T. Hascher, T.-S. Idel & W. Helsper (Hrsg.), *Handbuch Schulforschung* (S. 171–194). Wiesbaden: Springer Fachmedien. doi:10.1007/978-3-658-24729-4
- Javornik, Š. & Klemenčič Mirazchiyski, E. (2023). Factors Contributing to School Effectiveness: A Systematic Literature Review. *European Journal of Investigation in Health, Psychology and Education*, 13 (10), 2095–2111.
- Kazemi, A. (2016). Examining the Interplay of Justice Perceptions, Motivation, and School Achievement among Secondary School Students. *Social Justice Research*, 29, 103–118. doi:10.1007/s11211-016-0261-2
- Khan, N. (2019). The Impact of Organizational Climate on Teachers Commitment. *Journal of Education and Educational Development*, 6 (2), 327– 342.

- Kiel, E. (2017). Unterrichtsforschung im Kontext der empirischen Bildungsforschung. In R. Tippelt & B. Schmidt-Hertha (Hrsg.), *Handbuch Bildungsforschung* (S. 1–22). Wiesbaden: Springer. doi:10.1007/978-3-531-92015-3
- Klem, A. M. & Connell, J. P. (2014). Relationships matter: Linking teacher support to student engagement and achievement. *Journal of school health*, 74, 262–273. doi:10.1111/j.1746-1561.2004.tb08283.x
- KMK (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. Bonn: LinkLuchterhand.
- Kyriakides, L. & Creemers, B. P. M. (2008). Using a multidimensional approach to measure the impact of classroom-level factors upon student achievement: A study testing the validity of the dynamic model. *School Effectiveness and School Improvement*, 19 (2), 183–205. doi:10.1080/09243450802047873
- Kyriakides, L., Creemers, B. P. M. & Charalambous, E. (2018). *Equity and Quality Dimensions in Educational Effectiveness*. Cham: Springer International Publishing. doi:10.1007/978-3-319-72066-1
- Levine, D. U. & Lezotte, L. W. (1990). Unusually Effective Schools: A Review and Analysis of Research and Practice. *School Effectiveness and School Improvement*, 1 (3), 221–224. doi:10.1080/0924345900010305
- Lindorff, A., Sammons, P. & Hall, J. (2020). International Perspectives in Educational Effectiveness Research: A Historical Overview. In J. Hall, A. Lindorff & P. Sammons (Hrsg.), *International Perspectives in Educational Effectiveness Research* (S. 9–31). Cham: Springer Nature Switzerland. doi:10.1007/978-3-030-44810-3_14
- Liu, J., Peng, P., Zhao, B. & Lui, L. (2022). Socioeconomic Status and Academic Achievement in Primary and Secondary Education: a Meta-analytic Review. *Educational Psychology Review*, 34, 2867–2896. doi:10.1007/s10648-022-09689-y
- Maag Merki, K. (2016). Theoretische und empirische Analysen der Effektivität von Bildungsstandards, standardbezogenen Lernstandserhebungen und zentralen Abschlussprüfungen. In H. Altrichter & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (S. 151–182). Wiesbaden: Springer Fachmedien. doi:10.1007/978-3-531-18942-0
- Magnusson, D. & Stattin, H. (2007). The person in context: A holistic-interactionistic approach. In W. Damon & R. M. Lerner (Hrsg.), *Handbook of Child Psychology. Volume 1: Theoretical Models of Human Development*. New Jersey: John Wiley & Sons, Inc. doi:10.1002/9780470147658.chpsy0108
- Makharova, G., Nurzhanova, S., Adilbayeva, U., Dossanova, A. & Aimagambetova, M. (2022). The Level of Efficiency of Using Digital Resources for Developing Primary School Students' Linguodidactic Potential. *World Journal on Educational Technology: Current Issues*, 14 (1), 217–230. doi:10.18844/wjet.v14i1.6719
- Marks, G. N. (2024). No substantive effects of school socioeconomic composition on student achievement in Australia: a response to Sciffer, Perry and McConney. *Large-scale Assessments in Education*, 12 (1), 8. doi:10.1186/s40536-024-00196-w
- Marsh, H. W., Pekrun, R. & Dicke, T. (2023). Disentangling the long-term compositional effects of school-average achievement and SES: a substantive-methodological synergy. *Educational Psychology Review*, 35 (70), 1–36. doi:10.1007/s10648-023-09726-4

- Massenkeil, J. & Rothland, M. (2016). Kollegiale Kooperation im Lehrerberuf. Überblick und Systematisierung aktueller Forschung. *Schulpädagogik heute*, 7 (13), 1–28.
- McCracken, J. D. & Barcinas, J. D. T. (1991). Differences between rural and urban schools, student characteristics and student aspirations in Ohio. *Journal of Research in Rural Education*, 7 (2), 29–40.
- Opdenakker, M.-C. & Van Damme, J. (2000). Effects of schools, teaching staff and classes on achievement and well-being in secondary education: Similarities and differences between school outcomes. *School Effectiveness and School Improvement*, 11 (2), 165–196. doi:10.1076/0924-3453(200006)11:2;1-Q;FT165
- Özgenel, M. & Karsantik, I. (2020). Effects of school principals' leadership styles on leadership practices. *Malaysian Online Journal of Educational Sciences*, 8 (2), 1–13.
- Pham, G., Freunberger, R., Robitzsch, A. & Itzlinger-Bruneforth, U. (2016). Reliabilität und Stabilität des Index der sozialen Benachteiligung und Kompositionseffekt der Schulen. *Zeitschrift für Bildungsforschung*, 6, 345–364. doi:10.1007/s35834-016-0164-1
- Pham, G. & Robitzsch, A. (2014). „Fairer Vergleich“. *Technische Dokumentation – BIST-Ü Englisch*, 8. Schulstufe, 2013. Salzburg: BIFIE.
- R Core Team. (2023). *R: A language and environment for statistical computing*. Wien: R Foundation for Statistical Computing.
- Reynolds, D., Chapman, C., Kelly, A., Muijs, D. & Sammons, P. (2011). Educational effectiveness: The development of the discipline, the critiques, the defence, and the present debate. *Effective Education*, 3 (2), 109–127. doi:10.1080/19415532.2011.686168
- Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T. & Teddlie, C. (2014). Educational effectiveness research (EER): a state-of-the-art review. *School Effectiveness and School Improvement*, 25 (2), 197–230. doi:10.1080/09243453.2014.885450
- Sammons, P., Thomas, S. M. & Mortimore, P. (1997). *Forging links. Effective schools and effective departments*. London: Paul Chapman Publishing.
- Scheerens, J. (1990). School effectiveness research and the development of process indicators of school functioning. *School Effectiveness and School Improvement*, 1 (1), 61–80. doi:10.1080/0924345900010106
- Schmid, R. (2018). Pockets of excellence: Teacher beliefs and behaviors that lead to high student achievement at low achieving schools. *Sage Open*, 8 (3), 1–10. doi:10.1177/2158244018797238
- Schratz, M., Hartmann, M. & Schley, W. (2010). *Schule wirksam leiten. Analyse innovativer Führung in der Praxis*. Münster: Waxmann.
- Schreiner, C. & Breit, S. (2014). *Standardüberprüfung 2013 Mathematik, 4. Schulstufe. Bundesergebnisbericht*. Salzburg: BIFIE.
- Sciffer, M. G., Perry, L. B. & McConney, A. (2022). The substantiveness of socioeconomic school compositional effects in Australia: measurement error and the relationship with academic composition. *Large-scale Assessments in Education*, 10, 1–22. doi:10.1186/s40536-022-00142-8
- Semper, I., Mende, L. & Berkemeyer, N. (2017). Schul- und Unterrichtsforschung. In T. Burger & N. Miceli (Hrsg.), *Empirische Forschung im Kontext Schule* (S. 31–48). Wiesbaden: Springer. doi:10.1007/978-3-658-15437-0

- Stewart, L. (2009). Achievement differences between large and small schools in Texas. *The Rural Educator*, 30 (2), 20–28. doi:10.35608/ruraled.v30i2.450
- Stoll, L. & Fink, D. (1996). *Changing our schools: Linking school effectiveness and school improvement*. Buckingham: Open University Press.
- Teddlie, C. & Reynolds, D. (2000). *The international handbook of school effectiveness research*. London: Falmer Press. doi:10.4324/9780203454404
- Tiedemann, J. & Billmann-Mahecha, E. (2004). Kontextfaktoren der Schulleistung im Grundschulalter. Ergebnisse aus der Hannoverschen Grundschulstudie. *Zeitschrift für Pädagogische Psychologie*, 18 (2), 113–124. doi:10.1024/1010-0652.18.2.113
- Timmermans, A. C. & Thomas, S. M. (2015). The impact of student composition on schools' value-added performance. *School Effectiveness and School Improvement*, 26 (3), 487–498. doi:10.1080/09243453.2014.957328
- Toferer, B., Lang, B. & Salchegger, S. (2023). *PISA 2022. Kompetenzen in Mathematik, Lesen und Naturwissenschaft am Ende der Pflichtschulzeit im internationalen Vergleich*. Salzburg: Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen.
- Vogtenhuber, S., Lassnigg, L., Bruneforth, M., Edelhofer-Lielacher, E. & Siegle, T. (2015). B: Inputs – Personelle und finanzielle Ressourcen. In M. Bruneforth, L. Lassnigg, S. Vogtenhuber, C. Schreiner & S. Breit (Hrsg.), *Nationaler Bildungsbericht Österreich 2015, Band 1: Das Schulsystem im Spiegel von Daten und Indikatoren* (S. 37–70). Graz: Leykam. doi:10.17888/nbb2015-1-B.2
- Wang, J. & Wu, Y. (2023). Income inequality, cultural capital, and high school students' academic achievement in OECD countries: A moderated mediation analysis. *The British Journal of Sociology*, 74, 148–172. doi:10.1111/1468-4446.12997
- Woo, S. E., Jebb, A. T., Tay, L. & Parrigon, S. (2018). Putting the „person“ in the center: Review and synthesis of person-centered approaches and methods in organizational science. *Organizational Research Methods*, 21 (4), 814–845. doi:10.1177/1094428117752467
- Zou, B., Huang, L., Ma, W. & Qiu, Y. (2021). Evaluation of the effectiveness of EFL online teaching during the COVID-19 pandemic. *Sage Open*, 11 (4), 1–17. doi:10.1177/21582440211054491

Anhang

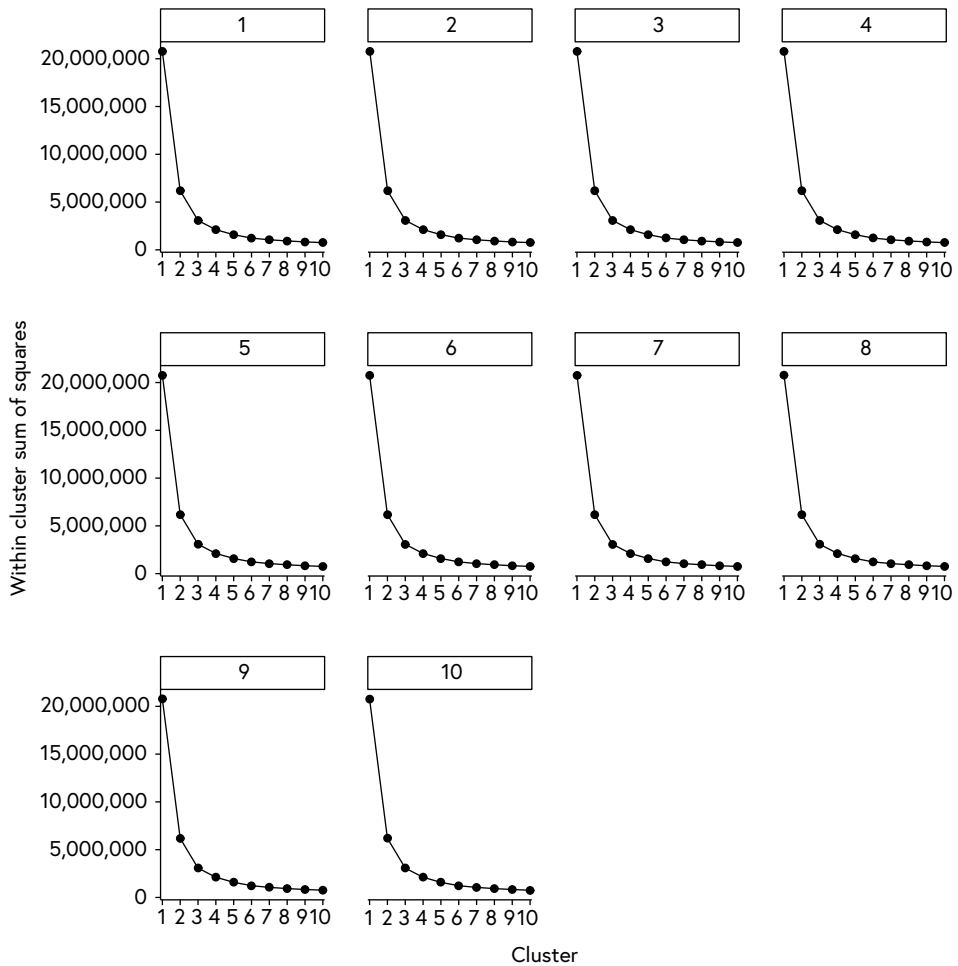


Abbildung 2: Elbow Plots der K-Means-Clusteranalyse für das Erhebungsjahr 2013.

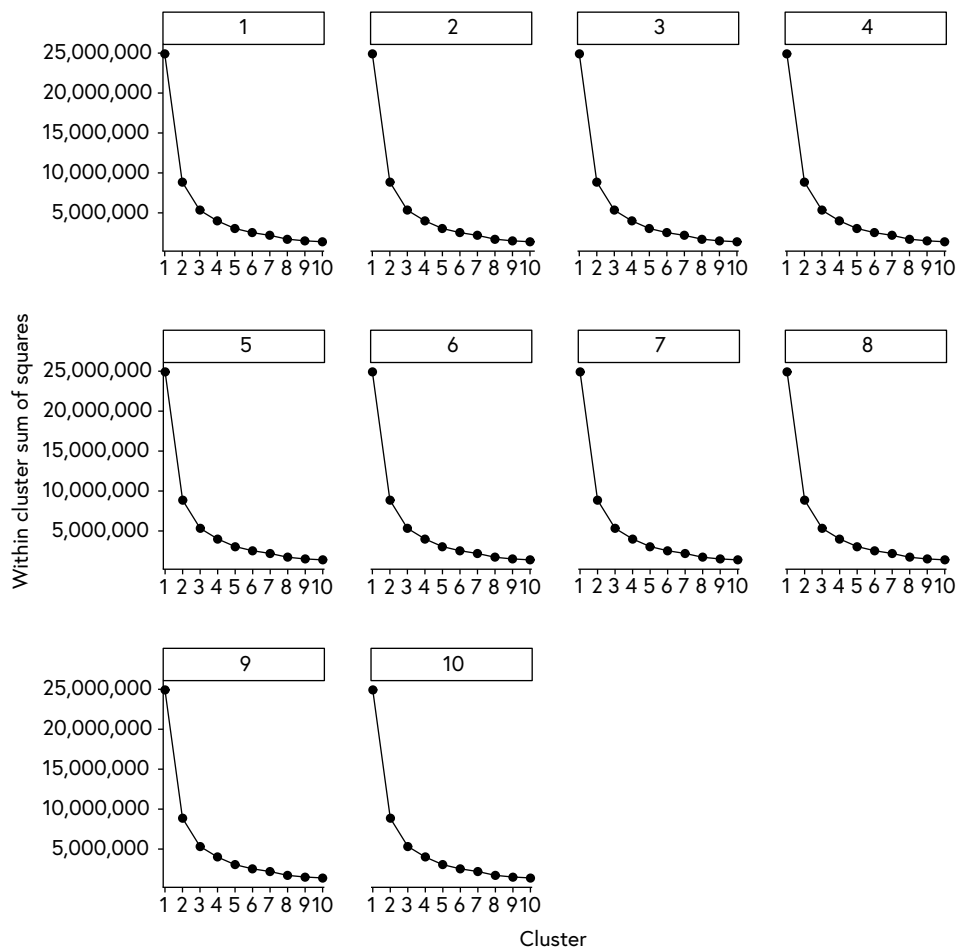


Abbildung 3: Elbow Plots der K-Means-Clusteranalyse für das Erhebungsjahr 2018.

Tabelle 8: Ebenen des dynamischen Modells, Items, Antwortkategorien und Rekodierungen der BIST-Ü 2013 und 2018 Kontextbefragungen.

Items			
2013	2018	Antwortkategorien	Rekodierung
In welchem Ausmaß gibt es die folgenden Förderangebote in Mathematik in diesem Schuljahr in den getesteten 4. Klassen? Förderangebot (Förderkurse) für lernschwache Schülerinnen/Schüler in Mathematik	Gibt es in diesem Schuljahr Förderunterricht für lernschwache Schülerinnen/Schüler der 4. Klasse in Mathematik?	2013: Anzahl der wöchentlichen Förderstunden an der jeweiligen Schule 2018: 1: ja, 2: nein	2013: Anzahl wöchentlicher Förderstunden > 0: 1 (vorhanden) Anzahl wöchentlicher Förderstunden = 0: 0 (nicht vorhanden) 2018: 1: 1 (vorhanden) 2: 0 (nicht vorhanden)
Über welche der folgenden Instrumente zur Qualitätssicherung und -entwicklung verfügen Sie an Ihrer Schule?	Welche der folgenden Instrumente zur Qualitätssicherung und -entwicklung fanden im Schuljahr 2017/18 an Ihrer Schule Anwendung?	Aufzählung von sieben Instrumenten; 0: nicht angekreuzt; 1: angekreuzt	Summe über angekreuzte/vorhandene Instrumente
Werden die Lehrerinnen/Lehrer an Ihrer Schule im laufenden Schuljahr unterstützt durch ...?		Aufzählung von verschiedenen Förderpersonen, pro Kategorie: 2013: 1: ja, ausreichend; 2: ja, jedoch zu wenig, nicht dem Bedarf entsprechend; 3: nein, an unserer Schule besteht kein Bedarf; 4: nein, aber an unserer Schule bestünde Bedarf 2018: 1: ja, ausreichend; 2: ja, jedoch zu wenig, nicht dem Bedarf entsprechend; 3: nein, aber an unserer Schule bestünde Bedarf; 4: nein, an unserer Schule besteht kein Bedarf	2013: 1: ausreichend vorhanden (1) 2: nicht ausreichend vorhanden (0) 3: kein Bedarf (1) 4: nicht ausreichend vorhanden (0) 2018: 1: ausreichend vorhanden (1) 2: nicht ausreichend vorhanden (0) 3: nicht ausreichend vorhanden (0) 4: kein Bedarf (1) Summe über alle Variablen
Gibt es an Ihrer Schule ein Ganztagsangebot in verschränkter Form, d. h. mit verteiltem Unterricht auf den ganzen Tag?		1: ja, 2: nein	1: 1 (vorhanden), 2: 0 (nicht vorhanden)
Gibt es an Ihrer Schule eine Nachmittags-/Tagesbetreuung (NABE/TABE)?		1: ja, 2: nein	1: 1 (vorhanden), 2: 0 (nicht vorhanden)

Jana Groß Ophoff, David Kemethofer, Christoph Helm, Ramona Zintl, Katharina Meusbürger, Martina Ott

Maßnahmen der Qualitätssicherung und -entwicklung als „Motor“ schulischen Lernens? Befunde zur Leistungsentwicklung

1 Einleitung

Die Diskussion um Qualität, Leistungsfähigkeit und ungleiche Bildungschancen im Schulsystem ist per se kein neues Phänomen und lässt sich in Form unterschiedlicher Phasen seit den 1960er-Jahren nachzeichnen (Eder & Altrichter, 2009; Specht, 2006). Seit Mitte der 1990er-Jahre und insbesondere im Zuge internationaler Vergleichsstudien von Schülerinnen- und Schülerleistungen intensivierte sich die Qualitätsdebatte im Bildungswesen (z. B. Klieme & Tippelt, 2008). Infolge der Ergebnisse aus dem *Programme for International Student Assessment* (PISA) für das Jahr 2003, wonach Schülerinnen und Schüler in Lesen, Mathematik und Naturwissenschaften deutlich schlechter in ihren Leistungen abgeschnitten hatten als noch drei Jahre zuvor, wurde im österreichischen Bildungswesen verstärkt auf eine neue Steuerungslogik gesetzt. An die Stelle einer bis dato dominierenden Orientierung am Input, z. B. in Form traditioneller Lehrpläne, rückte eine Orientierung an den Outputs von Bildung, also den durch Schule und Unterricht erzielten Ergebnissen (Ditton, 2017). Die damit verbundene Steuerungslogik räumt Schulen mehr Autonomie in der Schulprogrammarbeit und bei Entwicklungsprozessen ein. Die Dezentralisierung von Entscheidungskompetenzen bedeutete konkret, dass mehr Handlungsspielräume auf Schulebene in pädagogischen, organisatorischen und – je nach Schulerhalter – personellen Angelegenheiten eingeräumt wurden. Zugleich wurden der Dezentralisierung von Entscheidungskompetenzen verschiedene Bildungsmonitoring-Maßnahmen gegenübergestellt, mit denen die unmittelbaren und längerfristigen Effekte von Schule und Unterricht regelmäßig überprüft werden.

Dementsprechend definieren Döbert und Weishaupt (2012) Bildungsmonitoring als „kontinuierlichen, datengestützten Beobachtungs- und Analyseprozess des Bildungswesens“ mit dem Ziel, „Bildungspolitik und Öffentlichkeit über Rahmenbedingungen, Verlaufsmerkmale, Ergebnisse und Erträge von Bildungsprozessen“ zu informieren, was wiederum als Grundlage für „Zieldiskussionen und politische Entscheidungen“ (S. 158) dienen kann. Wesentlich ist hierbei, dass qualitätsrelevante Informationen mithilfe wissenschaftlicher

Methoden erhoben, analysiert und genutzt werden (Grünkorn, Klieme & Stanat, 2019). Das Bildungsmonitoring im deutschsprachigen Raum folgt entsprechend der Annahme, dass sich durch die Bereitstellung von Evidenz gezielt Stellschrauben zum Erreichen von definierten Effektivitätskriterien (Output) identifizieren lassen. Werden kompetenzorientierte Bildungsstandards als Maßstab herangezogen, können – so die Erwartung – auch Hinweise darauf gewonnen werden, welche Input- oder Prozessfaktoren angepasst werden sollten (Eder & Altrichter, 2009; Klieme et al., 2003). Von Evidenzen spricht man in diesem Zusammenhang, wenn aus der Auseinandersetzung mit Daten und Informationen Wissen generiert wird (Schratz et al., 2019); Outputkriterien sind in der Regel fachliche Leistungen von Schülerinnen und Schülern aber auch Einstellungen und Haltungen (Ditton, 2017). Hierzu werden verbindliche Ziele definiert, kommuniziert und deren Erreichen mittels verschiedenster Instrumente überprüft und bewertet. Die entsprechende Evidenz wird wiederum ins System zurückgespielt und soll auf den verschiedenen Systemebenen Entwicklungsaktivitäten stimulieren (Altrichter & Maag Merki, 2016). Vonseiten des Bundesministeriums für Bildung, Wissenschaft und Forschung (2021, S. 28) heißt es dazu, dass Schulen am Bildungserfolg ihrer Schülerinnen und Schüler unter Nutzung der möglichen Freiräume beurteilt werden. Das Zusammenspiel von einerseits zentral vorgegebenen Zielen und staatlichen Kontrollmechanismen und andererseits mehr Autonomie für Einzelschulen hat somit laut Fend (2011, S. 8) zu einer „neuen Grammatik“ der Qualitätssicherung und -entwicklung (= QSE) geführt.

Dieser Beitrag untersucht den Zusammenhang zwischen gängigen QSE-Maßnahmen an Schulen und den Leistungen der Schülerinnen und Schüler. Grundlage sind Daten aus den österreichischen Bildungsstandardüberprüfungen (BIST-Ü) im Sekundarbereich (8. Jahrgangsstufe) – Mathematik (2012, 2017) und Englisch (2013, 2019)¹. Neben Schülerinnen und Schülerleistungen erfassen die BIST-Ü-Begleitbefragungen der Schulleitungen² auch QSE-Maßnahmen an den Schulen. Wie schon Kemethofer et al. (2015) formulieren, ist auch unser Analyseverfahren „von der Annahme getragen, dass Schülerinnen und Schüler in [reform- und entwicklungs-]aktiveren Schulen [...] eine förderlichere Lernumgebung wahrnehmen und folglich bessere Leistungen erbringen“ (S. 27). Während die Analysen von Kemethofer et al. jedoch ausschließlich auf der 2012 durchgeführten BIST-Ü im Fach Mathematik aufbauen (und zwar aus einer querschnittlichen Perspektive), liegt der Mehrwert des vorliegenden Beitrags in der längsschnittlichen Auswertung des mittlerweile deutlich umfangreicheren Datenmaterials. Der Beitrag gibt zunächst einen Überblick über QSE-Maßnahmen im österreichischen Bildungswesen vor dem Hintergrund bildungspolitischer Reform- und Entwicklungsaktivitäten im Zuge der sog. „neuen Steuerung“ (Altrichter & Maag Merki, 2016). Im Anschluss an die für diesen Beitrag formulierten Forschungsfragen werden die methodischen Grundlagen und Ergebnisse berichtet, die abschließend in den Stand der Forschung eingeordnet und diskutiert werden.

-
- 1 Außerdem wurden 2016 die Bildungsstandards auf der 8. Schulstufe in Deutsch überprüft. Da in Deutsch jedoch nur eine Erhebung stattfand, werden die Daten der Deutschüberprüfung in dieser Studie nicht berücksichtigt.
 - 2 In den vorgestellten Analysen wurden die Leistungen der Schülerinnen und Schüler auf Schulebene aggregiert, was treffender als Schulleistung zu bezeichnen wäre. Zum besseren Verständnis verwenden wir aber in diesem Beitrag die gängigere Bezeichnung Schülerinnen- und Schülerleistung.

2 Qualitätssicherung und -entwicklung im österreichischen Bildungswesen

Die in Österreich in den letzten Jahren implementierten bildungspolitischen Maßnahmen (u. a. BMBWF, 2021), zielen darauf ab, ...

- das Leistungs- und Bildungsniveau der Schülerinnen und Schüler (und von Zielgruppen in der Erwachsenenbildung) zu erhöhen,
- für mehr Chancen- und Geschlechtergerechtigkeit zu sorgen,
- die Schulorganisation und die Bildungsverwaltung effizienter und effektiver zu gestalten und
- die Bildungs- und Berufschancen durch Förderung der Sprachkompetenz in der Unterrichtssprache Deutsch zu verbessern.

Die zahlreichen zu verzeichnenden Reform- und Entwicklungsaktivitäten sind Ausdruck ebendieser bildungspolitischen Bemühungen (Groß Ophoff & Kemethofer, 2023; Jesacher-Roessler & Kemethofer, 2022). In einer Analyse für den Nationalen Bildungsbericht 2018 klassifizierten Schober et al. (2019) diese Reformmaßnahmen in fachbezogene, überfachliche und grundlegende strukturelle Reformen. Während *fachbezogene Reformen* konkrete und messbare Lerninhalte definieren, die von den Schülerinnen und Schülern erreicht werden sollen, beziehen sich *überfachliche Reformen* auf eine „grundlegende Neuorganisation des Lernens und Lehrens“ (ebd., S. 465). *Strukturelle Reformen* betreffen schließlich organisatorische und pädagogische Veränderungen des Schulsystems bzw. der Lehrerinnen- und Lehrerbildung.

2.1 Ausgewählte Reformen im österreichischen Bildungswesen

Eine der bedeutendsten Reformen im österreichischen Bildungswesen konzentrierte sich auf den (Fach-)Unterricht (Schober et al., 2019): Unter dem Schlagwort Kompetenzorientierung sollte sich die Unterrichtskultur als Ganzes verändern. Anstatt der Vermittlung von Faktenwissen stehen nunmehr die „Fähigkeit zur flexiblen Anwendung des erworbenen Wissens und der erworbenen Fertigkeiten sowie die Anschlussfähigkeit an den weiteren Bildungsverlauf“ im Fokus (Kulmhofer-Bommer & Diekmann, 2021, S. 426). Anders ausgedrückt, Kompetenzen beschreiben „die bei Schülerinnen und Schülern verfügbaren oder von ihnen erlernbaren kognitiven Fähigkeiten und Fertigkeiten, die notwendig sind, um bestimmte Probleme bzw. Aufgaben lösen zu können“ (Köller, 2010, S. 530). Kompetenzen beschreiben also erwünschte Lernergebnisse und konkretisieren auf diese Weise den Bildungsauftrag von Schulen. Der Einzug der Kompetenzorientierung in den Schulalltag erfolgt in Österreich 2009 durch die Entwicklung und Implementierung von fachbezogenen Bildungsstandards, die konkrete Lernergebnisse definieren und beschreiben, was Schülerinnen und Schüler zu bestimmten Zeitpunkten in ihrer Schullaufbahn können sollen. Auf der Sekundarstufe wurden Bildungsstandards für Deutsch, Mathematik und Englisch für die achte Schulstufe definiert. In konkrete Testaufgaben übersetzt kann die Erreichung oder Nichterreichung

von Bildungsstandards gemessen werden und Auskunft zum Kompetenzerwerb einzelner Schülerinnen und Schüler, ganzer Klassen oder auch Schulen liefern. Dadurch stellen Bildungsstandards im Zusammenspiel mit regelmäßigen Überprüfungen nicht nur einen aussagekräftigen Bezugspunkt für die Beurteilung von Schulqualität dar, sondern ermöglicht(en) ferner eine veränderte Form der QSE (Maaz, Emmrich, Kropf & Gärtner, 2019). So wird in den BIST-Ü zwischen 2009 und 2019 österreichweit (also flächendeckend) und alternierend in den Fächern überprüft, „ob und in welchem Ausmaß Schülerinnen und Schüler die gewünschten Kompetenzen erreichen und wie sich der Ist-Stand der Kompetenzen der Schülerinnen und Schüler mit dem angestrebten Soll (Zielorientierung) deckt“ (BIFIE, 2020, S. 11). Um darüber Schulentwicklungs- und Professionalisierungsprozesse anzuregen, werden die Ergebnisse als erreichte Kompetenzen auch an die Schulen zurückgemeldet. Diesem Vorgehen liegt wie gesagt die Erwartung zugrunde, dass über die Bereitstellung solcher nach wissenschaftlichen Standards erfasster und ausgewerteter Evidenz Prozesse in Schule und Unterricht angeregt werden, die zur Sicherung bzw. Entwicklung von Qualität beitragen können – wenngleich davon auszugehen ist, dass diese von Lehrpersonen eher als (Roh-)Daten aufgefasst werden, die ein Sensemaking im Kontext der eigenen Schule erfordern (Groß Ophoff, Brown & Helm, 2023).

Der Gedanke der Kompetenzorientierung spiegelt sich auch in den Reformen zur Professionalisierung des pädagogischen Personals wider. Einerseits stellt kompetenzorientiertes Unterrichten veränderte Ansprüche an Lehrpersonen, andererseits gilt es, das pädagogische Personal auf insgesamt neue Anforderungen und Herausforderungen vorzubereiten. In der praktischen Umsetzung führte dies zu strukturellen Reformen (i. S. von Schober et al., 2019), u. a. in der Lehramtsausbildung, zu gesetzlichen Änderungen im Bereich der Fort- und Weiterbildung und zu neuen Aufgabenfeldern und Verantwortungsbereichen für Schulleitungen inkl. reformiertem Qualifizierungsprozess (Groß Ophoff & Kemethofer, 2023). Das zugrunde liegende Konzept der „PädagogInnenbildung NEU“ (QSR, 2014) betrifft alle Phasen der Lehrerinnen- und Lehrerbildung, wonach speziell durch die Dissemination von Evidenz und der damit verbundenen wissenschaftlichen Ausrichtung der Aus-, Fort- und Weiterbildung sowohl bei angehenden wie auch praktizierenden Lehrpersonen eine forschende und vertieft reflektierende Haltung unterstützt werden (BMUKK & BMWF, 2010). Im Zuge der pädagogischen Ausbildung sollen Pädagoginnen und Pädagogen allgemeine pädagogische Kompetenz, fachliche und didaktische Kompetenz, soziale Kompetenz, Gender- und Diversitätskompetenz sowie ein sich selbst reflektierendes Professionsverständnis kontinuierlich weiterentwickeln (QSR, 2014). Durch kontinuierliche Fort- und Weiterbildung soll sichergestellt werden, dass Lehrpersonen den sich stetig ändernden Anforderungen im Lehrberuf nachkommen und so zur systematischen und kompetenzorientierten Qualitätssicherung bzw. -entwicklung beitragen können. Dies erfordert u. a., dass Lehrpersonen Evidenzen und Qualitätsentwicklungsverfahren kennen und anwenden sowie Formen der Evaluation und Rechenschaftslegung im Bildungswesen auf das eigene Handlungsfeld hin beurteilen können (BMUKK & BMWF, 2010, S. 46).

Dementsprechend wird Evaluation als ein wichtiges Element evidenzbasierter Steuerungslogik und schulischer Qualitätskontrolle angesehen (Altrichter & Kemethofer, 2016; Alt-

richter & Maag Merki, 2016). Grundgedanke ist auch hier, dass die Evaluation von Schule und Unterricht Hinweise für Verbesserungen liefern kann, indem eine Überprüfung und Bewertung entlang vorgegebener Kriterien erfolgt. Jedoch kommt hierbei Prozess- und Kontextfaktoren eine größere Aufmerksamkeit zu als bei der Rückmeldung von Schülerinnen- und Schülerleistungen anhand von Bildungsstandards. Interne Evaluation wird im Zusammenhang mit schulischem Qualitätsmanagement z. B. von Gärtner (2013) im engeren Sinne definiert „als selbstbestimmte Messung bestimmter Aspekte von Qualität im Rahmen schulischer Entwicklungsplanung“ (S. 699). Abgrenzungsmerkmal zur externen Evaluation ist, dass die Schule „selbst der relevante Steuerungsakteur ist“ (ebd.). Im Rahmen von externer Evaluation erfolgt die Überprüfung und Bewertung durch schulfremde Personen. Externe Evaluationen können zwar auch von der Schule in Auftrag gegeben werden, traditionell sind diese jedoch in Form von Schulinspektionen (auch als Fremdevaluation oder Schulvisitation bezeichnet) institutionalisiert und für Schulen verpflichtend (Kemethofer, 2015). In Österreich wurde – trotz langer Tradition der Schulinspektion (vgl. Altrichter, Kemethofer & Schmiedinger, 2013) – erst kürzlich ein bundesweites Verfahren der externen Schulevaluation eingeführt (Chabera, 2021).

2.2 Qualitätssicherung und -entwicklung als „neue“ Aufgabe von Schulleitungen

Mit besagten neuen Steuerungsvorstellungen im Bildungswesen geht auch „ein reformiertes Verständnis der Rolle und Funktion der Schulleitung“ einher (Bonsen, 2016, S. 302): Schulleitungen agieren in einer „Vermittlungs- und Kontrollposition zwischen Administration und Einzelschule“ (ebd., S. 303). Mit anderen Worten sind Schulleitungen gefordert, administrative und politische Vorgaben umzusetzen und (im Konsens mit dem Kollegium) an die „jeweiligen Handlungsbedingungen vor Ort“ anzupassen (ebd., S. 303). Daher tragen sie „die besondere Verantwortung für die Entwicklung und Sicherung der Qualität schulischer und unterrichtlicher Arbeit“ (ebd., S. 304). Insofern handelt es sich bei dieser Akteursgruppe um diejenige, die einen umfassenden Überblick über die QSE-Maßnahmen an der eigenen Schule hat und entsprechend auskunftsfähig sein dürfte.

Daher wurden im Rahmen der BIST-Ü Schulleitungen wiederholt zu Maßnahmen schulischer QSE an ihrer eigenen Schule befragt, und zwar im Bereich der Personalentwicklung zu (i) Analyse des Bedarfs an beruflicher Weiterbildung der Lehrpersonen und der Schulleitung, dem Vorhandensein eines (ii) Konzepts zur beruflichen Weiterbildung der Lehrpersonen und der Schulleitung³ sowie der Verfügbarkeit von (iii) Mentorinnen und Mentoren für junge Lehrpersonen. Diese erfragten QSE-Maßnahmen beziehen sich darauf, dass die konkrete Ausgestaltung der Fortbildung am Schulstandort der Schulleitung obliegt. Damit verbunden ist die Aufgabe, bedarfsgerechte Fortbildung zu verankern und zu evaluieren, Stärken und Entwicklungsbedarfe zu identifizieren und darauf aufbauend ein Konzept zur beruflichen Fort- und Weiterbildung für das Kollegium und für sich selbst

3 2013 wurde dieses Item nicht miterhoben, weshalb der Index 2013 lediglich auf sechs Items basiert.

zu erarbeiten (BMBWF, 2019). Ebenso ist die Schulleitung damit betraut, das Mentoring für Berufseinsteigerinnen und Berufseinsteiger an der Schule zu koordinieren, zu beraten und zu begleiten sowie die Personalstelle vor Ablauf der Induktionsphase über den Verwendungserfolg zu informieren. Zusätzlich wurden die Schulleitungen im Rahmen von BIST-Ü dazu befragt, inwiefern an der jeweiligen Schule Evaluationsaktivitäten stattfinden, und zwar in Form (iv) interner Evaluation/Selbstbeurteilung, (v) externer Evaluation sowie von (vi) Feedback von Schülerinnen und Schülern, z. B. hinsichtlich des Unterrichts, der Lehrpersonen oder der vorhandenen Ressourcen. Alle diese verschiedenen Formen schulischer Qualitätssicherung und -entwicklung können in schulische Entwicklungsprozesse einfließen, weshalb u. a. auch erfragt wurde, inwiefern institutionalisierte Schulentwicklungsprozesse stattfinden, und zwar in Form einer (vii) Schulentwicklungsgruppe.

2.3 Forschungsstand

Insgesamt betrachtet ist die Befundlage zu den Effekten der im Rahmen der BIST-Ü erfragten QSE-Maßnahmen auf die Schülerinnen- und Schülerleistung keineswegs eindeutig: So quitiert beispielsweise Hattie (2024) Lehrerinnen- und Lehrer-Fortbildungen (i bzw. ii) eine „wünschenswerte“, also dem durchschnittlichen Leistungszuwachs von einem Schuljahr entsprechende mittlere Effektstärke. Weitere Metaanalysen berichten ebenfalls von signifikanten Zusammenhängen mit jedoch kleinerer Effektstärke (u. a. Basma & Savage, 2018; Blank & de las Alas, 2009; Sims et al., 2021). Für die mentorielle Begleitung der Induktionsphase (iii) ist die Befundlage spärlicher und nicht eindeutig (Abdallah & Alkaabi, 2023; Ingersoll & Strong, 2011): So finden sich sowohl Befunde, die dafür sprechen, dass Schülerinnen und Schüler tendenziell bessere Leistungen zeigen, wenn sie von jungen Lehrpersonen unterrichtet werden, die Unterstützungsangebote beim Berufseinstieg erhalten haben, aber auch Befunde, die solche Effekte nicht finden. Im Hinblick auf Schulinspektionen (v) werden in der Literatur ebenfalls sowohl positive als auch negative Effekte auf die Leistungsentwicklung berichtet (vgl. zusammenfassend Altrichter & Kemethofer, 2016). Hinsichtlich anderer Formen externer Evaluation wie z. B. Vergleichsarbeiten, zentrale Abschlussprüfungen oder auch interne Evaluationen (iv) lassen sich Wurster et al. (2017) zufolge keine Effekte datengestützter Entwicklungsaktivitäten auf die Schülerinnen- und Schülerleistung nachweisen. Bezüglich Feedback von Schülerinnen und Schülern (v) als eine spezifische Form interner Evaluation (Gärtner, Wurster & Pant, 2014) sprechen die Ergebnisse der Metaanalyse von Röhl (2021) dafür, dass dies nach Auskunft von Lehrpersonen zu einem veränderten Verhalten im Klassenzimmer und darüber vermittelt zu einer veränderten Einschätzung der Unterrichtsqualität (durch Schülerinnen und Schüler) führen kann, insbesondere wenn Lehrkräfte in der Reflexion der Rückmeldung und der Ableitung von Entwicklungsmaßnahmen gezielt unterstützt werden. Weitere Evidenz findet sich für den Zusammenhang zwischen Einschätzungen der Unterrichtsqualität durch Schülerinnen und Schüler und der Schülerinnen- und Schülerleistung (Fauth, Decristan, Rieser, Klieme & Büttner, 2014; Senden, Nilsen & Teig, 2023), wobei es hierbei kulturspezifische Unterschiede zu geben scheint (Herbert et al., 2022). In einer umfassenden Studie konnte Feldhoff (2011) schließlich aufzeigen, dass Steuergruppen (vii) schulische Entwicklungsprozesse fördern und den Unterricht positiv beeinflussen können.

Dies gelingt, „indem sie die schulentwicklungsbezogene Kooperation durch die Vorgabe klarer Ziele und durch ein regelmäßiges Feedback zur Erreichung der Ziele forcieren“ (ebd., S. 288).

Trotz einer nicht eindeutigen Befundlage zu den Wirkungen unterschiedlicher QSE-Maßnahmen steht außer Zweifel, dass sich Schulen aktiv entwickeln müssen, um auf nicht-beeinflussbare Veränderungen (wie etwa sozialer und technischer Wandel) zu reagieren. Kontinuierliche Qualitätsentwicklung ist demnach für den künftigen Schulerfolg unabdingbar. Im Umkehrschluss bedeutet dies, dass sich sowohl „gute“ Schulen als auch „weniger gute“ Schulen mit QSE-Maßnahmen auseinandersetzen müssen (Holtappels, van Ackeren, Bremm & Rutter, 2021). Eine kontinuierliche Auseinandersetzung mit der eigenen Qualität und entsprechenden Maßnahmen ist insbesondere deshalb bedeutend, als komplexere Veränderungsprozesse – und eine Wirkkette, die bei Schülerinnen- und Schülerleistungen endet, gehört zweifelsohne dazu – mindestens drei bis fünf Jahre, potenziell sogar länger dauern (Feldhoff & Radisch, 2021). Das kann ein Stück weit erklären, warum bisherige Analysen auf Basis querschnittlicher Daten zum Zusammenhang zwischen den genannten QSE-Maßnahmen an Schulen und den Mathematikleistungen auf Sekundarstufe 1 in der BIST-Ü 2012 praktisch keine direkten Effekte aufzeigen konnten (Kemethofer, Zuber, Helm, Demski & Rie, 2015).

Zusammenfassend ist festzuhalten, dass es für die Qualität von Schule und Unterricht vermutlich weniger von Bedeutung ist, ob eine bestimmte QSE-Maßnahme ergriffen wird, sondern vielmehr, ob überhaupt aktive Entwicklungsprozesse an Schulen stattfinden und in welchem Umfang. Grundlegende Annahme für die hier vorgestellten Befunde ist also, dass an „entwicklungsaktiveren“ Schulen über die Zeit auch bessere Schülerinnen- und Schülerleistungen zu verzeichnen sein sollten. Dabei gilt es zu prüfen, inwiefern ein „Mehr“ an QSE-Aktivität auch ein „Mehr“ an Schülerinnen- und Schülerleistungen mit sich bringt (linearer Zusammenhang), oder ob eher eine moderate (in Abgrenzung zu keiner oder intensiver) Entwicklungsaktivität mit Schülerinnen- und Schülerleistungen in (kurvilinearem) Zusammenhang steht – Hinweise hierauf finden sich z. B. bei Groß Ophoff et al. (2023).

3 Fragestellungen

Dieser Beitrag stellt eine Weiterführung des Analysevorgehens von Kemethofer et al. (2015) dar, indem die Schülerinnen- und Schülerleistungen auf schulischer Ebene im Längsschnitt untersucht werden und neben dem Fach Mathematik auch Englisch in den Blick genommen wird. Hierzu werden insbesondere die Zusammenhänge zwischen den zeitvariablen Anteilen der Schülerinnen- und Schülerleistung und der QSE-Aktivität in den Blick genommen, während die über die Zeit stabilen Anteile kontrolliert werden. Damit verbunden stellen sich folgende Forschungsfragen:

1. Gibt es Zusammenhänge zwischen schulischen QSE-Maßnahmen und Schülerinnen- und Schülerleistung und sind diese linear (im Sinne von „Je mehr QSE-Aktivität, desto bessere Schülerinnen- und Schülerleistung“, vgl. 5.1) oder kurvilinear (als Indikator dafür, dass z. B. besonders aktive Schulen womöglich nicht davon profitieren)?

2. Können auf Schulebene aggregierte Schülerinnen- und Schülerleistungen in den BIST-Ü vorhergesagt werden (vgl. 5.2 und 5.3) ...
 - a. durch die Leistungen des Schulstandorts in früheren BIST-Ü, also z. B. Englisch- durch frühere Mathematikleistungen,
 - b. durch die Summe der QSE-Maßnahmen und
 - c. finden sich die Effekte innerhalb des gleichen Erhebungszeitraums oder entfalten die QSE-Maßnahmen erst zeitversetzt ihren Effekt?

4 Methoden

Zur Beantwortung der Forschungsfragen wird auf Daten der österreichischen Bildungsstandardüberprüfungen in den Fächern Mathematik 2012 (M812) und 2017 (M817) sowie Englisch aus den Jahren 2013 (E813) und 2019 (E819) auf der Sekundarstufe zurückgegriffen und mit diesen eine Sekundärdatenanalyse durchgeführt. Das für die Untersuchung herangezogene Datenmaterial und die Operationalisierung im Rahmen der BIST-Ü 2012 bis 2019 werden im Folgenden dargestellt sowie die Analysestrategie erläutert.

4.1 Datenmaterial

An den flächendeckenden Erhebungen nahmen an allen öffentlichen und privaten Schulen Schülerinnen und Schüler der achten Schulstufe teil, mit Ausnahme von außerordentlichen Schülerinnen und Schülern und solchen mit sonderpädagogischem Förderbedarf. Zusätzlich zur standardisierten Überprüfung der Mathematik- und Englischkompetenzen wurden Kontextfragebögen eingesetzt, in denen Schulleitungen angeben konnten, welche Instrumente zur Qualitätssicherung und -entwicklung am eigenen Schulstandort umgesetzt werden. Tabelle 1 zeigt die Datenbasis der vier Vollerhebungen (siehe auch die entsprechenden Almanache auf www.iqs.gv.at):

Tabelle 1: Datenbasis

Bildungsstandardüberprüfung	N Schülerinnen und Schüler	N Schulleitungen
M812	79.678	1.405
E813	76.728	1.410
M817	72.704	1.344–1.374*
E819	74.241	1.369
Längsschnitt MSQE	-	1.359
Längsschnitt Schülerinnen- und Schülerleistung	-	1.211
<i>Anmerkung: *N variiert je nach Item.</i>		

4.2 Operationalisierung

In diesem Beitrag werden zwei Aspekte in den Blick genommen, nämlich einerseits die Angaben der Schulleitungen zu realisierten QSE-Maßnahmen und andererseits die Schülerinnen- und Schülerleistung in den jeweils untersuchten Fächern im Rahmen der BIST-Ü, und zwar aggregiert auf Schulebene.

4.2.1 Maßnahmen zur Qualitätssicherung und -entwicklung

Die Fragebögen für Schulleitungen erfragten zu allen vier Messzeitpunkten, welche der unter 2.2 und 2.3 genannten QSE-Maßnahmen am Schulstandort realisiert werden. Als Antwortoptionen stand jeweils eine dichotome Skala mit den Ausprägungen *ja* (= 1) bzw. *nein* (= 0) zur Verfügung. Die sieben Items wurden durch Mittelwertbildungen zu einem QSE-Aktivitätsindex zusammengefasst. Im Mittel setzten Schulen – je nach Messzeitpunkt – zwischen drei und fünf Maßnahmen zur Qualitätssicherung und -entwicklung um (siehe Abbildung 1; M812: $M = 2.96$, $SD = 1.43$; E813: $M = 2.62$, $SD = 1.28$; M817: $M = 4.28$, $SD = 1.45$; E819: $M = 3.62$, $SD = 1.40$).

4.2.2 Schülerinnen- und Schüler- und Schulleistung

Die Schulleistung wurde über die Fachleistung der Schülerinnen und Schüler einer Schule in den beiden Unterrichtsgegenständen Mathematik und Englisch erfasst. Im Rahmen des Tests zu den Bildungsstandards in Mathematik in 2012 und 2017 musste jede Schülerin und jeder Schüler etwa 70 Aufgaben lösen, die auf zwei Testhefte zu je 40 Minuten verteilt waren. Jedes Testheft hatte ungefähr den gleichen Schwierigkeitsgrad, gemischt mit Multiple-Choice-Aufgaben und offenen Aufgaben. Im Fach Englisch wurde sowohl 2013 als auch 2019 ein Testheftdesign zur Messung von drei Kompetenzbereichen eingesetzt, sodass die Anzahl an Items in einem Testheft je nach Kompetenzbereich und Testform variiert. So enthielten die Testhefte in Hören bis zu 33 Aufgaben (ausschließlich geschlossenes Format), die Testhefte in Lesen etwa 16 Aufgaben (maximal zwei Aufgaben mit [halb]offenem Format) und die Testhefte in Schreiben immer zwei Aufgaben (ausschließlich offenes Format), wofür eine Bearbeitungszeit von insgesamt etwa 90 bzw. 100 Minuten vorgesehen war. Zu Vergleichszwecken wurden die Testergebnisse auf eine für Large-Scale Assessments übliche Skala transformiert, die in der Baseline-Testung 2009 einen Mittelwert von 500 und eine Standardabweichung von 100 ausweist. Im Durchschnitt erreichten die Schülerinnen und Schüler folgende Punkte: M812: $M = 524.14$, $SD = 56.95$; E813: $M = 504.46$, $SD = 58.81$; M817: $M = 531.83$, $SD = 54.15$; E819: $M = 568.93$, $SD = 57.37$ (siehe Abbildung 1)⁴. Die Schülerinnen- und Schülerleistung wurde unter Verwendung des Schülerinnen- und Schüler-Gewichts auf Schulebene aggregiert, um die Schulleistung zu erhalten.

Als Kovariate wird in allen Analysen der Schultyp ($N = 1.094$ allgemeinbildende Pflichtschulen = APS; $N = 266$ allgemeinbildende höhere Schulen = APS) aufgenommen. Weitere Informationen zu den Daten (z. B. Zusammensetzung der Schülerinnen und Schüler) und

4 Da die Skalen über die Fächer nicht miteinander verlinkt sind, sind die Schülerinnen- und Schülerleistungen zwischen den Fächern nicht vergleichbar.

den Instrumenten zur Erfassung der Mathematik- und Englischleistungen der Schülerinnen und Schüler finden sich in den jeweiligen nationalen Ergebnisberichten (Bundesinstitut BIFIE, 2020; Schreiner & Breit, 2012, 2013; Schreiner et al., 2017) sowie bei Breit und Schreiner (2016).

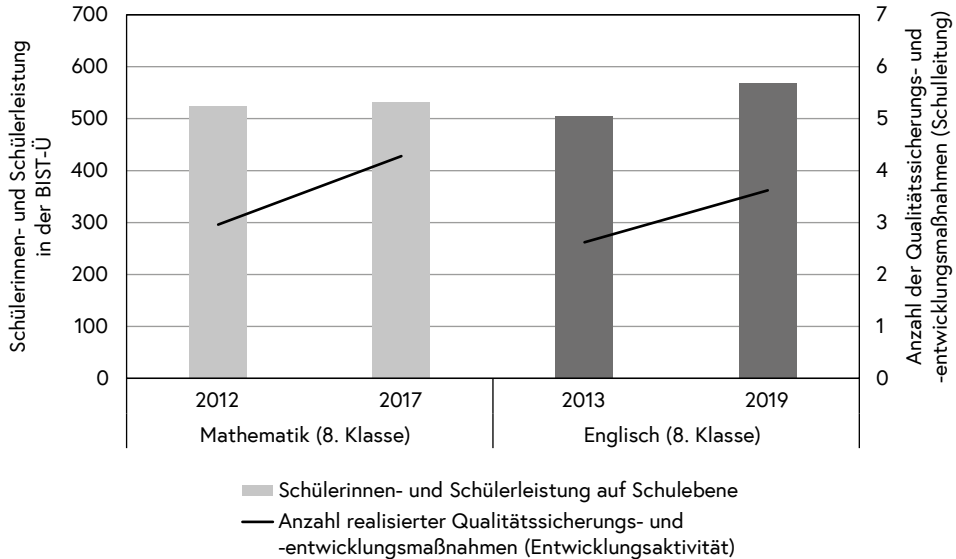


Abbildung 1: Zusammenfassende Darstellung der deskriptiven Statistiken (Mittelwerte)

4.3 Analysestrategie

Alle Analysen wurden mit dem R-Paket MplusAutomation (Hallquist & Wiley, 2018) in Kombination mit Mplus 8 (Muthén & Muthén, 2017) durchgeführt. Die Modelle wurden mit einer robusten Maximum-Likelihood-Schätzung (MLR) geschätzt, wobei die Standardfehler robust gegenüber nichtnormalen Verteilungen und der Verwendung von Ordinalvariablen sind (Finney & DiStefano, 2006). Fehlende Werte wurden in den nachfolgenden Analysen mithilfe der Full-Information-Maximum-Likelihood-(FIML-)Methode adressiert. Bei einem N von 1.360 Schulen lag der Anteil an fehlenden Werten bei allen untersuchten Variablen unter 1 %. Zur Bestimmung der Modellpassung dienten gängige Gütekriterien (Hu & Bentler, 1999) – der Bentler's Comparative Fit Index ($CFI \geq .95$), der Tucker-Lewis Index ($TLI \geq .95$), der Root Mean Square Error of Approximation ($RMSEA \leq .06$) und der Standardised Root Mean Square Residual ($SRMR \leq .08$).

4.3.1 Voranalysen

Zunächst wurde geprüft, ob der Zusammenhang zwischen den am Schulstandort umgesetzten QSE-Maßnahmen und der Schülerinnen- und Schülerleistung einem linearen oder einem quadratischen Verlauf folgt (vgl. 4.1 Prüfung der Linearitätsannahme). Hierfür wurden multivariate Regressionsanalysen durchgeführt, die neben dem Haupteffekt der Maßnahmen auf die Schülerinnen- und Schülerleistung auch den Effekt eines quadratischen

Terms der Maßnahmen auf statistische Signifikanz prüfen. Die Prädiktoren, die zur Bildung des quadratischen Terms herangezogen wurden, wurden vorab zentriert (Aiken & West, 1991). Das Fehlen eines statistisch signifikanten quadratischen Terms ist zwar kein Beleg für die Linearität des Zusammenhangs, schließt aber einen plausiblen, alternativen Zusammenhang aus.

4.3.2 Bivariate Analysen

Zur Prüfung der zugrunde liegenden Struktur der Daten wurden in einem weiteren Schritt bivariate Korrelationen zwischen den beiden Untersuchungsvariablen, also der Schülerinnen- und Schülerleistung auf Schulebene und dem Mittelwert der QSE-Maßnahmen, über die vier Messzeitpunkte hinweg berechnet (vgl. 5.2 Bivariate Korrelationsanalysen).

4.3.3 Multivariate Analysen

Zur Untersuchung der Zusammenhänge zwischen den am Schulstandort umgesetzten QSE-Maßnahmen mit der Schülerinnen- und Schülerleistung (vgl. 5.3 Längsschnittliche Cross-Lagged-Panel-Modelle) wurden hierzu sowohl klassische Cross-Lagged-Panel-Modelle als auch Random-Intercept-Cross-Lagged-Panel-Modelle geschätzt (vgl. Abbildung 2 und Abbildung 3).

Das „klassische“ Cross-Lagged-Panel-Modell mit Lag-1-Effekten (bei Lüdtke & Robitzsch, 2022 auch CL1 genannt) wird in erster Linie verwendet, um Hypothesen zu testen, die sich auf die kreuzverzögerten Effekte zwischen zwei Variablen in einem Längsschnitt-Panel-Design beziehen. Dieses Modell konzentriert sich auf die Schätzung des kausalen Effekts eines Prädiktors (hier QSE-Maßnahmen und Schülerinnen- und Schülerleistung zu Zeitpunkt t) zu einem bestimmten Zeitpunkt auf ein Ergebnis zu einem späteren Zeitpunkt (hier QSE-Maßnahmen und Schülerinnen- und Schülerleistung zu Zeitpunkt $t + 1$), wobei die autoregressiven Effekte beider Variablen und alle zeitvariablen Kovariaten (in der vorliegenden Studie keine) kontrolliert werden (z. B. Lüdtke & Robitzsch, 2022). Mit dem CLPM kann die Direktionalität der Effekte, also die Richtung der kausalen Effekte zwischen der Prädiktorvariable zum Zeitpunkt $t - 1$ und der Ergebnisvariable zum Zeitpunkt t untersucht werden. Zudem geben CLPM Aufschluss darüber, ob Veränderungen des Prädiktors zu nachfolgenden Veränderungen der Ergebnisvariable führen. Im vorliegenden Fall wird geprüft, ob eine Veränderung in den QSE-Maßnahmen zu einer Veränderung in der Schülerinnen- und Schülerleistung führt und umgekehrt. Das CLPM kontrolliert die autoregressiven Effekte beider Variablen und stellt damit sicher, dass die geschätzten kreuzverzögerten Effekte nicht durch die eigenen Vergangenheitswerte der Variablen verfälscht werden.

Das Random-Intercept-Cross-Lagged-Panel-Modell wird – wie auch das klassische CLPM – eingesetzt, um Hypothesen über die kausalen Beziehungen (kreuzverzögerte Effekte, Direktionalität) zwischen zwei und mehr Variablen in einem Längsschnitt-Panel-Design zu testen. Im Gegensatz zum CPLM ermöglicht das RI-CPLM allerdings, nicht gemessene zeitinvariante Störfaktoren zu berücksichtigen. Dies geschieht durch die getrennte Analyse von zeitinvarianten und -variablen Anteilen in den beiden Konstrukten – wodurch die hierarchische Struktur der Daten (Messzeitpunkte innerhalb von Schulstandorten) Be-

rücksichtigung findet (z. B. Hamaker, Kuiper & Grasmann, 2015). Darüber hinaus erlaubt das Modell die Untersuchung der Stabilität von Beziehungen im Zeitverlauf, indem untersucht wird, ob die Auswirkungen von Variablen über mehrere Messwellen hinweg bestehen bleiben oder sich verändern. Im vorliegenden Fall wird – wie beim CLPM – geprüft, ob eine Veränderung in den QSE-Maßnahmen zu einer Veränderung in der Schülerinnen- und Schülerleistung führt und umgekehrt. Das RI-CLPM kontrolliert ebenfalls für die autoregressiven Effekte beider Variablen, aber zusätzlich auch für nicht gemessene zeitinvariante Störfaktoren (z. B. Schultyp), indem die Varianzen der Untersuchungsvariablen in zeitinvariante Anteile (Varianz zwischen den Schulstandorten) und zeitvariable Anteile (Varianz innerhalb der Schulstandorte über die Zeit) geteilt wird.

Das Cross-Lagged-Panel-Modell wurde als vollständig autoregressives Modell spezifiziert, d. h. die beiden Untersuchungsvariablen wurden zum jeweiligen Messzeitpunkt auf jeweils alle davorliegenden Messzeitpunkte regressiert. Das Random-Intercept-Cross-Lagged-Panel-Modell wurde nach Hamaker et al. (2015) spezifiziert, sodass die Varianz der beiden Untersuchungsvariablen zum jeweiligen Messzeitpunkt in ihre stabile/zeitinvariante und ihre variablen/zeitvarianten Anteile zerlegt werden.

4.3.4 Post-hoc-Analysen

Um die Frage nach moderierenden Einflüssen bspw. des Kontexts der Schulen auf die Beziehung zwischen QSE-Maßnahmen und Schülerinnen- und Schülerleistung zu untersuchen, wurden Multigroup-Analysen für den Schultyp durchgeführt. Die entsprechenden Analysen wurden sowohl für die klassischen als auch für die Random-Intercept-CLPM geschätzt. Darüber hinaus wurden differenzielle Effekte einzelner QSE-Maßnahmen untersucht, indem alle Analysen auch auf Ebene der Einzelitems durchgeführt wurden. Entsprechende Ergebnistabellen werden allerdings ausschließlich im Anhang berichtet.

5 Ergebnisse

Zunächst werden die Ergebnisse der Prüfung der Linearitätsannahme (5.1) vorgestellt, auf denen aufbauend die Analyse des Effekts von QSE-Maßnahmen auf die Schülerinnen- und Schülerleistung über die Zeit vorgenommen wird (5.2 ff.).

5.1 Prüfung der Linearitätsannahme

Die Prüfung der Linearitätsannahme bzgl. der Zusammenhänge zwischen den am Schulstandort umgesetzten Maßnahmen und der Schülerinnen- und Schülerleistung zeigt in allen Modellen, dass der quadratische Term – unter Berücksichtigung des Schultyps als Kovariate – statistisch nicht signifikant ist (vgl. Tabelle 2). Daher werden die Folgeanalysen unter der Linearitätsannahme fortgeführt.

Tabelle 2: Prüfung auf nichtlinearen Zusammenhang zwischen QSE-Maßnahmen und Schülerinnen- und Schülerleistung

Erhebung	UV	AV	std. Schätzer	SE	p
2012	QSE12	M812WLE	-.07	.08	.39
	QSE12^2		.02	.08	.86
	Schultyp		.66	.02	.00
2013	QSE13	E813WLE	-.03	.06	.66
	QSE13^2		.01	.06	.94
	Schultyp		.79	.01	.00
2017	QSE17	M817WLE	-.13	.09	.16
	QSE17^2		.10	.10	.30
	Schultyp		.60	.02	.00
2019	QSE19	E819WLE	-.09	.08	.31
	QSE19^2		.08	.08	.30
	Schultyp		.72	.01	.00

Anmerkungen: Abkürzungen: M812 = Bildungsstandardüberprüfung Klasse 8 in Mathematik im Jahr 2012; M817 = Mathematikerhebung in 2017; E813 = Bildungsstandardüberprüfung in Englisch im Jahr 2013; E819 = Englischerhebung in 2019. WLE = Weighted Likelihood Estimate (Personenschätzer für die Leistung im Kompetenztest); QSE = Anzahl der Maßnahmen zur Qualitätssicherung und -entwicklung an der jeweiligen Schule.

5.2 Bivariate Korrelationsanalysen

Die hohen bivariaten Korrelationen der Schülerinnen- und Schülerleistung über die Zeit (linker oberer Abschnitt in Tabelle 4, $r = .72-.84$, $p < .01$) verweisen auf eine hohe Stabilität der Schülerinnen- und Schülerleistung auf Schulebene. Dagegen sind die Indexe untereinander und über die Zeit nur moderat korreliert (rechter unterer Abschnitt in Tabelle 3, $r = .26-.48$, $p < .01$), weshalb für die Anzahl an QSE-Maßnahmen in Schulen zwischen den Erhebungsjahren eine eher hohe Variabilität anzunehmen ist – einmal werden mehrere Maßnahmen in einem Jahr angeführt, dann wieder wenige. Zwischen den am Schulstandort umgesetzten QSE-Maßnahmen und der Schülerinnen- und Schülerleistung zeigt sich eine Reihe von schwachen, aber statistisch signifikanten (teils auch zeitversetzten) Zusammenhängen (linker unterer Abschnitt in Tabelle 3, $r = .07-.13$, $p < .01$). Sie stützen vorerst die Hypothese des Effektes der QSE-Maßnahmen auf die Schülerinnen- und Schülerleistung.

Tabelle 3: Korrelationen zwischen und innerhalb der Messzeitpunkte von Schülerinnen- und Schülerleistung und QSE-Maßnahmen

	M812WLE	E813WLE	M817WLE	E819WLE	QSE12	QSE13	QSE17	QSE19
M812WLE	1.00							
E813WLE	.79	1.00						
M817WLE	.84	.75	1.00					
E819WLE	.74	.81	.72	1.00				
QSE12	.07	.13	.07	.13	1.00			
QSE13	.07	.10	.08	.09	.48	1.00		
QSE17	n. s.	n. s.	n. s.	n. s.	.29	.32	1.00	
QSE19	n. s.	.08	n. s.	.08	.26	.27	.43	1.00

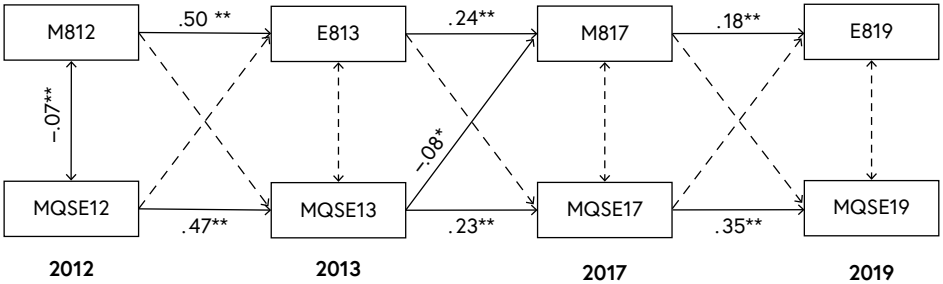
Anmerkungen: Abkürzungen: M812 = Bildungsstandardüberprüfung Klasse 8 in Mathematik im Jahr 2012; M817 = Mathematikerhebung in 2017; E813 = Bildungsstandardüberprüfung in Englisch im Jahr 2013; E819 = Englischerhebung in 2019. WLE = Weighted Likelihood Estimate (Personenschätzer für die Leistung im Kompetenztest); QSE = Anzahl der Maßnahmen zur Qualitätssicherung und -entwicklung an der jeweiligen Schule.

5.3 Längsschnittliche Cross-Lagged-Panel-Modelle

Im Folgenden werden zunächst die Ergebnisse für das klassische Cross-Lagged-Panel-Modell berichtet und anschließend die Ergebnisse für das Random-Intercept-Cross-Lagged-Panel-Modell vorgestellt.

5.3.1 Klassisches Cross-Lagged-Panel-Modell

Die Analyse weist für das Modell einen sehr guten Fit aus ($N = 1.360$, $Par = 46$, $\chi^2 = 7.10$, $df = 6$, $\chi^2 p \text{ value} = .31$, $\chi^2/df = 1.18$, $CFI = 1$, $TLI = .99$, $RMSEA = .01$, $SRMR = .01$). Die postulierten Annahmen spiegeln also die Struktur in den Daten gut wider; oder anders, die Daten können durch das theoretische Modell ausreichend gut reproduziert werden. Die in Abbildung 2 und in Tabelle 4 präsentierten Ergebnisse zeigen, dass sowohl die Schülerinnen- und Schülerleistung als auch das durchschnittliche Ausmaß der Maßnahmen zur Qualitätssicherung und -entwicklung jeweils durch die zurückliegenden Messpunkte am stärksten vorhergesagt werden (siehe die Between-Wave-Pfade: $\beta = .18$ bis $.50$, $p < .01$). Über diese autoregressiven Pfade (sowie den signifikanten Einfluss des Schultyps) hinaus sind keine signifikanten Zusammenhänge beobachtbar, insbesondere keine kreuzverzögerten Effekte der Schülerinnen- und Schülerleistung auf die QSE-Maßnahmen oder umgekehrt der QSE-Maßnahmen auf die Schülerinnen- und Schülerleistung. Lediglich die Mathematikleistung 2017 auf Schulebene wird statistisch signifikant, aber nur schwach negativ durch die QSE-Maßnahmen 2013 vorhergesagt ($\beta = -.08$, $p < .05$).



Anmerkungen: Alle acht Indikatorvariablen wurden für den Schultyp kontrolliert. Die autoregressiven Pfade zu früheren Erhebungen sowie die Regressionskoeffizienten der Kontrollvariable „Schultyp“ sind nicht abgebildet. Die entsprechenden Koeffizienten sind in Tabelle 5 ersichtlich.
Abkürzungen: M812 = Bildungsstandardüberprüfung Klasse 8 in Mathematik im Jahr 2012; M817 = Mathematiküberprüfung in 2017; E813 = Bildungsstandardüberprüfung in Englisch im Jahr 2013; E819 = Englischerhebung in 2019. MQSE = Durchschnittliche Anzahl (Mittelwert) der Maßnahmen zur Qualitätssicherung und -entwicklung an der jeweiligen Schule in den Jahren 2012 bis 2019.

Abbildung 2: Standardisierte Schätzer des klassischen Cross-Lagged-Panel-Modells

Tabelle 4: Standardisierte Schätzer des klassischen Cross-Lagged-Panel-Modells

Pfad	AV	UV	std. Schätzer	SE	p
Between Wave	MQSE19	MQSE17	.35	.03	.00
	MQSE19	MQSE13	.09	.03	.00
	MQSE19	MQSE12	.11	.03	.00
	MQSE17	MQSE13	.23	.03	.00
	MQSE17	MQSE12	.17	.03	.00
	MQSE13	MQSE12	.47	.03	.00
	E819WLE	M817WLE	.18	.03	.00
	E819WLE	E813WLE	.40	.03	.00
	E819WLE	M812WLE	.15	.03	.00
	M817WLE	E813WLE	.24	.03	.00
	M817WLE	M812WLE	.67	.02	.00
	E813WLE	M812WLE	.49	.02	.00
Cross-Lagged	MQSE19	M817WLE	-.02	.03	.59
	MQSE17	E813WLE	-.08	.04	.05
	MQSE13	M812WLE	.00	.03	.98
	E819WLE	MQSE17	.01	.02	.62
	M817WLE	MQSE13	.01	.01	.32
	E813WLE	MQSE12	-.00	.01	.82

Within Wave	MQSE19	E819WLE	.02	.03	.38
	MQSE17	M817WLE	.01	.03	.66
	MQSE13	E813WLE	-.02	.03	.51
	MQSE12	M812WLE	-.07	.03	.01
Covariate	MQSE19	AHS	.07	.03	.03
	MQSE17	AHS	.07	.04	.08
	MQSE13	AHS	.06	.03	.10
	MQSE12	AHS	.20	.03	.00
	E819WLE	AHS	.20	.03	.00
	M817WLE	AHS	-.02	.03	.33
	E813WLE	AHS	.46	.02	.00
	M812WLE	AHS	.65	.01	.00

Anmerkungen: Abkürzungen: M812 = Bildungsstandardüberprüfung Klasse 8 in Mathematik im Jahr 2012; M817 = Mathematikerhebung in 2017; E813 = Bildungsstandardüberprüfung in Englisch im Jahr 2013; E819 = Englischerhebung in 2019. WLE = Weighted Likelihood Estimate (Personenschätzer für die Leistung im Kompetenztest); MQSE = Durchschnittliche Anzahl (M) der Maßnahmen zur Qualitätssicherung und -entwicklung an der jeweiligen Schule.

Moderationsanalyse. Die Multigroup-Analysis (Tabelle 5; $N = 1.360$, $Par = 76$, $\chi^2 = 18.95$, $df = 12$, $\chi^2 p \text{ value} = .09$, $\chi^2/df = 1.57$, $CFI = .99$, $TLI = .99$, $RMSEA = .03$, $SRMR = .02$) zeigt, dass keine differenziellen Effekte für die Schultypen vorliegen, d. h., der Schultyp übt keinen moderierenden Einfluss auf die Zusammenhänge zwischen den durchschnittlichen QSE-Maßnahmen und den Schülerinnen- und Schülerleistung aus. In beiden Schultypen (AHS, APS) sind durchschnittliche QSE-Maßnahmen und Schülerinnen- und Schülerleistung statistisch nicht signifikant miteinander assoziiert. Nur hinsichtlich der autoregressiven Pfade, die im Falle der AHS nicht durchgängig signifikant sind, lassen sich Unterschiede zwischen den Schultypen beobachten. So wird etwa die Schülerinnen- und Schülerleistung in Englisch 2019 bei AHS überwiegend durch die Schülerinnen- und Schülerleistung in Englisch des Jahres 2013 erklärt. Bei den APS tragen auch Leistungen in Mathematik aus Vorjahren (2017 und 2012) zur Erklärung der Englischleistungen bei.

5.3.2 Random-Intercept-Cross-Lagged-Panel-Modell

Auch dieses Modell (vgl. Abbildung 3) weist einen sehr guten Fit auf ($N = 1.360$, $Par = 37$, $\chi^2 = 134.04$, $DF = 15$, $\chi^2 p \text{ value} < .01$, $\chi^2/df = 8.93$, $CFI = .98$, $TLI = .95$, $RMSEA = .08$, $SRMR = .03$). Die in Tabelle 6 präsentierten Ergebnisse verdeutlichen, dass primär die autoregressiven Pfade (Between Wave) sowie die Schultypenunterschiede (zugunsten der AHS) statistisch signifikant sind. Bedingt durch Zerlegung der Konstrukte in zeitinvariante und -variable Anteile hängt die Schülerinnen- und Schülerleistung auf Schulebene nun über die Zeit teilweise negativ zusammen ($\beta = -.57$ bzw. $-.57$, $p < .01$). Die Mittelwertindexe für die QSE-Maßnahmen sind zwischen den Erhebungszeitpunkten signifikant positiv miteinander assoziiert ($\beta = .08$ bis $.29$, $p < .05$). Die standardisierten Koeffizienten zeigen eine

Tabelle 5: Standardisierte Schätzer des klassischen Cross-Lagged-Panel-Modells für die APS und AHS getrennt (Multigroup-Analysis)

Pfad	AV	UV	APS			AHS		
			std. Est.	SE	p	std. Est	SE	p
Between Wave	MQSE19	MQSE17	.33	.03	.00	.44	.05	.00
	MQSE19	MQSE13	.13	.03	.00	-.02	.06	.81
	MQSE19	MQSE12	.11	.03	.00	.09	.06	.15
	MQSE17	MQSE13	.24	.03	.00	.19	.07	.01
	MQSE17	MQSE12	.16	.03	.00	.18	.06	.00
	MQSE13	MQSE12	.47	.03	.00	.46	.05	.00
	E819WLE	M817WLE	.22	.04	.00	.08	.06	.22
	E819WLE	E813WLE	.31	.04	.00	.61	.05	.00
	E819WLE	M812WLE	.20	.04	.00	.01	.07	.93
	M817WLE	E813WLE	.19	.03	.00	.15	.05	.00
	M817WLE	M812WLE	.64	.02	.00	.62	.04	.00
	E813WLE	M812WLE	.62	.02	.00	.52	.05	.00
Cross-Lagged	MQSE19	M817WLE	-.02	.03	.41	.05	.05	.38
	MQSE17	E813WLE	-.04	.03	.16	-.10	.05	.06
	MQSE13	M812WLE	-.02	.03	.56	.08	.05	.12
	E819WLE	MQSE17	.02	.02	.47	.03	.05	.58
	M817WLE	MQSE13	.01	.02	.66	.06	.04	.14
	E813WLE	MQSE12	-.01	.02	.68	.03	.06	.60
Within Wave	MQSE19	E819WLE	.04	.03	.22	.01	.06	.94
	MQSE17	M817WLE	.01	.03	.75	.03	.06	.68
	MQSE13	E813WLE	-.01	.03	.72	-.04	.06	.51
	MQSE12	M812WLE	-.10	.03	.00	.06	.07	.39

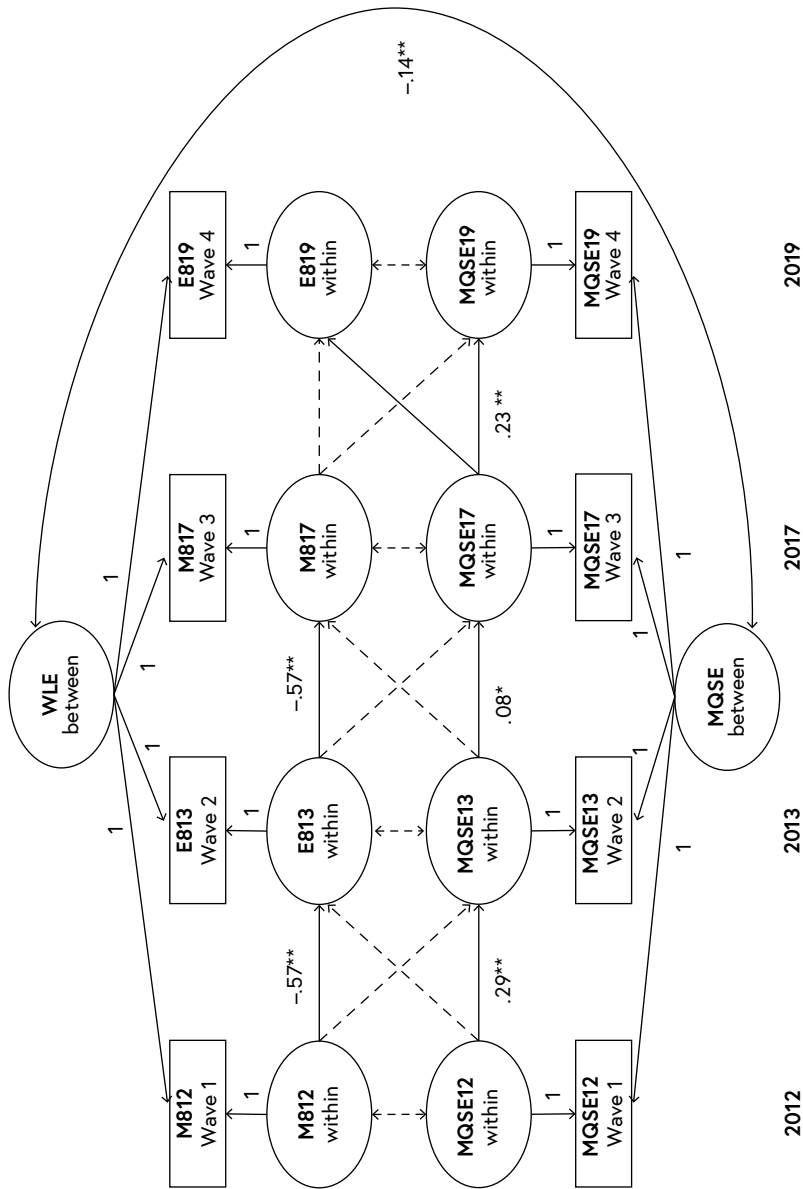
Anmerkungen: Abkürzungen: M812 = Bildungsstandardüberprüfung Klasse 8 in Mathematik im Jahr 2012; M817 = Mathematikerhebung in 2017; E813 = Bildungsstandardüberprüfung in Englisch im Jahr 2013; E819 = Englischerhebung in 2019. WLE = Weighted Likelihood Estimate (Personenschätzer für die Leistung im Kompetenztest); MQSE = Durchschnittliche Anzahl (M) der Maßnahmen zur Qualitätssicherung und -entwicklung an der jeweiligen Schule; APS = Allgemeinbildende Pflichtschule, AHS = Allgemeinbildende höhere Schule.

statistisch signifikante, moderat negative Korrelation ($r = -.14, p = .01$) zwischen (a) dem über die Zeit hinweg stabilen Anteil der QSE-Maßnahmenindexe und (b) dem über die Zeit hinweg stabilen Anteil in der Schülerinnen- und Schülerleistung (vgl. Across-Wave-Pfad in Tabelle 6). Dieser deutet darauf hin, dass in Schulen mit niedriger Schülerinnen- und Schülerleistung überzufällig häufiger QSE-Maßnahmen gesetzt werden. Allerdings ist dieser Zusammenhang sehr schwach ausgeprägt. Bezüglich der zeitvarianten Anteile (vgl. Cross-Lagged-Pfade in Tabelle 5) zeigen sich keine signifikanten Zusammenhänge. Diese Befunde decken sich im Wesentlichen mit den Befunden aus dem klassischen Cross-Lagged-Panel-Modell, auch dort sind mit einer Ausnahme keine kreuzverzögerten Zusammenhänge beobachtbar.

Tabelle 6: Standardisierte Schätzer des Random-Intercept-Cross-Lagged-Panel-Modells

Pfad	AV	UV	std. Schätzer	SE	p
Across Wave	WLE	MQSE	-.14	.05	.00
Between Wave	MQSE19	MQSE17	.23	.04	.00
	MQSE17	MQSE13	.08	.04	.04
	MQSE13	MQSE12	.29	.04	.00
	E819WLE	M817WLE	.05	.05	.29
	M817WLE	E813WLE	-.57	.05	.00
	E813WLE	M812WLE	-.57	.06	.00
Cross-Lagged	MQSE19	M817WLE	-.01	.03	.82
	MQSE17	E813WLE	.02	.04	.59
	MQSE13	M812WLE	.05	.05	.31
	E819WLE	MQSE17	.02	.03	.61
	M817WLE	MQSE13	.10	.09	.22
	E813WLE	MQSE12	.05	.05	.36
Within Wave	MQSE19	E819WLE	.01	.03	.68
	MQSE17	M817WLE	.04	.04	.35
	MQSE13	E813WLE	.10	.12	.44
	MQSE12	M812WLE	-.02	.03	.62
Covariate	MQSE19	AHS	.13	.03	.00
	MQSE17	AHS	.09	.03	.00
	MQSE13	AHS	.15	.03	.00
	MQSE12	AHS	.19	.03	.00
	E819WLE	AHS	.67	.01	.00
	M817WLE	AHS	.63	.02	.00
	E813WLE	AHS	.77	.01	.00
	M812WLE	AHS	.67	.01	.00

Anmerkungen: Abkürzungen: M812 = Bildungsstandardüberprüfung Klasse 8 in Mathematik im Jahr 2012; M817 = Mathematikerhebung in 2017; E813 = Bildungsstandardüberprüfung in Englisch im Jahr 2013; E819 = Englischerhebung in 2019. WLE = Weighted Likelihood Estimate (Personenschätzer für die Leistung im Kompetenztest); MQSE = Durchschnittliche Anzahl (M) der Maßnahmen zur Qualitätssicherung und -entwicklung an der jeweiligen Schule.



Anmerkungen: Alle acht Indikatorvariablen wurden für den Schultyp kontrolliert. Die entsprechenden Regressionskoeffizienten sind nicht abgebildet; sie sind in Tabelle 6 ersichtlich. Abkürzungen: M812 = Bildungsstandardüberprüfung Klasse 8 in Mathematik im Jahr 2012; M817 = Mathematikserhebung in 2017; E813 = Bildungsstandardüberprüfung in Englisch im Jahr 2013; E819 = Englischserhebung in 2019; MQSE = Durchschnittliche Anzahl (M) der Maßnahmen zur Qualitätssicherung und -entwicklung an der jeweiligen Schule in den Jahren 2012 bis 2019.

Abbildung 3: Standardisierte Schätzer des Random-Intercept-Cross-Lagged-Panel-Modells

Moderationsanalyse. Die Multigroup-Analysis für das Random-Intercept-Cross-Lagged-Panel-Modell (Tabelle 7; $N = 1.360$, $Par = 58$, $\chi^2 = 90.47$, $df = 30$, $\chi^2 p \text{ value} < .01$, $\chi^2/df = 3.01$, $CFI = .98$, $TLI = .96$, $RMSEA = .05$, $SRMR = .04$) zeigt – wie auch für das klassische Cross-Lagged-Panel-Modell, dass mit einer Ausnahme (siehe unten) keine differenziellen kreuzverzögerten Effekte für die Schultypen vorliegen, d. h., der Schultyp übt keinen moderierenden Einfluss auf die kreuzverzögerten Zusammenhänge zwischen den Maßnahmen der Qualitätssicherung und -entwicklung und der Schülerinnen- und Schülerleistung auf Schulebene aus. Auch was die Between-Wave-Koeffizienten und die Within-Wave-Koeffizienten betrifft, sind keine Schultypenunterschiede beobachtbar; in manchen Fällen sind die Zusammenhänge allerdings für die AHS stärker ausgeprägt. Einzig der Across-Wave-Koeffizient zeigt sich für die APS-Schulen statistisch signifikant negativ ($\beta = -.17$, $p < .01$), während dieser für die AHS-Schule nicht signifikant ist. Das bedeutet, dass in APS-Schulen mit niedrigeren Leistungen, nicht aber in AHS-Schulen mit niedriger Schülerinnen- und Schülerleistung, vermehrt QSE-Maßnahmen umgesetzt werden. Umgekehrt ist der kreuzverzögerte Effekt von der Schülerinnen- und Schülerleistung 2013 auf die QSE-Maßnahmen 2017 für die AHS-Schulen signifikant ($\beta = -.16$, $p = .03$), nicht aber für die APS-Schulen.

Tabelle 7: Standardisierte Schätzer des Random-Intercept-Cross-Lagged-Panel-Modells für die APS und AHS getrennt (Multigroup-Analysis)

Pfad	AV	UV	APS			AHS		
			std. Est.	SE	p	std. Est	SE	p
Across Wave	WLE	MQSE	-.17	.05	.00	.10	.11	.35
Between Wave	MQSE19	MQSE17	.18	.04	.00	.34	.07	.00
	MQSE17	MQSE13	.08	.05	.10	.10	.09	.27
	MQSE13	MQSE12	.28	.04	.00	.34	.08	.00
	E819WLE	M817WLE	-.14	.06	.02	-.39	.10	.00
	M817WLE	E813WLE	-.34	.07	.00	-.60	.09	.00
	E813WLE	M812WLE	-.16	.07	.02	-.25	.12	.04
Cross-Lagged	MQSE19	M817WLE	-.00	.04	.99	.07	.06	.28
	MQSE17	E813WLE	.02	.04	.59	-.16	.07	.03
	MQSE13	M812WLE	.02	.05	.64	.06	.07	.41
	E819WLE	MQSE17	.05	.04	.26	-.00	.08	.95
	M817WLE	MQSE13	.05	.07	.43	.11	.09	.20
	E813WLE	MQSE12	-.01	.04	.89	.01	.08	.91
Within Wave	MQSE19	E819WLE	.04	.03	.27	.04	.03	.27
	MQSE17	M817WLE	.03	.05	.53	.03	.05	.53
	MQSE13	E813WLE	-.00	.05	.93	.00	.05	.93
	MQSE12	M812WLE	-.02	.04	.59	-.02	.03	.60

Anmerkungen: Abkürzungen: M812 = Bildungsstandardüberprüfung Klasse 8 in Mathematik im Jahr 2012; M817 = Mathematikerhebung in 2017; E813 = Bildungsstandardüberprüfung in Englisch im Jahr 2013; E819 = Englischerhebung in 2019. WLE = Weighted Likelihood Estimate (Personenschätzer für die Leistung im Kompetenztest); MQSE = Durchschnittliche Anzahl (M) der Maßnahmen zur Qualitätssicherung und -entwicklung an der jeweiligen Schule; APS = Allgemeinbildende Pflichtschule, AHS = Allgemeinbildende höhere Schule.

6 Diskussion

Aus der Perspektive forschungsmethodologischer Erwägungen bietet die vorliegende Studie eine Reihe von Vorzügen und Qualitätseigenschaften, die für die Validität der Ergebnisse sprechen. Insbesondere erweitert die Studie die existierende Forschungslandschaft durch die Implementierung eines längsschnittlichen Untersuchungsdesigns auf Schulebene, welches einem Forschungsansatz nahesteht, der kausale Inferenz ermöglicht. Ein wesentliches Merkmal dieser Studie ist das umfassende Datenmaterial aus den Bildungsstandardüberprüfungen, realisiert durch eine Vollerhebung an österreichischen Schulen. Durch diesen längsschnittlichen Ansatz sind die Ergebnisse nicht nur weitgehend verallgemeinerbar, sondern weisen zudem eine die Zeit übergreifende bzw. messzeitpunktunabhängige Relevanz auf. Die differenzierte Erfassung der Schülerinnen- und Schülerleistungen durch domänenspezifische, objektivierte und standardisierte Leistungstests unterstreicht die methodische Strenge der Studie.

In der Zusammenschau der Befunde lässt sich jedoch trefflich sagen, dass sie mehr Fragen aufwerfen, als letztlich beantwortet werden. Für die QSE-Aktivitätsindexe finden sich keine Hinweise auf einen u-förmigen Zusammenhang zwischen Schülerinnen- und Schülerleistung und QSE-Maßnahmen. Schließlich zeigen die längsschnittlichen Analysen, dass QSE-Maßnahmen nur bedingt in Zusammenhang mit höherer Schülerinnen- und Schülerleistung stehen. Das gilt für beide gerechneten Modelle, insofern sind die Befunde in dieser Hinsicht konsistent. Allerdings konnte kein Effekt über die Zeit zwischen der Schülerinnen- und Schülerleistung einerseits und den QSE-Maßnahmen andererseits identifiziert werden. Der nicht statistisch nachweisbare Zusammenhang der QSE-Maßnahmen auf die Schülerinnen- und Schülerleistung (aggregiert auf Schulebene) im Zeitverlauf und die höhere Anzahl an durchschnittlich berichteten QSE-Maßnahmen bei Schulen mit geringerer Leistung können in zwei Richtungen interpretiert werden. Im positiven Sinne interpretiert bedeutet dies, dass leistungsschwache Schulen zu erkennen scheinen, dass sie etwas verändern müssen (oder sie werden von außerhalb z. B. durch die Schulaufsicht dazu aufgefordert). Dass sich aber viele potenzielle Effekte dieser Maßnahmen nicht binnen zwei bis drei Jahren entfalten können, liegt auf der Hand: Beispielsweise hat eine Schule den Wunsch nach Veränderung und führt eine interne Evaluation durch. Nachfolgend müssen die Ergebnisse entsprechend analysiert sowie Maßnahmen abgeleitet und umgesetzt werden. Bis die Maßnahmen sichtlich greifen können, vergeht zwangsläufig Zeit (Feldhoff & Radisch, 2021) – daher ist es auch wenig überraschend, wenn sich das in den Daten noch nicht (!) widerspiegelt. Im eher negativen Sinne könnte man die Ergebnisse auch dahingehend interpretieren, dass QSE-Maßnahmen auch über einen längeren Zeitraum hinweg nicht zu einer Leistungssteigerung von Schülerinnen und Schülern führen oder sich zumindest ein entsprechender Zusammenhang z. B. aufgrund zahlreicher vermittelnder Prozesse und Störvariablen nicht ohne Weiteres abbilden lässt.

In diesem Zusammenhang ist anzumerken, dass die im Rahmen der BIST-Ü erfragten QSE-Maßnahmen nicht das volle Spektrum möglicher Aktivitäten abbilden und ebenso wenig etwas über die Umsetzungsqualität aussagen. Über die eingesetzten Items wurden

zwar Maßnahmen im Bereich der Personalentwicklung (Bedarfsanalyse und Konzept der Fort- und Weiterbildung an der Schule, Mentoring für Berufseinsteigerinnen und Berufseinsteiger) wie auch der Organisationsentwicklung (interne/externe Evaluation, Schulentwicklungsgruppe) erfasst. Jedoch können zu Maßnahmen im Bereich der Unterrichtsentwicklung hingegen nur eingeschränkt Aussagen getroffen werden (Schülerinnen- und Schüler-Feedback zum Unterricht). Letzteres beschreibt allerdings Rolff (2018) als tragendes Element schulischer Entwicklungsprozesse. Gezielt auf Unterrichtsentwicklung fokussierende Maßnahmen wären etwa professionelle Lerngemeinschaften (z. B. Kansteiner, Stamann, Buren & Theurl, 2020), kollegiale Hospitationen (z. B. Kreis, 2015) oder Classroom Walkthroughs (z. B. Protheroe, 2009). Überdies liegen keine Informationen zur Qualität und Intensität der einzelnen Maßnahmen vor. Im Kontext der Schulentwicklungsforschung wird beispielsweise davon gesprochen, dass eine Maßnahme zunächst initiiert, dann implementiert und schließlich institutionalisiert werden muss, damit Veränderungen auf Schulebene überhaupt sichtbar werden (Holtappels, 2013). Anhand der verfügbaren Daten lassen sich allerdings keine Rückschlüsse dahingehend ziehen, in welchem Ausmaß von Schulleitungen berichtete QSE-Maßnahmen fester Bestandteil der Schulkultur sind. Darüber hinaus liegt es nahe, dass einzelne QSE-Maßnahmen ineinandergreifen oder sich aufeinander beziehen. Zum Beispiel könnte eine Schule auf Basis der BIST-Ü-Ergebnisse die Notwendigkeit für Unterrichtsentwicklung ausmachen. In einem ersten Schritt könnte die Schulleitung eine interne/externe Evaluation für konkretere Hinweise initiieren. Befragungen der Schülerinnen und Schüler zum Unterricht könnten wiederum ein Element der Evaluation sein. Die Evaluationsergebnisse könnten in der Folge Fortbildungsbedarfe aufzeigen. Gemeinsam mit einer neu ins Leben gerufenen Schulentwicklungsgruppe wird schließlich ein Fortbildungskonzept erarbeitet. Begleitet könnte der Prozess etwa durch die zuständige Schulqualitätsmanagerin oder den zuständigen Schulqualitätsmanager bzw. die zuständige Schulentwicklungsberaterin oder den zuständigen Schulentwicklungsberater werden usw. Es handelt sich also unverkennbar um komplexe Prozesse, und deren Entfaltung z. B. der Schülerinnen- und Schülerleistung auf Schulebene erfordert Zeit (Feldhoff & Radisch, 2021). Dies lässt sich als Plädoyer dafür werten, weiterhin systematisch den Zusammenhang zwischen unterschiedlichen QSE-Maßnahmen, deren Umsetzungsqualität und der Schul- bzw. Schülerinnen- und Schülerleistung zu untersuchen, v. a. mit Blick darauf, ob sich unter etwaigen veränderten Rahmenbedingungen deutlichere Zusammenhänge aufzeigen lassen oder eben weiterhin nicht.

An dieser Stelle ist ebenfalls auf methodische Limitationen aufgrund der verfügbaren Daten hinzuweisen: In künftigen Forschungsarbeiten sollte das erfasste Spektrum der QSE-Maßnahmen um aussagekräftigere Items erweitert werden. Ebenso sollte künftige Forschung auch die Qualität der realisierten QSE-Maßnahmen in den Blick nehmen – hierfür bieten sich insbesondere Mixed-Methods-Zugänge an. Ferner gilt es zu untersuchen, inwiefern im Rahmen etwaiger Schul- und Unterrichtsentwicklungsmaßnahmen externe Expertise, z. B. in Form von Schulentwicklungsberatung (Altrichter et al., 2021), in Anspruch genommen wurde. Ein weiterer kritischer Aspekt der Studie ist die ausschließliche Fokussierung auf direkte Zusammenhänge zwischen QSE-Maßnahmen und Schülerinnen- und Schülerleistung, ohne die theoretische Plausibilität zu berücksichtigen, dass diese Maßnahmen

eher als distale, denn als proximale Faktoren fungieren. Vielmehr ist anzunehmen, dass sie indirekt über die Unterrichtsqualität und das Lernverhalten der Schülerinnen und Schüler (etwa gemäß der Context-Input-Process-Output-Logik, vgl. Ditton, 2017) auf die Schülerinnen- und Schülerleistung einwirken. Beispielsweise finden sich in der Literatur Hinweise auf einen moderierenden Effekt des erlebten kollegialen Vertrauens und des organisationalen Lernens auf die Schülerinnen- und Schülerleistung (Groß Ophoff et al., 2023). Auch diesem Aspekt sollte in künftiger Forschung mehr Aufmerksamkeit zuteil werden.

Ungeklärt bleibt im Zusammenhang mit der vorliegenden Studie außerdem, wie die „angeordneten“ Entwicklungsprozesse an den Schulen tatsächlich realisiert werden, zumal im Zuge der Bildungsreformgesetzgebung 2017 die Bemühungen um schulische Qualitätssicherung und -entwicklung weiter ausgebaut und institutionalisiert wurden. Beispielsweise wurde zu Beginn des Schuljahrs 2021/22 österreichweit ein gemeinsames Qualitätsmanagementsystem für Schulen (QMS) eingeführt. Als Referenz dient seit 2020 (und damit nach der letzten BIST-Ü im Jahr 2019) der Qualitätsrahmen für Schulen (BMBWF, 2020). Die darin beschriebenen Qualitätsdimensionen und Qualitätsbereiche bilden eine verbindliche Grundlage für die schulische Arbeit und liefern zugleich den inhaltlichen Rahmen für den regelmäßig zu erstellenden Schulentwicklungsplan. Dieser sollte sämtliche Entwicklungsvorhaben der Schule berücksichtigen und von der Schulleitung gemeinsam mit dem Kollegium (ggf. einer Schulentwicklungsgruppe) ausgearbeitet werden. In dem mindestens alle drei Jahre stattfindenden Bilanz- und Zielvereinbarungsgespräch zwischen Schulaufsicht und Schulleitung werden die im QMS dokumentierten Inhalte (z. B. Schulentwicklungsplan, Daten aus dem Bildungsmonitoring, Q-Handbuch) in den Blick und ggf. zum Anlass für QSE-Maßnahmen genommen (BMBWF, 2021). Außerdem wurde infolge der Bildungsreformgesetzgebung 2017 mit dem Schuljahr 2021/22 eine Neukonzeption der Überprüfung durch die sogenannte individuelle Kompetenzmessung PLUS (iKM^{PLUS}) initiiert, die vergleichbar zu den Vergleichsarbeiten in Deutschland (Groß Ophoff, 2019) eine jährlich verpflichtende, österreichweite Überprüfung der Bildungsstandards auf der dritten, vierten, siebten und achten Schulstufe vorsieht. Auf Systemebene werden die Ergebnisse im Dreijahresrhythmus publiziert (Opriessnig, Neubacher, Bruneforth & Mayer, 2024). Auch die sog. Kompetenzorientierung im Unterricht wird durch weitere Vorgaben noch stärker institutionalisiert als bisher, z. B. in Form neuer Lehrpläne ab dem Schuljahr 2023/24. Diese fokussieren auf die Entwicklung von fachlichen und überfachlichen Kompetenzen, präzisiert durch Kompetenzraster, zur Verfügung gestellte Lernaufgaben und abgestimmte Leistungsbeurteilung (vgl. BMBWF, 2022).

Die genannten Reformbemühungen sind Teil des sogenannten Pädagogik-Pakets, das darauf abzielt, Kinder und Jugendliche durch Bildung auf eine immer komplexer werdende Welt vorzubereiten (BMBWF, 2022). Und mit Blick auf den beispielsweise von Pietsch, Aydin und Gümüş (2023) thematisierten Rechenschaftslegungsdruck ist hervorzuheben, dass die verschiedenen genannten Maßnahmen auch in Österreich als „low-stakes“ intendiert sind, also die resultierenden Ergebnisse v. a. als Entwicklungsanreize für die Schulen dienen sollen (Jesacher-Roessler & Kemethofer, 2022). Zugleich zeichnet sich das österreichische

Bildungssystem durch eine komplexe Governance-Struktur mit hoher Regelungsdichte aus (Brauckmann-Sajkiewicz et al., 2022), das eine Wahrnehmung von Kontrolle an den Schulen durchaus befördern kann und möglicherweise individualisierten Entwicklungsbestrebungen entgegenwirkt. In diesem Zusammenhang gilt es zu untersuchen, inwiefern die zunehmende Betonung der Accountability-Komponente in den genannten Reformen als kontrollierend wahrgenommen und sich somit sogar abträglich auf die Bereitschaft, schulische und unterrichtliche Entwicklungsprozesse zu initiieren oder gar zu verstetigen, auswirkt. Entsprechende Hinweise werden für verschiedene Maßnahmen und Akteure im Zusammenhang mit neuer, d. h. evidenzbasierter Steuerung berichtet (für einen Überblick z. B. Groß Ophoff & Cramer, 2022), was sich u. a. durch die Selbstbestimmungstheorie von Deci und Ryan (Deci & Ryan, 1987; Ryan & Sapp, 2005) erklären lässt.

Literatur

- Abdallah, A. K. & Alkaabi, A. M. (2023). Induction programs' effectiveness in boosting new teachers' instruction and student achievement: A critical review. *International Journal of Learning, Teaching and Educational Research*, 22 (5), 493–517.
- Aiken, L. S. & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. (S. xi, 212). Newbury Park, CA: Sage Publications, Inc.
- Altrichter, H. & Kemethofer, D. (2016). Stichwort: Schulinspektion. *Zeitschrift für Erziehungswissenschaft*, 19 (3), 487–508. doi:10.1007/s11618-016-0688-0
- Altrichter, H., Kemethofer, D. & Schmidinger, E. (2013). Neue Schulinspektion und Qualitätsmanagement im Schulsystem. *Erziehung und Unterricht*, 163 (9-10), 961–978.
- Altrichter, H., Krainz, U., Kemethofer, D., Jesacher-Rößler, L., Hautz, H. & Brauckmann-Sajkiewicz, S. (2021). Schulentwicklungsberatung und Schulentwicklungsberatungsforschung. In Bundesministerium für Bildung und Forschung (BMBWF) (Hrsg.), *Nationaler Bildungsbericht Österreich 2021, Teil 3. Ausgewählte Entwicklungsfelder* (S. 375–421). doi:10.17888/nbb2021
- 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.
- Basma, B. & Savage, R. (2018). Teacher professional development and student literacy growth: A systematic review and meta-analysis. *Educational Psychology Review*, 30 (2), 457–481. doi:10.1007/s10648-017-9416-4
- Blank, R. K. & de las Alas, N. (2009). *The effects of teacher professional development on gains in student achievement: How meta analysis provides scientific evidence useful to education leaders*. Council of Chief State School Officers. Verfügbar unter: <https://eric.ed.gov/?id=ED544700>

- Bonsen, M. (2016). Schulleitung und Führung in der Schule. In H. Altrichter & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (S. 301–323). Wiesbaden: Springer.
- Brauckmann-Sajkiewicz, S., Lassnigg, L., Feldhoff, T., Maag Merki, K., Jude, N. & Radisch, F. (2022). Über österreichische Schulleiter*innen, die „Veränderung“ können (sollen) – erste Folgerungen aus der S-CLEVER-Studie. In G. Fenkart, G. Khan-Svik, K. Krainer & N. Maritzen (Hrsg.), *Die Kunst des Widerstands. Festschrift für Marlies Krainz-Dürr* (S. 113–125). Innsbruck: StudienVerlag.
- Breit, S. & Schreiner, C. (Hrsg.). (2016). *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung*. Wien: Facultas.
- Bundesinstitut BIFIE. (2020). *Standardüberprüfung 2019. Englisch, 8. Schulstufe. Bundesergebnisbericht*. Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/b2f25d7876f79a9aae779089a6a2eda898fe41a2/BiSt_UE_E8_2019_Bundesergebnisbericht.pdf
- Bundesministerium für Bildung, Wissenschaft und Forschung (BMBWF) (2019). *Schulleitungsprofil. Eine praxisbezogene Orientierung für effektives Schulleitungshandeln*. Verfügbar unter: https://www.bmbwf.gv.at/dam/jcr:1ed495ba-5df0-4b57-b0fb-54cd121143f6/schulleitungsprofil_a4_bf.pdf
- Bundesministerium für Bildung, Wissenschaft und Forschung (BMBWF) (2020). *Der Qualitätsrahmen für Schulen*. https://www.qms.at/images/Qualitaetsrahmen_fuer_Schulen.pdf.
- Bundesministerium für Bildung, Wissenschaft und Forschung (BMBWF) (Hrsg.). (2021). *Nationaler Bildungsbericht Österreich 2021*. Wien. doi:10.17888/nbb2021-3-4
- Bundesministerium für Bildung, Wissenschaft und Forschung (BMBWF) (2022). *Das Pädagogik-Paket. Zeitgemäß. Transparent. Fair*. <https://www.bmbwf.gv.at/dam/jcr:326dfad6-a8b9-4e56-9d67-b4bdcc343bb1/pb.pdf>
- Bundesministerium für Unterricht, Kunst und Kultur & Bundesministerium für Wissenschaft und Forschung (2010). *LehrerInnenbildung NEU. Die Zukunft der pädagogischen Berufe. Die Empfehlungen der ExpertInnengruppe*. Verfügbar unter: https://www.qsr.or.at/dokumente/1870-20140529-092820-Empfehlungen_der_ExpertInnengruppe_Endbericht_092010_2_Auflage.pdf
- Chabera, B. (2021). Externe Schulevaluation als Instrument datengestützter Qualitätsentwicklung – Grundlagen und Perspektiven. *Erziehung und Unterricht*, 171 (9–10), 82–89.
- Deci, E. L. & Ryan, R. M. (1987). The support of autonomy and the control of behavior. *Journal of personality and social psychology*, 53 (6), 1024–1037. doi:10.1037/0022-3514.53.6.1024
- Ditton, H. (2017). Zum Wirkungszusammenhang der schulischen Handlungsebenen. In U. Steffens, K. M. Merki & H. Fend (Hrsg.), *Schulgestaltung: Aktuelle Befunde und Perspektiven der Schulqualitäts- und Schulentwicklungsforschung. Grundlagen der Qualität von Schule 2* (S. 59–84). Münster: Waxmann.
- Döbert, H. & Weishaupt, H. (2012). Bildungsmonitoring. In A. Wacker, U. Maier & J. Wissinger (Hrsg.), *Schul- und Unterrichtsreform durch ergebnisorientierte Steuerung: Empirische Befunde und forschungsmethodische Implikationen* (S. 155–173). Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/978-3-531-94183-7_7

- Eder, F. & Altrichter, H. (2009). Qualitätsentwicklung und Qualitätssicherung im österreichischen Schulwesen: Bilanz aus 15 Jahren Diskussion und Entwicklungsperspektiven für die Zukunft. In W. Specht (Hrsg.), *Nationaler Bildungsbericht Österreich 2009, Band 2. Fokussierte Analysen bildungspolitischer Schwerpunktthemen* (S. 305–322). Graz: Leykam.
- Fauth, B., Decristan, J., Rieser, S., Klieme, E. & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, 29, 1–9. doi:10.1016/j.learninstruc.2013.07.001
- Feldhoff, T. (2011). *Schule organisieren*. Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/978-3-531-93384-9
- Feldhoff, T. & Radisch, F. (2021). Why must everything be so complicated? Demands and challenges on methods for analyzing school improvement processes. In A. Oude Groote Beverborg, T. Feldhoff, K. Maag Merki & F. Radisch (Hrsg.), *Concept and design developments in school improvement research: Longitudinal, multilevel and mixed methods and their relevance for educational accountability* (S. 9–26). Cham: Springer International Publishing. doi:10.1007/978-3-030-69345-9
- Fend, H. (2011). Die Wirksamkeit der Neuen Steuerung – theoretische und methodische Probleme ihrer Evaluation. *Zeitschrift für Bildungsforschung*, 1 (1), 5–24. doi:10.1007/s35834-011-0003-3
- Finney, S. J. & DiStefano, C. (2006). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Hrsg.), *Structural equation modeling: A second course (Quantitative methods in education and the behavioral sciences)* (S. 269–314). Charlotte, NC: Information Age Publishing, Inc.
- Gärtner, H. (2013). Praxis und Verhältnis interner und externer Evaluation im Schulsystem im internationalen Vergleich. *Zeitschrift für Erziehungswissenschaft*, 4 (16), 693–712.
- Gärtner, H., Wurster, S. & Pant, H. A. (2014). The effect of school inspections on school improvement. *School Effectiveness and School Improvement*, 25 (4), 489–508. doi:10.1080/09243453.2013.811089
- Groß Ophoff, J. (2019). Vergleichsarbeiten. In E. Kiel, B. Herzig, U. Maier & U. Sandfuchs (Hrsg.), *Handbuch Unterrichten an allgemeinbildenden Schulen* (S. 434–443). Bad Heilbrunn: UTB Klinkhardt.
- Groß Ophoff, J., Brown, C. & Helm, C. (2023). Do pupils at research-informed schools actually perform better? Findings from a study at English schools. *Frontiers in Education*, 7. Verfügbar unter: <https://www.frontiersin.org/articles/10.3389/educ.2022.1011241>
- Groß Ophoff, J. & Cramer, C. (2022). The engagement of teachers and school leaders with data, evidence and research in Germany. In C. Brown & J. R. Malin (Hrsg.), *The emerald international handbook of evidence-informed practice in education* (S. 175–196). Bingley: Emerald Publishing Limited.
- Groß Ophoff, J. & Kemethofer, D. (2023). Von Österreich lernen? Neue Reformen und ihre Herausforderungen. *Schulmagazin* 5–10, 9–10, 14–17.
- Grünkorn, J., Klieme, E. & Stanat, P. (2019). Bildungsmonitoring und Qualitätssicherung. In O. Köller, M. Hasselhorn, F. W. Hesse, K. Maaz, J. Schrader, H. Solga et al. (Hrsg.), *Das Bildungswesen in Deutschland. Bestand und Potenziale* (S. 263–298). Bad Heilbrunn: Julius Klinkhardt.

- Hallquist, M. N. & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in M plus. *Structural equation modeling: a multidisciplinary journal*, 25 (4), 621–638.
- Hamaker, E. L., Kuiper, R. M. & Grasman, R. P. (2015). A critique of the cross-lagged panel model. *Psychological methods*, 20 (1), 102.
- Hattie, J. (2024). *Visible learning 2.0* (S. Wernke & K. Zierer, Übers.; 2. Aufl.). Bielefeld: Schneider.
- Herbert, B., Fischer, J. & Klieme, E. (2022). How valid are student perceptions of teaching quality across education systems? *Learning and Instruction*, 82, 101652. doi:10.1016/j.learninstruc.2022.101652
- Holtappels, H. G. (2013). Innovation in Schulen – Theorieansätze und Forschungsbefunde zur Schulentwicklung. In M. Rürup & I. Bormann (Hrsg.), *Innovationen im Bildungswesen* (S. 45–69). Wiesbaden: Springer Fachmedien. doi:10.1007/978-3-531-19701-2_2
- Holtappels, H. G., van Ackeren, I., Bremm, N. & Rutter, S. (2021). Schulen in herausfordernden Lagen entwickeln und unterstützen. Fazit und Empfehlungen aus dem Projekt „Potenziale entwickeln – Schulen stärken“. In I. van Ackeren, H. G. Holtappels, N. Bremm & A. Hillebrand-Petri (Hrsg.), *Schulen in herausfordernden Lagen – Forschungsbefunde und Schulentwicklung in der Region Ruhr. Das Projekt „Potenziale entwickeln – Schulen stärken“* (S. 333–344). Weinheim: Beltz Juventa. doi:10.25656/01:21387
- Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6 (1), 1–55.
- Ingersoll, R. M. & Strong, M. (2011). The impact of induction and mentoring programs for beginning teachers: A critical review of the research. *Review of Educational Research*, 81 (2), 201–233. doi:10.3102/0034654311403323
- Jesacher-Roessler, L. & Kemethofer, D. (2022). Evidence-informed practice in Austrian education. In C. Brown & J. R. Malin (Hrsg.), *The emerald handbook of evidence-informed practice in education* (S. 317–334). Bingley: Emerald Publishing Limited. doi:10.1108/978-1-80043-141-620221035
- Kansteiner, K., Stamann, C., Buhren, C. G. & Theurl, P. (Hrsg.). (2020). *Professionelle Lerngemeinschaften als Entwicklungsinstrument im Bildungswesen*. Weinheim: Beltz Juventa.
- Kemethofer, D. (2015). *Wie wirken Schulinspektionen? Empirische Analysen zu den Wirkungsmechanismen von Schulinspektionen* (Unveröffentlichte Dissertation). Johannes Kepler Universität Linz. Verfügbar unter: <https://epub.jku.at/obvulihs/download/pdf/963650?originalFilename=true>
- Kemethofer, D., Zuber, J., Helm, C., Demski, D. & Rie, C. (2015). Effekte von Schulentwicklungsmaßnahmen auf Schülerinnen- und Schülerleistungen im Fach Mathematik. *SWS-Rundschau*, 55 (1), 26–47.
- Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M. et al. (2003). *Zur Entwicklung nationaler Bildungsstandards*. Bonn: Bundesministerium für Bildung und Forschung (BMBF).
- Klieme, E. & Tippelt, R. (2008). Qualitätssicherung im Bildungswesen. Eine aktuelle Zwischenbilanz. In E. Klieme & R. Tippelt (Hrsg.), *Qualitätssicherung im Bildungs-*

- wesen. *Eine aktuelle Zwischenbilanz*. 53. Beiheft der Zeitschrift für Pädagogik (S. 7–13). Weinheim: Beltz. Verfügbar unter: <https://www.pedocs.de/volltexte/2013/7266/>
- Köller, O. (2010). Bildungsstandards. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch Bildungsforschung* (S. 529–548). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Kreis, A. (2015). Kollegiale Hospitation – Chancen und Realisierungsmöglichkeiten. In K. Kansteiner & C. Stamann (Hrsg.), *Personalentwicklung in der Schule zwischen Fremdsteuerung und Selbstbestimmung* (S. 185–199). Bad Heilbrunn: Klinkhardt.
- Kulmhofer-Bommer, A. & Diekmann, N. (2021). Qualitätsentwicklung im österreichischen Schulsystem – Kompetenzorientierung als Leitkonzept und dessen Implementierung. In Bundesministerium für Bildung, Wissenschaft und Forschung (BMBWF) (Hrsg.), *Nationaler Bildungsbericht Österreich 2021, Teil 3. Ausgewählte Entwicklungsfelder* (S. 425–470). Wien: BMBWF. doi:10.17888/nbb2021-3
- Lüdtke, O. & Robitzsch, A. (2022). A comparison of different approaches for estimating cross-lagged effects from a causal inference perspective. *Structural Equation Modeling: A Multidisciplinary Journal*, 29 (6), 888–907. doi:10.1080/10705511.2022.2065278
- Maaz, K., Emmrich, R., Kropf, M. & Gärtner, H. (2019). Bildungsstandards als innovative Elemente moderner Bildungssysteme. In J. Zuber, H. Altrichter & M. Heinrich (Hrsg.), *Bildungsstandards zwischen Politik und schulischem Alltag* (S. 25–44). Wiesbaden: Springer. doi:10.1007/978-3-658-22241-3_2
- Muthén, L. K. & Muthén, B. O. (2017). MPlus Version 8. [Software]. Muthén & Muthén.
- Opriessnig, S., Neubacher, M., Bruneforth, M. & Mayer, S. (2024). Das Wesentliche im Blick haben: Mit nachhaltigen Beobachtungen der Grundkompetenzen Schul- und Qualitätsentwicklung stützen – Die iKM^{PLUS} Zyklusberichte. *Erziehung und Unterricht*, 174 (5–6), 54–61.
- Pietsch, M., Aydin, B. & Gümüş, S. (2023). Putting the instructional leadership–student achievement relation in context: A meta-analytical big data study across cultures and time. *Educational Evaluation and Policy Analysis*. doi:10.3102/01623737231197434
- Protheroe, N. (2009). Using classroom walkthroughs to improve instruction. *Principal*, 88 (4), 30–34.
- Qualitätssicherungsrat (2014). *Pädagoginnen- und Pädagogenbildung NEU. Basisinformationen mit besonderem Fokus auf die Studienarchitektur der allgemeinbildenden Sekundarstufe*. Verfügbar unter: https://www.qsr.or.at/dokumente/1863-20140610-133004-Basisinformationen_mit_besonderem_Fokus_auf_die_Studienarchitektur_der_allgemeinbildenden_Sekundarstufe_QSR_Maerz_2014.pdf
- Röhl, S. (2021). Effects of student feedback on teaching and classes: An overview and meta-analysis of intervention studies. In W. Rollett, H. Bijlsma & S. Röhl (Hrsg.), *Student feedback on teaching in schools: Using student perceptions for the development of teaching and teachers* (S. 139–156). Cham: Springer International Publishing. doi:10.1007/978-3-030-75150-0_9
- Rolff, H.-G. (2018). Die Rolle der Schulleitung in Schulentwicklungsprozessen. In C. G. Bühren & H.-G. Rolff (Hrsg.), *Handbuch Schulentwicklung und Schulentwicklungsberatung* (S. 203–218). Weinheim: Beltz.

- Ryan, R. M. & Sapp, A. (2005). Zum Einfluss testbasierter Reformen: High stakes testing (HST). Motivation und Leistung aus der Sicht der Selbstbestimmungstheorie. *Unterrichtswissenschaft*, 33 (2), 143–159.
- Schober, B., Schultes, M.-T., Kollmayer, M. & Lüftenegger, M. (2019). Implementierung von Reformen im Bildungsbereich. In S. Breit, F. Eder, K. Krainer, C. Schreiner, A. Seel & C. Spiel (Hrsg.), *Nationaler Bildungsbericht Österreich 2018, Band 2. Fokussierte Analysen und Zukunftsperspektiven für das Bildungswesen* (S. 455–484). Graz: Leykam. doi:10.17888/nbb2018-2-11
- Schratz, M., Wiesner, C., Rößler, L., Schildkamp, K., George, A. C., Hofbauer, C. et al. (2019). Möglichkeiten und Grenzen evidenzorientierter Schulentwicklung. In S. Breit, F. Eder, K. Krainer, C. Schreiner, A. Seel & C. Spiel (Hrsg.), *Nationaler Bildungsbericht Österreich 2018, Band 2. Fokussierte Analysen und Zukunftsperspektiven für das Bildungswesen* (S. 403–454). Graz: Leykam. doi:10.17888/nbb2018-2-10
- Schreiner, C. & Breit, S. (Hrsg.). (2012). *Standardüberprüfung 2012. Mathematik, 8. Schulstufe. Bundesergebnisbericht*. Salzburg: IQS. Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/854e7e8b9e19d93e24de67589369acb34e0b849c/BiSt-UE_M8_2012_Bundesergebnisbericht.pdf
- Schreiner, C. & Breit, S. (Hrsg.). (2013). *Standardüberprüfung 2013 Englisch, 8. Schulstufe. Bundesergebnisbericht*. Salzburg: IQS. Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/f0043f3c93d4019fdb7b26c1f4aebd00d88230cf/BiSt-UE_E8_2013_Bundesergebnisbericht_gesamt.pdf
- Schreiner, C., Breit, S., Pointinger, M., Pacher, K., Neubacher, M. & Wiesner, C. (Hrsg.). (2017). *Standardüberprüfung 2017. Mathematik, 8. Schulstufe. Bundesergebnisbericht*. Salzburg: IQS. Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/4a28609fd6414dbc257274688ffa37e44e4a3cf7/BiSt-UE_M8_2017_Bundesergebnisbericht.pdf
- Senden, B., Nilsen, T. & Teig, N. (2023). The validity of student ratings of teaching quality: Factorial structure, comparability, and the relation to achievement. *Studies in Educational Evaluation*, 78, 101274. doi:10.1016/j.stueduc.2023.101274
- Sims, S., Fletcher-Wood, H., O'Mara, A., Cottingham, S., Stansfield, C., Herwegen, J. V. & Anders, J. (2021). *What are the characteristics of effective teacher professional development? A systematic review and meta-analysis*. London: Education Endowment Foundation. Verfügbar unter: <https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews/teacher-professional-development-characteristics>
- Specht, W. (2006). Von den Mühen der Ebene. Entwicklung und Implementation von Bildungsstandards in Österreich. In F. Eder, A. Gastager & F. Hofmann (Hrsg.), *Qualität durch Standards? Beiträge zum Schwerpunktthema der 67. Tagung der AEPF* (S. 13–37). Münster: Waxmann.
- Wurster, S., Richter, D. & Lenski, A. E. (2017). Datenbasierte Unterrichtsentwicklung und ihr Zusammenhang zur Schülerleistung. *Zeitschrift für Erziehungswissenschaft*, 20 (4), 628–650.

Anhang

Analysen auf Itemebene

Klassisches CLPM

Model	MQSE19on M817	MQSE19on M817_p	MQSE17on E813	MQSE17on E813_p	MQSE13on M812	MQSE13on M812_p	E819on MQSE17	E819on MQSE17_p	E817on MQSE13	E817on MQSE13_p	E813on MQSE12	E813on MQSE12_p
1	-0.040	0.221	-0.161	0.000	-0.011	0.749	-0.017	0.233	0.003	0.832	-0.009	0.545
2	-0.089	0.007	-0.168	0.000	0.086	0.015	-0.022	0.131	-0.013	0.360	0.004	0.767
3	0.030	0.362	-0.014	0.747	0.009	0.806	0.009	0.523	-0.013	0.364	-0.005	0.743
4	0.013	0.685	0.042	0.333	-0.002	0.959	0.008	0.592	0.020	0.153	-0.033	0.020
5	0.102	0.002	0.127	0.003	0.137	0.000	0.026	0.085	0.025	0.077	0.010	0.494
6	-0.029	0.351	-0.111	0.012	-0.072	0.028	0.033	0.030	0.003	0.798	0.014	0.322
7	-0.051	0.129	-0.026	0.573	-0.103	0.002	0.010	0.564	0.002	0.894	0.011	0.461

Model	N	Par	Chi2	DF	Chi2p	Chi2_df	CFI	TLI	RMSEA	SRMR
1	1.360	46	7.489	6	0.278	1.250	1.000	.998	.014	.007
2	1.360	46	5.380	6	0.496	0.900	1.000	1.001	.000	.005
3	1.360	46	5.524	6	0.479	0.920	1.000	1.001	.000	.005
4	1.360	46	6.470	6	0.373	1.080	1.000	.999	.008	.005
5	1.360	46	8.826	6	0.184	1.470	1.000	.997	.019	.006
6	1.360	46	14.958	6	0.021	2.490	.998	.991	.033	.010
7	1.360	46	4.432	6	0.618	0.740	1.000	1.002	.000	.004

Random Intercept CLPM

Model	MQSE13on M812	MQSE13on M812_p	MQSE17on E813	MQSE17on E813_p	MQSE19on M817	MQSE19on M817_p	E813on MQSE12	E813on MQSE12_p	E817on MQSE13	E817on MQSE13_p	E819on MQSE17	E819on MQSE17_p	WLEwMQSE	WLEwMQSE_p
1	0.029	0.524	-0.004	0.913	0.025	0.402	0.006	0.887	0.086	0.251	-0.030	0.305	-0.228	0.000
2	0.198	0.000	-0.031	0.414	-0.067	0.022	0.012	0.756	0.198	0.006	-0.054	0.080	-0.193	0.000
3	0.036	0.429	-0.006	0.870	-0.010	0.731	0.028	0.466	0.007	0.919	0.028	0.348	0.026	0.637
4	0.022	0.616	-0.012	0.734	0.035	0.217	0.033	0.468	0.085	0.259	0.011	0.709	-0.044	0.579
5	0.101	0.020	-0.030	0.401	0.050	0.092	-0.044	0.320	0.108	0.143	0.044	0.142	0.316	0.000
6	-0.060	0.195	0.090	0.018	-0.038	0.202	0.047	0.285	-0.057	0.454	0.080	0.011	-0.122	0.012
7	-0.098	0.024	0.017	0.636	-0.020	0.544	-0.024	0.610	-0.106	0.159	0.018	0.600	-0.227	0.000

Model	N	Par	Chi2	DF	Chi2p	Chi2_df	CFI	TLI	RMSEA	SRMR
1	1.360	37	141.377	15	0	9.43	.978	.947	.079	.030
2	1.360	37	151.643	15	0	10.11	.976	.942	.082	.033
3	1.360	37	133.284	15	0	8.89	.979	.949	.076	.028
4	1.360	37	141.471	15	0	9.43	.977	.945	.079	.029
5	1.360	37	133.817	15	0	8.92	.980	.952	.076	.028
6	1.360	37	142.087	15	0	9.47	.978	.948	.079	.031
7	1.360	37	137.617	15	0	9.17	.978	.948	.078	.029

Schulebene

Mechanismen der Klassenzusammensetzung und ihre Effekte auf die Schulleistung

1 Einleitung

Die Bildung von Schulklassen stellt einen wichtigen Prozess im Bildungswesen dar, da deren Auswirkungen für Schülerinnen und Schüler weitreichend sein können. Die Art und Weise, wie Schülerinnen und Schüler verschiedenen Klassen zugeordnet werden, kann die Lernumgebung und -erfahrungen der Schülerinnen und Schüler beeinflussen, was sich in der weiteren Folge auf ihre fachlichen Leistungen sowie ihre überfachliche Entwicklung auswirken kann. Dies belegen beispielsweise die Forschungsarbeiten zum Ability Grouping (z. B. Hattie, 2009), der Schul- und Klassenkomposition (z. B. Harker & Tymms, 2004) oder zum „Big Fish Little Pond“-Effekt (z. B. Marsh, Trautwein, Lüdtke, Baumert & Köller, 2007). Die Bildung von Schulklassen hat damit eine Bedeutung für pädagogische Praktiken und Bildungspolitik. Entsprechend verdienten Forschung zur Bildung von Klassen sowie die daraus resultierenden Effekte der Klassenzusammensetzung besondere Aufmerksamkeit. Während international betrachtet die Auswirkungen der Zusammensetzung von Schulklassen auf kognitive und nonkognitive Schüleroutcomes bereits häufig untersucht wurden (z. B. Hattie, 2009; Kulik & Kulik, 1992; Lou, Abrami, Spence, Poulsen, Chambers & d'Apollonia, 1996; siehe auch die Reviews von Braunsteiner, Fischer, Kernbichler, Prengel & Wohlhart, 2019, oder Becker, Neumann & Dumont, 2017), stellen Untersuchungen zu Mechanismen, die erklären, wie Schul- und Klassenzusammensetzungen entstehen, ein wenig bearbeitetes Forschungsfeld dar. Insbesondere für Österreich liegen zu den Mechanismen der Klassenbildung sowie zu den Effekten der Klassenzusammensetzung auf die Schülerleistung bisher nur wenige Arbeiten vor (Sertl & Leditzky, 2016; Weber & Kemethofer, 2024; Weber et al., 2019). Die vorliegende Studie zielt daher auf die Schließung dieser Forschungslücke ab. Wir wollen damit nicht nur das Forschungsfeld vorantreiben, sondern, soweit die Datenlage zulässt, auch empirische Evidenz und daraus ableitbare Empfehlungen für die Schulpraxis, insbesondere für Schulleitungen, liefern (z. B. welche Strategien der Klassenbildung mit negativen oder positiven Entwicklungen in der Schulleistung einhergehen können).

Aus theoretischer Perspektive wird im vorliegenden Beitrag argumentiert, dass die Bildung von Schulklassen von einer Vielzahl von Faktoren beeinflusst werden kann, die sowohl explizite Entscheidungen der Schulleitungen als auch implizite gesellschaftliche und bildungs-

politische Tendenzen umfassen (z. B. Sertl & Leditzky, 2016; Slavin, 1990; Weber et al., 2019). Diese Faktoren reichen von der örtlichen Herkunft und damit verbundenen Unterschieden im sozioökonomischen Hintergrund der Schülerinnen und Schüler bis hin zu pädagogischen Überlegungen wie der Leistungshomogenisierung von Lerngruppen. Diese Faktoren, die bei der Bildung von Klassen Berücksichtigung finden, können einerseits direkt auf die Leistungszusammensetzung der Klasse abzielen und somit intendiert zu einer leistungshomogenen oder -heterogenen Lerngruppe führen. Auf der anderen Seite können Entscheidungen bei der Bildung von Klassen auch nichtintendierte Effekte auf die leistungsbezogene Zusammensetzung der Klasse haben, sofern bei der Klassenbildung berücksichtigte Kriterien mit Leistung konfundiert sind. Solche nichtintendierten Effekte auf die leistungsbezogene Zusammensetzung können etwa entstehen, wenn die Muttersprache oder das Religionsbekenntnis bei der Klassenbildung berücksichtigt werden. Diese Faktoren können zwar die innerschulische Organisation erleichtern (z. B. muttersprachlicher Unterricht, Religionsunterricht), jedoch auch Leistungsunterschiede zwischen Klassen nach sich ziehen. Während für die intendierte Leistungshomogenisierung etwa im Rahmen der Forschung zum Ability Grouping eine längere Forschungstradition bereits einen Fundus an Befunden und Theorien erarbeitet hat, ist für Faktoren, die nicht explizit auf die Leistungshomogenisierung (oder auch -heterogenisierung) abzielen, jedoch nichtintendierte Effekte auf die leistungsbezogene Zusammensetzung von Klassen haben könnten, sowohl der Forschungsstand als auch der theoretische Diskussionsstand unterbelichtet. Die vorliegende Studie zielt daher darauf ab, die vielfältigen Mechanismen, die zur Klassenbildung führen, offenzulegen und die daraus resultierenden Effekte auf die Schulleistung zu analysieren.

Um ein umfassendes Verständnis der Mechanismen und Effekte der Klassenzusammensetzung zu erlangen, wird in diesem Beitrag auf Daten der Vollerhebungen im Rahmen der Bildungsstandardüberprüfungen in Österreich in den Jahren 2012 bis 2019 zurückgegriffen. Konkret fließen in unsere Analysen die fachlichen Schülerleistungen wie auch Angaben von Schulleitungen ein. In den Kontextfragebögen der Bildungsstandarderhebungen wurden die Schulleitungen gebeten, Auskunft über die von ihnen verwendeten Kriterien der Klassenzuweisung von Schülerinnen und Schülern bei Eintritt in die Sekundarstufe I zu berichten. Die Datenlage erlaubt, über einen Zeitraum von sieben Jahren (a) ein besonderes Augenmerk auf die unterschiedlichen Strategien und Kriterien zu legen, die Schulleitungen bei der Klassenzusammensetzung anwenden, und (b) zu prüfen, ob und in welchem Ausmaß die Art der Klassenzusammensetzung (homogene/heterogene Gruppenbildung) mit der Gesamtleistung der Schulen (d. h. der Abweichung der mittleren Schulleistung vom Erwartungswert) zusammenhängt. Um mögliche konfundierende Effekte auf die Schulleistung auszuschließen, wird im Beitrag die Schulleistung aus dem „fairen Vergleich“ verwendet, in welcher über Merkmale des Schulstandorts und der Schülerschaft kontrolliert wird (George & Robitzsch, 2015).

Der Beitrag gliedert sich in mehrere Abschnitte, beginnend mit einer detaillierten Betrachtung der verschiedenen Mechanismen der Klassenzusammensetzung. Darauf folgt eine Aufbereitung der Theorie und des internationalen Forschungsstands zu Auswirkungen von Klassenzusammensetzungen (Ability Grouping auf kognitive und nichtkognitive

Schülerinnen- und Schüler-Outcomes). Abschließend werden das Design, die Analysen und die Ergebnisse der vorgelegten Studie präsentiert und diskutiert, um Implikationen für die Bildungsforschung und -praxis abzuleiten.

2 Mechanismen der Klassenzusammensetzung

Auswirkungen der Zusammensetzung von Schulklassen auf kognitive und nonkognitive Schüleroutcomes stellen in der Bildungsforschung ein häufig untersuchtes Phänomen dar. Derartige Effekte werden bspw. in der Forschung zu „Kompositionseffekten“ oder dem „Big Fish Little Pond“-Effekt untersucht. Dagegen stellen Untersuchungen zu Mechanismen, die erklären, wie Schul- und insbesondere Klassenzusammensetzungen entstehen, laut Weber et al. (2019) ein nationales wie internationales Forschungsdesiderat dar. Weber et al. (2019) geben im Nationalen Bildungsbericht 2018 einen Überblick über mögliche Ursachen der zwischenschulischen Segregation (z.B. soziale und ethnische Zusammensetzung der Einzugsgebiete, Schulwahl durch Eltern, Aufhebung des Schulsprengels in Städten, Schulprofilbildung), wobei ähnliche Faktoren auch für die Klassenbildung eine Rolle spielen können (z. B. Elternwünsche, Bildung von Schwerpunktklassen). Darüber hinaus kommt hier jedoch den (expliziten) Entscheidungen der Schulleitungen, die auf unterschiedlichen Kriterien basieren und unterschiedlichen Handlungslogiken folgen (z. B. organisatorische Überlegungen wie etwa die Vermeidung von klassenübergreifendem Religionsunterricht, siehe etwa Weber et al., 2019), eine wichtige Rolle zu. So weist das vom BMBWF erstellte Schulleitungsprofil (BMBWF, 2019) darauf hin, dass die Klassen- und Lerngruppenbildung im Verantwortungsbereich der Schulleitung liegt (auch wenn entsprechende Entscheidungen mit dem Kollegium abgestimmt werden sollen). Auch die Studie von Sertl und Leditzky (2016) hebt (für die Primarstufe) die wesentliche Rolle der Schulleitungen hervor. Die dort berichteten Kriterien werden im Folgenden präsentiert – ergänzt um Kriterien, die für die Sekundarstufe (Bildung von leistungshomogenen Lerngruppen z. B. durch Leistungsgruppen) von Bedeutung sind.

2.1 Örtliche Herkunft der Schülerinnen und Schüler

Schulleitungen neigen dazu, Schülerinnen und Schüler aus denselben oder benachbarten Gemeinden und Stadtteilen derselben Klasse zuzuweisen, um bspw. gemeinsame Anreisen und die soziale Interaktion außerhalb der Schule (gemeinsame Freizeitgestaltung, Austausch von Lernmaterialien) zu ermöglichen. So berichten Sertl und Leditzky (2016), dass die Berücksichtigung der Wünsche der Eltern durch die Schulleitungen dazu führen kann, dass Kinder dieselbe Klasse wie ihre Freunde besuchen. Dieser Logik folgend kann es dazu kommen, dass in einer Klasse überproportional viele Schülerinnen und Schüler aus einer Region (z. B. von ländlichen Gebieten, einer Gemeinde, einem Stadtteil) in einer Schulklasse und in einer anderen Schulklasse überproportional viele Schülerinnen und Schüler aus einer anderen Region (z. B. aus städtischen Gebieten, einer anderen Gemeinde, einem anderen Stadtteil) zusammengefasst werden, wodurch das mittlere Leistungsniveau variieren kann, weil die soziale Segregation u. a. mit der Wohngegend (höhere Dichte sozioökonomisch

benachteiligter Familien im städtischen Gebiet bzw. in bestimmten Stadtteilen; Sertl & Leditzky, 2016; Vogtenhuber, Lassnigg & Bruneforth, 2012) assoziiert ist.

2.2 Fördern und Fordern ist aus Sicht der Schulleitungen in homogenen Gruppen einfacher

Ein häufig in der Literatur (vgl. Becker et al., 2017; Köller & Baumert, 2001) angeführtes Argument für die klassenweise Trennung von leistungsschwächeren und -stärkeren Schülerinnen und Schülern ist, dass die Förderung bzw. Forderung der Schülerinnen und Schüler in homogenen Lerngruppen leichter umgesetzt werden kann als in heterogenen Lerngruppen. So kann es sein, dass leistungshomogene Klassen gebildet werden, weil Schulleitungen befürchten, dass leistungsheterogene Klassen didaktisch herausfordernder sind (siehe dazu auch den Abschnitt zum Ability Grouping).

2.3 Möglichkeit der Bildung von Schwerpunktklassen

Die Bildung von Schwerpunktklassen als Teil der Schulprofilierung spielt ebenfalls eine Rolle. Unterschiedliches Wahlverhalten und Aufnahmekriterien führen zur Bildung dieser Klassen, was wiederum die soziale und leistungsmäßige Zusammensetzung beeinflusst (Altrichter, Heinrich & Soukup-Altrichter, 2014; Eder, 2011; Specht, 2011). Sertl und Leditzky (2016) beschreiben die entsprechende Dynamik wie folgt: Indem vor allem Eltern aus der bildungsbewussten Mittelschicht – und zwar auch aus anderen Schulsprengeln – dorthin drängen, kommt es zu Konkurrenz- und Verdrängungseffekten beim Wettbewerb um die wenigen Schulplätze an diesen Einrichtungen. Weber et al. (2019) konnten in diesem Zusammenhang feststellen, dass etwa die Berücksichtigung von fachlichen Neigungen/Schwerpunkten bzw. Schulversuchen bei der Klassenzusammenstellung auf der Primarstufe mit einer höheren innerschulischen sozialen Segregation – also Ungleichverteilung von Schülerinnen und Schülern unterschiedlicher sozioökonomischer Herkunft zwischen den Klassen von Schulen – einhergeht und auf der Sekundarstufe (sowohl AHS als auch NMS/HS) dieses Kriterium der Klassenzusammensetzung mit einer höheren innerschulischen ethnischen Segregation verbunden ist.

2.4 Möglichkeit der Bildung von Leistungsgruppen

Im Rahmen der NMS-Reform „Pädagogik-Paket“ (BMBWF, 2020) wurde Schulleitungen die Möglichkeit eingeräumt, an ihren Standorten in den Unterrichtsfächern Deutsch, Mathematik und Englisch dauerhafte Lerngruppen, getrennt nach zwei unterschiedlichen Leistungsniveaus („Standard“ bzw. „Standard AHS“) einzurichten. Solche Leistungsgruppen gab es bereits in der Hauptschule, dem Vorgänger der (Neuen) Mittelschule. Allerdings waren die Leistungsgruppen damals verpflichtend.

2.5 Bewusstes Zusammenfassen nach Religionsbekenntnis und Muttersprache

Sertl und Leditzky (2016) halten in ihrer qualitativen Studie zur Klassenzusammensetzung an Österreichs Volksschulen fest, dass das Religionsbekenntnis der Schülerinnen und Schüler ein weiteres Hauptkriterium für die Klassenzusammensetzung darstellt. In ihrer Studie argumentieren die Autorin und der Autor, dass diese religiöse Segmentierung dazu führen kann, dass Schulklassen gebildet werden, in denen Kinder, die Deutsch sprechen und eine Affinität zur Bildung aufweisen, in der Überzahl sind, während sie in anderen Klassen unterrepräsentiert sind. So stellten die Autorin und der Autor fest, dass an einigen der untersuchten Schulen Kinder, die dieselbe Muttersprache sprechen, absichtlich in denselben Klassen zusammengelegt wurden, um die Bildung von Sprachgruppen im Unterricht zu erleichtern. Allerdings gab es auch Schulleitungen, die explizit anders vorgegangen sind. Sie achteten bewusst darauf, Kinder, deren Muttersprache nicht Deutsch war, gleichmäßig auf alle Klassen zu verteilen. Dies sollte der Entstehung von milieuspezifischen Gruppierungen („Ghetto-Bildung“) unter den Schülerinnen und Schülern entgegenwirken. Die Analysen von Weber et al. (2019) deuten für die Volksschule darauf hin, dass die Klassenzusammensetzung nach Herkunft bzw. Muttersprache mit einer höheren innerschulischen ethnischen Segregation einhergeht.

2.6 Ungleichheitsreduzierende Bemühungen von Schulleitungen

Die qualitative Studie von Sertl und Leditzky (2016) zeigt aber auch, dass manche Schulleitungen bemüht sind, eine möglichst ausgewogene und faire Verteilung der Schülerinnen und Schüler zu erreichen, um bestehende Ungleichheiten abzufedern bzw. neuen Ungleichheiten vorzubeugen. Beispielsweise werden Kinder, die bereits den Vorschulunterricht besucht haben, sowie jene, die während der Aufnahmegespräche als unaufmerksam, noch nicht entwicklungsentsprechend oder leistungsschwach eingestuft wurden, gleichmäßig auf die verschiedenen ersten Klassen verteilt. Allerdings geben Sertl und Leditzky (2016) zu bedenken, dass fraglich ist, ob trotz dieser Bemühungen die Klassenbildung die oft bestehenden Unterschiede, die durch Faktoren wie Wohnort/-gegend und Schulprofil entstehen, kompensieren kann.

Die skizzierten Mechanismen legen nahe, dass es auch im österreichischen Schulwesen zu (sozialer und ethnischer) Segregation innerhalb von Schulen (also zwischen Klassen) kommt, die Auswirkungen auf den Unterricht und die kognitiven und nichtkognitiven Schüleroutcomes haben kann. Die diesbezügliche internationale Forschung findet sich insbesondere in der längeren Tradition des sogenannten Ability Groupings, die, wie der Name sagt, auf leistungsbezogene Kriterien der Schul- und Klassenzusammensetzung fokussiert und daher insbesondere für solche (und damit konfundierte) Kriterien theoretische Argumentationslinien bzw. Erklärungen über erwartete Effekte auf die fachlichen Schulleistungen bereithalten.

3 Argumentationslinien für die Erklärung von Effekten leistungsbezogener Kriterien der Klassenzusammensetzung auf die Schulleistung (Ability Grouping)

Ability Grouping (AG) beschreibt Maßnahmen zur Reduktion von Leistungsheterogenität innerhalb von Lerngruppen (z. B. Becker et al., 2017; Slavin, 1990). Die Gruppierung erfolgt oft auf Basis von Schülerleistungen, also Leistungen, die die Schülerinnen und Schüler in den Zubringerschulen erbracht haben, Intelligenzquotienten der Schülerinnen und Schüler oder Lehrerinnen- und Lehrer-Urteile und kann sich auf unterschiedliche Ebenen (zwischen Schultypen, innerhalb von Schulen, innerhalb von Klassen) beziehen. AG folgt der grundlegenden Idee, dass eine optimale Unterrichtsgestaltung erleichtert wird, wenn das Fähigkeitsniveau der Schülerinnen und Schüler homogener ist. Die Gruppierung von Schülerinnen und Schülern nach Fähigkeiten ist Merkmal vieler Schulsysteme weltweit – insbesondere auf der Sekundarstufe – und kann verschiedene Formen annehmen. Die Literatur (z. B. Slavin, 1990) unterscheidet zwischen strukturell-permanenten (between-class ability grouping, between-school tracking) und informell-temporären (within-class ability grouping) Maßnahmen:

- Zu den strukturell-permanenten Maßnahmen gehört das Tracking, z. B. die Bildung von permanenten Leistungsgruppen innerhalb von Schulen oder die Bildung von Schultypen und Schulprofilen. In Deutschland und Österreich bspw. erfolgt die Fähigkeitsgruppierung in Form eines Systems, das Schülerinnen und Schüler je nach ihrem Fähigkeitsniveau in separate Schultypen einteilt. Dieses Schultypen-Tracking gilt als die stärkste Form der Fähigkeitsgruppierung. Eine schwächere Form ist das fachbezogene Tracking oder Setting, bei dem Schülerinnen und Schüler innerhalb der Schule in fachspezifische Lerngruppen aufgeteilt werden, wobei sich diese Lerngruppen von Fach zu Fach unterscheiden.
- Zu den informell-temporären Maßnahmen gehört die Binnendifferenzierung, also die Gruppierung von Schülerinnen und Schülern für einzelne oder mehrere Unterrichtseinheiten.

Laut Slavin (1990) sind die strukturell-permanenten Maßnahmen primär auf der Sekundarstufe zu finden.

3.1 Theorien zu Effekten des Ability Groupings auf Schülerleistungen

Zur Beantwortung der Forschungsfrage nach den Effekten des AG auf Schülerleistungen ziehen wir im nächsten Schritt theoretische Überlegungen heran, die eine Antwort darauf geben, ob aus Sicht verschiedener etablierter Theorien das AG Vorteile oder gar Nachteile für das Lernen der Schülerinnen und Schüler nahelegt. Die nachfolgend dargestellten theoretischen Konzepte wurden primär im Kontext strukturell-permanenter Maßnahmen (insbesondere des Trackings) entwickelt und erforscht. In diesen Systemen dürften sich die beschriebenen Wirkmechanismen (z. B. der Big-Fish-Little-Pond-Effekt) besonders

entfalten bzw. stark ausgeprägt sein, da Schülerinnen und Schüler über längere Zeiträume in festgelegten Leistungsgruppen verbleiben, was die Effekte stabilisiert bzw. verstärkt. In informell-temporären Maßnahmen wie der Binnendifferenzierung könnten die beschriebenen Mechanismen zwar auch auftreten, jedoch wären sie wahrscheinlich weniger stark ausgeprägt, da die Gruppenzusammensetzungen flexibler sind und Schülerinnen und Schüler häufiger zwischen verschiedenen Leistungsniveaus wechseln können.

3.1.1 Die Annahme höherer Unterrichtsqualität

Theoretische Argumente für die Wirkungsweise von AG lassen sich in den von den Proponentinnen und Proponenten sowie den Opponentinnen und Opponenten des AG vorgebrachten Vor- und Nachteilen wiederfinden (Slavin, 1990, S. 473). Dabei beziehen sich die Argumente weitgehend auf die eingangs bereits erwähnte Grundidee des AG, die Ermöglichung einer höheren Unterrichtsqualität durch homogenere Schülergruppen (vgl. auch Becker et al., 2017; Köller & Baumert, 2001):

- *Höhere Unterrichtsqualität/mehr adaptiver Unterricht:* AG erleichtert für Lehrkräfte das Unterrichten. Es ermöglicht eine Anpassung der Unterrichtsmethode an die Bedürfnisse der Lerngruppe; bspw. individuellen Unterricht für kleine, langsame Gruppen. Die Gefahr, von Schülerinnen und Schülern zu viel oder zu wenig zu verlangen, wird vermieden.
- *Mehr individualisiertes Lernen:* AG ermöglicht Schülerinnen und Schülern individualisiertes Lernen, d. h., entsprechend ihren Fähigkeiten Lernfortschritte zu erzielen.
- *Mehr Kompetenzerlebnisse:* Durch vermehrte Individualisierung und adaptiven Unterricht erhöht AG Kompetenzerlebnisse und reduziert Misserfolge.
- *Höhere Beteiligung am Unterricht:* AG hilft, Interesse und Anreize aufrechtzuerhalten, da sich leistungsfähige Schülerinnen und Schüler durch die Rücksichtnahme und Beteiligung weniger begabter Schülerinnen und Schüler nicht langweilen. Umgekehrt nehmen langsamere Schülerinnen und Schüler stärker am Unterricht teil, wenn sie sich nicht von den leistungsstarken Schülerinnen und Schülern abgehängt fühlen.

Dem stehen u. a. folgende vermutete Nachteile gegenüber:

- *Wenig anregende Lernumgebung für leistungsschwache Lerngruppen:* Für leistungsschwächere Schülerinnen und Schüler ist der gemeinsame Unterricht und das gemeinsame Lernen mit leistungsstärkeren Schülerinnen und Schülern wichtig, da so ihr Lernen stimuliert und angeregt wird. In homogenen, leistungsschwachen Lerngruppen besteht die Gefahr, dass die soziale Lernumgebung wenig kognitiv aktivierend ist (Renkl, 1997).
- *Stigmatisierung von leistungsschwachen Lerngruppen:* Darüber hinaus haftet homogenen, leistungsschwächeren Klassen häufig ein Stigma an. Lehrpersonen können Einwände, Vorbehalte und Stereotype gegenüber den leistungsschwächeren Lerngruppen haben. Das führt dazu, dass die Schülerinnen und Schüler in diesen Klassen entmutigt werden.
- *Potenzial des adaptiven Unterrichts wird nicht genutzt:* Der Idee, dass AG zu mehr adaptivem Unterricht führt, wird entgegengehalten, dass Lehrerinnen und Lehrer

nicht in der Lage sind – weil ihnen die Kompetenzen, die Motivation und/oder die Zeit fehlen oder sie die Notwendigkeit nicht sehen –, den Unterricht an unterschiedliche Leistungsniveaus anzupassen.

- Schließlich können Unterschiede in der (materiellen und immateriellen) *Ressourcenverteilung* zu differenziellen Lern- und Entwicklungsmilieus beitragen, wobei hier die Zuteilung von Lehrkräften (und die damit verbundenen Unterschiede in der Lehrkraft- bzw. Unterrichtsqualität) in Abhängigkeit von der Klassenzusammensetzung eine Rolle spielen kann (Harker & Tymms, 2004; Specht, 2011; Weber et al., 2019).

3.1.2 Die Differentiation-Polarisation-Theorie

Eine Theorie zum AG ist die „Differentiation-Polarisation Theory“ von Hargreaves (1967). Entsprechend dieser Theorie spielt das AG eine wichtige Rolle in der Polarisierung der Einstellung der Schülerinnen und Schüler in „Pro- und Anti-Schul-Gruppen“. Die Annahme ist, dass leistungsschwache Schülerinnen und Schüler unter einem Statusverlust leiden und Gefühle des Scheiterns entwickeln. Dies deshalb, weil ihre Zuweisung zur leistungsschwachen Schülergruppe ihre geringe Leistungsfähigkeit explizit macht. Als Konsequenz entwickeln sie negative Einstellungen gegenüber Schule. Im Gegensatz dazu haben leistungsstarke Schülerinnen und Schüler häufiger positive Schulerfahrungen, da die Zugehörigkeit zur leistungsstarken Gruppe ihnen einen besonderen Status verleiht. Bei Harker und Tymms (2004) wird ein sehr ähnlicher Effekt als Peereffekt bezeichnet, der unabhängig von der leistungsbedingten Zuweisung dann vorliegt, wenn sich in Lerngruppen mit einem geringen Leistungsniveau durch soziale Interaktionen vermehrt geringere schulbezogene Motivationslagen unter den Schülerinnen und Schülern ausbilden. Studien von Abraham (1989) und Van Houtte (2006) bestätigen die Differentiation-Polarisation-Theorie bzw. die Peereffekte. Über die Zeit hinweg verstärkt das AG die Freundschaften, sodass es zu einer weiteren Polarisierung der gruppenspezifischen Einstellungen kommt. Das heißt, leistungstärkere Schülerinnen und Schüler werden noch enthusiastischer, während leistungsschwache Schülerinnen und Schüler mehr und mehr resignieren. Auch für Deutschland liegen Befunde vor, die auf Zusammenhänge zwischen der Veränderung der Klassenzusammensetzung, konkret der Peers, und der Entwicklung der Schülermotivation hindeuten (Reindl, 2021).

3.1.3 Die Labelling-Theorie

Eine weitere Theorie, die die negativen Effekte des AG zu erklären versucht, stellt die „Labelling Theory“ von Becker (1963) dar. Dieser Theorie folgend kann erwartet werden, dass Lehrkräfte dazu tendieren, Schülerinnen und Schüler entsprechend ihrer Leistungsfähigkeit zu labeln bzw. Stereotype zu verwenden. Schülerinnen und Schüler aus schwachen Leistungsgruppen werden häufiger als langsam, dumm oder schwierig „gelabelt“ und häufiger von Lehrkräften kritisiert als Schülerinnen und Schüler aus den leistungsstarken Gruppen. Das kann dazu führen, dass leistungsschwache Schülerinnen und Schüler sich unsicher und dumm fühlen und negative Einstellungen zur Schule sowie zu den Personen, die sie in diese Schubladen stecken, entwickeln. Darüber hinaus kann es zu Unterrichtseffekten kommen, wenn beispielsweise Lehrkräfte ihren Unterricht derart an die Klassen-

zusammensetzung anpassen, dass in Klassen mit einem geringeren Leistungsniveau weniger leistungsförderliche Aktivitäten gesetzt werden (Harker & Tymms, 2004).

3.1.4 Der Big-Fish-Little-Pond-Effekt

Schließlich stellt der Bezugsgruppeneffekt (engl.: „Big Fish Little Pond“-Effekt, BFLPE; Marsh et al., 2007) eine zentrale Theorie dar, die eine Erklärung für AG-Effekte, insbesondere im Hinblick auf die Entwicklung nichtkognitiver Schülerinnen- und Schüler-Outcomes wie des fachlichen Selbstkonzepts, und damit auch indirekt auf das fachliche Lernen der Schülerinnen und Schüler, bereithält. Der BFLPE beschreibt das Phänomen, dass Schülerinnen und Schüler mit gleicher Leistungsfähigkeit ein geringeres akademisches Selbstkonzept entwickeln, würden sie in leistungsstarken Gruppen/Schulklassen lernen, als es der Fall wäre, wenn sie in leistungsschwachen Gruppen/Schulklassen lernen würden. Der BFLPE repräsentiert den Nettoeffekt, der sich aus mehreren psychologischen internalen und externalen Vergleichsprozessen ergibt. Der interne Bezugsrahmen-Vergleich beschreibt einen psychologischen Prozess, bei dem Schülerinnen und Schüler ihre eigenen Leistungen in einem Bereich (z. B. Mathematik) mit denen in anderen Bereichen (z. B. Naturwissenschaften) vergleichen. Wenn die mathematischen Leistungen einer Schülerin oder eines Schülers höher sind als die naturwissenschaftlichen, dann wird das mathematische Selbstkonzept dieser Schülerin oder dieses Schülers höher sein als das naturwissenschaftliche, selbst wenn die mathematischen Leistungen der Schülerin bzw. des Schülers im Vergleich zu denen anderer Schülerinnen und Schülern nicht gut sind. Der externe Bezugsrahmen-Vergleich untergliedert in die „across-group“ und die „within-group comparison“. Im „across-group“-Vergleich stellt eine Schülerin oder ein Schüler ihre bzw. seine Leistungen jenen Leistungen von Schülerinnen und Schülern aus anderen Gruppen gegenüber. Im „within-group“-Vergleich dagegen vergleicht die Schülerin oder der Schüler ihre bzw. seine Leistungen mit jenen aus der eigenen Lerngruppe. Beide Vergleiche führen zu gegensätzlichen Urteilen über die eigenen fachlichen Leistungen. Entsprechend dieser Theorie (Richer, 1976) werden leistungsstarken Schülerinnen und Schülern in homogenen, leistungsstarken Klassen immer wieder ihre eigenen fachlichen Unzulänglichkeiten vor Augen geführt; nämlich dann, wenn sie sich mit den leistungsstarken, noch stärkeren Mitschülerinnen und Mitschülern vergleichen. Dies führt der Theorie nach dazu, dass diese Schülerinnen und Schüler immer unsicherer bezüglich ihrer eigenen Leistungsfähigkeit werden. Das hat negative Effekte auf das fachliche Leistungsselbstkonzept. Andererseits können sich diese leistungsstarken Schülerinnen und Schüler aus homogenen, leistungsstarken Gruppen mit schwächeren Schülerinnen und Schülern aus anderen homogenen, leistungsschwächeren Gruppen vergleichen und so wieder positive Gefühle, mehr Selbstsicherheit bzgl. ihrer eigenen fachlichen Leistungen erlangen („basking in reflected glory“-Effekt, kurz BIRGE, übersetzt: „Ruhm im Licht des Erfolgs anderer“; Liu, Wang & Parkins, 2005). Ähnliche Prozesse gelten für leistungsschwächere Schülerinnen und Schüler. Sie fühlen sich besser hinsichtlich ihrer eigenen Leistungen, wenn sie sich mit ihren eigenen Klassenkolleginnen und Klassenkollegen in leistungsschwachen Gruppen vergleichen. Wenn sie sich allerdings über ihre Lerngruppe hinaus mit Schülerinnen und Schülern aus leistungsstarken Klassen vergleichen, dann entwickeln sie Unsicherheiten hinsichtlich ihrer eigenen Leistungen.

In diesem Zusammenhang sind die Ergebnisse der Studie aus Kapitel 9 (Auer et al.) des vorliegenden Bands interessant. Sie zeigen für Österreich, dass der BFLP-Effekt auf der Sekundarstufe I stärker ausgeprägt ist als auf der Primarstufe.

Den skizzierten theoretischen Annahmen zur Wirkungsweise des AG folgen nun nationale und internationale Befunde der AG-Forschung, um zu zeigen, inwiefern sich die geschilderten Annahmen auch empirisch belegen lassen. Im ersten Schritt wird dabei auf Studien zu Auswirkungen auf die Unterrichtsqualität eingegangen, danach auf Studien zu Effekten auf die Schulleistung.

3.2 Internationale und nationale Befunde zu Effekten des Ability Groupings

Die nachfolgend berichteten Forschungsbefunde stammen in den meisten Fällen aus Studien, die strukturell-permanente Maßnahmen („between-class ability grouping“, „between-school tracking“) untersuchten. Studien zu Effekten der Binnendifferenzierung liegen nur vereinzelt vor.

3.2.1 Befunde zu AG-Effekten auf die Unterrichtsqualität

Slavin (1990, S. 474) verweist auf mehrere Studien, deren Befunde tendenziell darauf hinweisen, dass in homogenen, leistungsschwachen Schulklassen die Unterrichtsqualität signifikant geringer ausgeprägt ist als in homogenen, leistungsstarken Schulklassen. Gleichzeitig hebt Slavin die Herausforderung hervor, Unterrichtsqualität in diesem Setting zu bestimmen, da bspw. weniger durchgenommener Unterrichtsstoff in leistungsschwachen Klassen durchaus auch ein Qualitätskriterium im Sinne des adaptiven Unterrichts sein kann. Ältere Studien, beispielsweise von Dar und Resh (1986), zeigen, dass AG zwar zu einer ressourcenreicheren Lernumgebung für leistungsstarke Schülerinnen und Schüler führt, allerdings auch zu einer deprivierten Lernumgebung für leistungsschwache Schülerinnen und Schüler, weil die intellektuellen Kapazitäten der Mitschülerinnen und -schüler fehlen, die jedoch eine wichtige Lernressource darstellen. Zudem scheint der Unterricht in leistungsschwachen Klassen stärker fragmentiert zu sein und langsamer im Stoff voranzuschreiten. Im Gegensatz dazu haben Schülerinnen und Schüler in leistungsstarken Klassen eine höhere Wahrscheinlichkeit, Unterricht zu bekommen, der kritisches Denken und Problemlösen fördert (Oakes, 1985). Auch aktuellere Studien zur wahrgenommenen Unterrichtsqualität berichten von deutlichen Unterschieden zwischen Schulformen bzw. -typen, also der stärksten Formen des AG. Auch hier lassen sich Evidenzen für eine geringere Unterrichtsqualität finden. Zwar wird in den Schultypen mit durchschnittlich leistungsschwächeren Schülerinnen und Schülern (z. B. Hauptschule, Mittelschule) von einer höheren Schülerinnen- und Schülerzentrierung berichtet, gleichzeitig aber auch von deutlich mehr Disziplinstörungen, einem geringen Unterrichtstempo und einer stärkeren Überforderung im Unterricht (z. B. Baumert, Kunter, Brunner, Krauss, Blum & Neubrand, 2004; Clausen, 2002; Gruehn, 2000; Helm, Kemethofer, Moosbrugger, Bröderbauer & Luthe, 2016; Klieme & Baumert, 2001; Kunter et al., 2005; van Ophuysen & Wendt, 2010).

In den letzten Jahren wurden vermehrt Studien durchgeführt, die versuchten, Veränderung in der Unterrichtsqualität durch die Klassenzusammensetzung vorherzusagen. Dies gelingt nur bedingt, denn mehrheitlich können in den Studien (z. B. Desimone, Hochberg & McMaken, 2016; Kuger, Kluczniok, Kaplan & Rossbach, 2016; Maulana, Opdenakker & Bosker, 2014; Polikoff, 2015; Rowley et al., 2019) keine signifikanten Zusammenhänge mit Merkmalen der Klassenzusammensetzung (z. B. Anteil an leistungsstarken Schülerinnen und Schülern, Anteil an Schülerinnen und Schülern mit Migrationshintergrund, Anteil an sozial benachteiligten Schülerinnen und Schülern) beobachtet werden. In einer Studie aus Indonesien (Maulana et al., 2014) geht jedoch ein höherer Anteil leistungsstarker Schülerinnen und Schüler in der Klasse mit einer positiveren Entwicklung der Unterrichtsqualität einher.

Opponentinnen und Opponenten des AG befürchten eine geringere Unterrichtsqualität in leistungsschwächeren Schulklassen, die häufig aufgrund des überproportionalen Einsatzes geringer qualifizierter Lehrkräfte in leistungsschwächeren Schulklassen vermutet wird. So deutet internationale Forschung (Luschei & Jeong, 2018) darauf hin, dass Lehrkräfte mit den besten Reputationen den leistungsstarken Klassen zugeteilt werden, während die weniger erfahrenen und oft weniger erfolgreichen Lehrkräfte den leistungsschwachen (Förder-)Klassen zugeteilt werden. Weber et al. (2019) zeigen auf Basis der Daten der Bildungsstandardüberprüfung in Mathematik, 4. Schulstufe, aus dem Jahr 2013 sowie Deutsch, 8. Schulstufe, aus dem Jahr 2016, dass dies auch für Österreich gilt. Zudem belegen internationale Studien (bspw. Sharplin, 2014; Sheppard, Padwa, Kelly & Krakehl, 2020; Taylor, Banilower & Clayton, 2020; Weldon, 2016; Ziegler & Richter, 2017), dass fachfremder Unterricht insbesondere in leistungsschwachen Schulen und Schulen in deprivierter Lage stattfindet (Specht, 2011). Darüber hinaus scheinen älteren Studien zufolge (Rosenbaum, 1976; Vanfossen, Jones & Spade, 1987) Lehrkräfte weniger Vorbereitung und weniger Enthusiasmus für leistungsschwache Klassen mitzubringen.

Zusammenfassend schließt Slavin (1990) aus den vorliegenden Studien, dass Schülerinnen und Schüler in homogenen, leistungsschwachen Schulklassen nicht dieselbe Unterrichtsqualität genießen wie leistungsstärkere Schülerinnen und Schüler.

3.2.2 Internationale Studien zur Wirkung des AG auf kognitive und nicht-kognitive Schülerinnen- und Schüler-Outcomes

Mit Blick auf den Forschungsstand zu den Effekten leistungshomogener und leistungsheterogener Lerngruppen auf das fachliche Lernen von Schülerinnen und Schülern liegen mehrere Metaanalysen (z. B. Hattie, 2009; Kulik & Kulik, 1992; Lou et al., 1996; Steenbergen-Hu, Makel & Olszewski-Kubilius, 2016; siehe auch die Reviews von Braunsteiner et al., 2019, oder Becker et al., 2017) vor. Die Metaanalysen berichten von unbedeutenden, vernachlässigbaren Effekten des AG auf fachliche Schülerleistungen. So berichtet Hattie (2009), dass der durchschnittliche Effekt der AG relativ bei $d = 0,05$ liegt. Auch zeigt eine rezente Metaanalyse von Steenbergen-Hu et al. (2016), die rund 100 Jahre Forschung zu AG bündelt, dass zwar Kleingruppenbildung innerhalb von Klassen sowie schulstufenübergreifende Gruppenbildung mit positiven Effekten auf die Leistungen der Schülerinnen und Schüler

verbunden sind, jedoch klassenübergreifende Gruppenbildung – was dem Konzept der Leistungsgruppen der Hauptschule entsprach – keine positiven Effekte aufweist.¹ Becker et al. (2017) verweisen in ihrem Review schließlich darauf, dass der Nutzen, der durch die Fähigkeitsgruppierung entsteht, vor allem leistungsstarken Schülerinnen und Schülern zugutekommt, während weniger leistungsstarke Schülerinnen und Schüler möglicherweise in gemischten Gruppen besser lernen. Neben den Metaanalysen und Review-Studien liegen aktuelle Befunde aus der Post-COVID-19-Zeit vor: Lavrijsen et al. (2022) finden heraus, dass die Heterogenität der Klasse positiv mit dem Einsatz von differenziertem Unterricht durch die Lehrkräfte verbunden ist. Zudem wird festgestellt, dass die Schülerinnen und Schüler in Klassen mit einem hohen Leistungsdurchschnitt bessere Leistungen erbringen; wobei leistungsstarke Schülerinnen und Schüler am meisten davon profitieren, in einer Klasse mit einem hohen Durchschnittsniveau zu sein. Tan und Dimmock (2022) zeigen, dass die Einteilung in Leistungsgruppen zwar die wahrgenommene Unterrichtsqualität beeinflusst, nicht aber den Zusammenhang zwischen der Unterrichtspraxis der Lehrkräfte und den Mathematikleistungen der Schülerinnen und Schüler. Daraus schließen die Autorin und Autor, dass es wichtiger ist, die Beziehung zwischen der Unterrichtsqualität und den Schülerleistungen in homogenen Klassen zu untersuchen als die Beziehung zwischen dem Ability Grouping und den Schülerleistungen an sich. Wu und Becker (2023) stellen fest, dass das Leistungsniveau auf Schulebene den Rückgang der Lebenszufriedenheit signifikant negativ voraussagt, was darauf hindeutet, dass die Lebenszufriedenheit von Schülerinnen und Schülern in Schulen mit einer höherem Leistungsniveau stärker abnimmt als die ihrer Altersgenossen in Schulen mit niedrigerem Niveau.

3.2.3 Studien aus Deutschland zur Wirkung des AG auf Schülerinnen- und Schülerleistungen

Für Deutschland scheint die Befundlage heterogen zu sein. Becker et al. (2017) verweisen allerdings auch darauf, dass nur wenige aussagekräftige Längsschnittstudien vorliegen (Becker, Lüdtke, Trautwein, Köller & Baumert, 2012; Köller & Baumert, 2001; Retelsdorf, Becker, Köller & Möller, 2012). Die Befundlage zeigt einerseits, dass das AG in Deutschland (insbesondere in Schultypen im Sekundarschulbereich) zu unterschiedlichen Leistungsentwicklungen der Schülerinnen und Schüler führt und den Schülerinnen und Schülern daher unterschiedliche Möglichkeiten bietet, ihre Fähigkeiten zu entwickeln. Laut Becker et al. (2017) stehen die Studien aus Deutschland im Einklang mit internationalen Forschungsarbeiten; auch sie legen nahe, dass die stärker akademisch orientierten Schultypen höhere Lernzuwächse fördern (sowohl im Fach Mathematik als auch im Fach Englisch). Als Beispiel kann hier die längsschnittliche Studie von Hosenfeld, Köller und Baumert (1999) angeführt werden. Die Autoren zeigen für das Fach Mathematik, dass – bei Konstanzhaltung des individuellen Vorwissens – Schülerinnen und Schüler im Gymnasium deutlich höhere Wissenszuwächse über die Zeit verzeichnen als Schülerinnen und Schüler anderer Schulformen. Dies bestätigt die Annahme, dass AG insbesondere für leistungsstarke Schülergruppen von Vorteil ist. Leistungsschwache Schülerinnen und Schüler hingegen profitieren

1 Anzumerken ist, dass Steenbergen-Hu et al. (2016) auf Basis weniger randomisierter kontrollierter Studien auch einen leichten Effekt für klassenübergreifende Gruppenbildung zeigen konnten.

laut der Studie von Hosenfeld et al. (1999) deutlich weniger vom mehrgliedrigen Schulsystem. Eine weitere Studie, die diese Vorteile für leistungsstarke Schülerinnen und Schüler beobachten kann, ist jene von Köller und Baumert (2001). Sie untersuchen zusätzlich zum Effekt der Schulform auch den Effekt der Klassenzusammensetzung. Sie kommen zu dem Schluss, dass jenseits der Schulform die auf Schulebene aggregierte Mathematikleistung in der 7. Klasse keinen Effekt auf die Mathematikleistung der Schülerinnen und Schüler in der 10. Schulklasse hat, sodass kein positiver Effekt des Leistungspotenzials der Schulklasse beobachtet werden kann. Im Kontext der AG-Forschung sind auch die Befunde zu Effekten der Gesamtschule in Deutschland interessant. Braunsteiner et al. (2019) fassen die Ergebnisse von Studien zur Leistungsheterogenität in der Gesamtschule (Lehmann & Lenkeit, 2008; Scharenberg, 2012) mit dem Befund zusammen, dass vor allem leistungsschwächere Schülerinnen und Schüler in ihrer Leistungsentwicklung von leistungsheterogenen Lerngruppen profitieren. Leistungsstarke Schülerinnen und Schüler profitieren weniger stark.

Erwähnenswert sind darüber hinaus Befunde aus Studien zu Schülerinnen- und Schülergruppen mit besonderen Bedarfen. Kocaj, Kuhl, Kroth, Pant und Stanat (2014) zeigen auf Basis der Daten des deutschlandweiten IQB-Ländervergleichs für die Primarstufe, dass für Schülerinnen und Schüler mit sonderpädagogischem Förderbedarf eine inklusive Beschulung vorteilhafter ist als an Förderschulen. So zeigten Schülerinnen und Schüler mit sonderpädagogischem Förderbedarf signifikant höhere Kompetenzwerte in Lesen, Zuhören und in der Mathematik als vergleichbare Kinder an Förderschulen. In einer Untersuchung zur Beschulung von begabten Schülerinnen und Schülern in Begabtenklassen im Vergleich zu Regelklassen zeigen Schneider, Stumpf, Preckel und Ziegler (2012), dass nach Kontrolle der IQ-Unterschiede ein etwa vergleichbares Leistungsniveau in beiden Klassenarten beobachtbar ist. Intellektuell besonders begabte Schülerinnen und Schüler dürften daher in beiden Kontexten gut zurechtkommen (siehe auch die geringe Effektstärke von Begabtenklassen bei Hattie, 2009).

3.2.4 Studien aus Österreich zur Wirkung des AG auf Schülerinnen- und Schülerleistungen

Basierend auf österreichischen Daten aus internationalen Leistungsvergleichsstudien konnte Eder (2002) zeigen, dass das Leistungsniveau zwar wie erwartet von der ersten bis zur dritten Leistungsgruppe abnimmt, jedoch deutliche Überschneidungen zwischen den Leistungsgruppen beobachtbar waren. So konnten zweite Leistungsgruppen identifiziert werden, die über eine höhere Durchschnittsleistung als AHS-Klassen verfügten. Auch gab es dritte Leistungsgruppen mit einem jeweils höheren Leistungsdurchschnitt als manche zweite Leistungsgruppe. Nahezu idente Befunde präsentieren Bruneforth et al. (2016) anhand der Bildungsstandardüberprüfungen in Mathematik (2012) und Englisch (2013). Zusammenfassend kann festgehalten werden, dass in Österreich die gewünschte Leistungshomogenisierung in den ehemaligen Hauptschulen durch die Leistungsgruppen nur bedingt funktioniert (da substantielle Leistungsüberschneidungen zwischen Leistungsgruppen bestehen). Als mögliche Ursache führt Eder (2002) an, dass die Zuweisung zu den Leistungsgruppen oft keinen objektiven und vergleichbaren Kriterien folgt. Auch zur Frage, ob die Abschaffung der Leistungsgruppen im Rahmen der Einführung der Neuen Mittelschule

zu bedeutsamen Leistungsänderungen führt, können Weber et al. (2015) und Weber und Kemethofer (2024) zeigen, dass die Umstellung vom Hauptschul- auf das NMS-System keine bedeutsamen Effekte hat, die über alle NMS-Generationen hinweg beobachtbar wären. Dies könnte als Beleg für fehlende Effekte des AG in Österreichs Schulwesen gedeutet werden. Allerdings ist anzumerken, dass der Effekt des Verzichts auf Leistungsgruppen mit Effekten anderer NMS-Konzeptmerkmale oder epochalen Effekten konfundiert sein könnte (siehe dazu Eder & Svecnik, 2015). Schließlich ist die von Braunsteiner et al. (2019) auf Basis der Daten der Bildungsstandardüberprüfung 2015 der 4. und 8. Schulstufe in Österreich gemachte Beobachtung hinsichtlich der Bewältigung von Heterogenität durch Schulen interessant. Braunsteiner et al. (2019) stellen fest, dass Schulen mit einem sehr ähnlichen Erwartungswert und damit sehr ähnlichen Bedingungen hinsichtlich der Schülerzusammensetzung teils sehr unterschiedliche Schulleistungen hervorgebracht haben. „Aus der hohen Streuung kann geschlossen werden, dass Schulen unterschiedlich gut mit Heterogenität umgehen“ (ebd., S. 23). Insgesamt widersprechen diese Befunde dem vermuteten (förderlichen) Effekt von leistungshomogenen Lerngruppen in Österreich.

Während die genannten Theorien und Befunde Erklärungsansätze insbesondere für Effekte der Schulklassenbildung nach der Leistungsfähigkeit von Schülerinnen und Schülern (homogene vs. heterogene Zusammensetzung) sowie für damit zusammenhängende Kriterien (z. B. Migrationshintergrund, soziale Herkunft, Zubringerschulen) bieten, sind sie weniger explanativ für Effekte der Klassenbildung nach Kriterien, die leistungsunabhängiger sind (z. B. Zugehörigkeit zur Religion und Kultur). Obwohl für Österreich gezeigt werden kann, dass auch diese Zugehörigkeiten mit der Schülerleistung assoziiert sind (z. B. Herzog-Punzenberger, 2018), werden im nachfolgenden Abschnitt theoretische Argumente, die mögliche Effekte einer Klassenzusammensetzung nach Religions- und Kulturzugehörigkeit auf die Schülerinnen- und Schülerleistung erklären, dargestellt.

4 Argumentationslinien für die Erklärung von Effekten nicht-leistungsbezogener Kriterien der Klassenzusammensetzung auf die Schulleistung

Die Frage, wie sich die Klassenzusammensetzung nach ethnisch-kultureller Herkunft (z. B. Migrationshintergrund, Religionszugehörigkeit) auf die Schülerinnen- und Schülerleistung auswirkt, ist in der Literatur nicht umfassend erforscht, doch es gibt einige theoretische Ansätze, die relevant sein könnten.

Eine bedeutende Theorie in diesem Kontext ist die Kontakt-Theorie von Allport (2003). Diese besagt, dass der Kontakt zwischen verschiedenen sozialen Gruppen unter bestimmten Bedingungen Vorurteile abbauen und gegenseitiges Verständnis fördern kann. Auf Schulklassen angewendet bedeutet dies, dass eine Zusammensetzung aus Schülerinnen und Schülern unterschiedlicher Kulturen und Religionszugehörigkeit zu einem besseren interkulturellen Verständnis beitragen könnte. Ein solches Umfeld könnte ein harmonischeres Lernklima schaffen, was sich positiv auf die Schulleistungen auswirken könnte.

Ein weiterer relevanter Ansatz ist die Soziale Identitätstheorie von Tajfel und Turner (2004). Laut dieser Theorie neigen Menschen dazu, ihre soziale Identität durch die Zugehörigkeit zu bestimmten Gruppen zu definieren. In Klassen mit homogener Religionszugehörigkeit und oder kultureller Zugehörigkeit könnte dies zu einer starken Gruppenkohäsion führen, die das Gemeinschaftsgefühl und die Unterstützung unter den Schülerinnen und Schülern stärkt. Andererseits könnten heterogene Klassen die Gefahr der Bildung von Subgruppen und potenziellen Konflikten bergen, was die Schulleistungen negativ beeinflussen könnte.

Diese theoretischen Überlegungen und Argumente zeigen, dass die Zusammensetzung von Schulklassen nach kultureller Zugehörigkeit und/oder Religionszugehörigkeit sowohl Chancen als auch Herausforderungen mit sich bringt. Dies bedeutet für die vorliegende Untersuchung, dass auf Basis der bestehenden Theorielage unklar ist, ob überhaupt Effekte auf die Schulleistung zu erwarten sind.

Vor dem Hintergrund der skizzierten theoretischen Annahmen bzw. Überlegungen sowie der dargestellten Befundlage gehen wir im Folgenden der Frage nach, (a) welche Kriterien Schulleitungen bei der Klassenbildung heranziehen und (b) ob und in welchem Ausmaß die Bildung von homogenen/heterogenen Schulklassen im Zusammenhang mit der Schulleistung aus dem fairen Vergleich steht.

5 Methode

5.1 Stichprobe und Methodik

Um der Frage nach den Zusammenhängen bestimmter Selektionsmechanismen bei der Klassenzusammensetzung auf der Sekundarstufe I mit der Schulleistung nachzugehen, werden im vorliegenden Kapitel Daten der Bildungsstandardüberprüfungen der 8. Schulstufe in Mathematik 2012 (M812) und 2017 (M817) sowie in Englisch 2013 (E813) und 2019 (E819) analysiert. Die Bildungsstandardüberprüfung der 8. Schulstufe in Deutsch 2016 wird nicht betrachtet, da die Datengrundlage für das vorliegende Vorhaben nicht vollständig gegeben ist.

In den vier in die Analysen einbezogenen Erhebungen M812, E813, M817 und E819 werden jeweils nur Schulen mit mindestens zwei Klassen berücksichtigt, da ansonsten an den Schulen keine bewusste Aufteilung bzw. Selektion von Schülerinnen und Schülern in unterschiedliche Klassen vorgenommen werden könnte. Jede in den Analysen betrachtete Klasse besteht aus mindestens zwei Schülerinnen und Schülern. Da im vorliegenden Kapitel die Zusammenhänge der Zusammensetzung von Klassen mit der Schulleistung untersucht werden, werden in der analysierten Stichprobe weiters Schulen ausgeschlossen, in denen der Unterricht nicht in der Klassengemeinschaft, sondern in Gruppen (das sind in der Regel Leistungsgruppen der Hauptschulen) stattfindet. Damit sind in der analysierten Stichprobe bei der Erhebung M812 460 Schulen, bei der Erhebung E813 512 Schulen, bei

M817 1.056 und bei E819 1.086. Die Anzahl der in die Stichprobe aufgenommenen Schulen steigt durch die Auflösung der Leistungsgruppen an den Hauptschulen – die für die Zwecke dieser Analyse ausgeschlossen werden.

Die Information zu den an den Schulen vorgenommenen Selektionsmechanismen, mittels derer die Schülerinnen und Schüler in die Klassen aufgeteilt werden, stammt aus den Kontextfragebögen für Schulleiterinnen und Schulleiter, die begleitend zu den Bildungsstandards erhoben werden. Die betrachteten Fragen finden sich in den Kontextfragebögen zu den Erhebungen M812 (Frage 8), E813 (Frage 8) und D816 (Frage 9). Schulleiterinnen und Schulleiter werden auf einer dreistufigen Skala „gar nicht“, „etwas“, und „entscheidend“ befragt, welche der folgenden fünf Merkmale der Schülerinnen und Schüler zur Klassenbildung an ihrer Schule beigetragen haben: Leistungshomogenität, Leistungsheterogenität, Religion, Migration oder die Zubringerschule. Diese Merkmale werden im Folgenden als Selektionsmechanismen (zur Klassenbildung) bezeichnet. Die dreistufige Skala wird dichotomisiert, sodass sich Schulen ergeben, in denen die Selektionsmechanismen „etwas“ oder „entscheidend“ zur Klassenbildung beitragen, und Schulen, in denen die Selektionsmechanismen „gar nicht“ herangezogen werden. Schließlich werden für die Analysen zwei Gruppen von Schulen gebildet: In der einen Gruppe von Schulen wird der Selektionsmechanismus regelmäßig zur Klassenbildung genutzt – Schulen in dieser Gruppe geben in mindestens zwei der drei Kontexterhebungen an, den Mechanismus zu nutzen. In der anderen Gruppe befinden sich Schulen, in denen der Mechanismus nicht genutzt wird, die Schulen haben maximal eine Nennung zur Nutzung des Mechanismus in den drei Kontextfragebögen. Die Einteilung der Schulen in diese zwei Gruppen erfolgt für jeden der fünf Mechanismen. Die Zusammenfassung von Schulen, die einen Mechanismus regelmäßig anwenden, gegenüber den Schulen, die den Mechanismus gar nicht oder selten anwenden, dient dazu, die Indikatoren der Selektionsmechanismen zeitlich robust zu gestalten. So werden in die Gruppe der Schulen, die eine Selektion durchführen, nur Schulen aufgenommen, die dies über mehrere Messzeitpunkte hinweg und somit vermutlich als bewusste Entscheidung tun.

Bei der Gruppenbildung zum Selektionsmechanismus der Leistungshomogenität werden zusätzlich Schulen ausgeschlossen, welche in die 6.–8. Generationen der NMS-Umstellung fallen, also jene, welche zum Zeitpunkt der letzten Erhebung (E819) noch nicht vollständig auf NMS umgestellt waren. Eine Nichtnennung des Selektionsmechanismus Homogenität könnte bei diesen Schulen trotzdem die Bildung von Leistungsklassen bedeuten, da dies standardmäßig in Mittelschulen umgesetzt wird.

Für alle Schulen in der Stichprobe werden der Leistungswert (WLE) und die erwartete Leistung (Erwartungswert) aus den Daten genutzt. Der Erwartungswert gibt die aufgrund von bestimmten Hintergrundmerkmalen der Schule und ihrer Schülerschaft erwartete Leistung der Schule wieder (George & Robitzsch, 2015). Sowohl der Leistungswert als auch die erwartete Leistung werden auf einer stetigen 500er-Skala berichtet. Als Maß der Leistungshomogenität innerhalb der Klassen bzw. Leistungsheterogenität zwischen den Klassen wird für jede Schule in der Stichprobe die Intraklassen-Korrelation (ICC) auf

Grundlage der gewichteten Schülerleistungen berechnet (Field, 2005). Der ICC wird größer, wenn Schülerinnen und Schüler in einer Klasse ähnliche Leistungen zeigen, sowie auch, wenn die Leistungen zwischen Klassen verschiedener sind.

Die Analyseergebnisse werden getrennt in den vier Strata APS Stadt, APS Land, AHS Stadt und AHS Land sowie nach den fünf Selektionsmechanismen Leistungshomogenität, -heterogenität, Migration, Religion und Zubringerschule dargestellt. Bei der Einteilung der Schulstandorte in die vier Strata wird also nach Schultyp „Allgemeinbildende höhere Schule“ (AHS) oder „Allgemeine Pflichtschule“ (APS) unterschieden so wie Schulstandorte in Gemeinden mit einer Größe unter 50.000 Einwohnerinnen und Einwohnern als „Land“ bezeichnet werden, dagegen fallen Schulstandorte in Städten mit mehr als 50.000 Einwohnerinnen und Einwohnern in die Kategorie „Stadt“.

Bei den so aufgeteilten Grafiken zeigt die jeweils linke Seite Werte der Schulen, welche einen Mechanismus anwenden, die rechte Seite zeigt Werte für Schulen, welche den Mechanismus nicht anwenden. Jeder Punkt im Plot stellt die für eine Gruppe von Schulen (Anwendung des Mechanismus oder nicht) und eine Erhebung (M812, E813, M817, E819) gemittelte Abweichung der Leistungen der Schulen von den erwarteten Leistungen der Schulen dar. Ein Wert größer 0 zeigt dabei, dass die Gruppe der Schulen im Mittel besser abschneidet, als von ihnen erwartet wird. Zusätzlich wird über alle vier Erhebungen hinweg die gemittelte Abweichung für Schulen, die den Mechanismus anwenden bzw. die ihn nicht anwenden, dargestellt (fette Linien). Aufgrund der oben beschriebenen, nicht zufälligen Bildung der dieser Studie zugrunde liegenden Stichprobe wird in den Gruppen von einer Prüfung der Signifikanz der Abweichungen vom Erwartungswert bewusst abgesehen. Zur Einordnung kann festgehalten werden, dass die Abweichungen vom Erwartungswert aller Schulen im Mittel über die Erhebungen hinweg eine Standardabweichung von 25 Punkten aufweisen. Eine mittlere Abweichung von 5 Punkten in einer Gruppe entspricht somit ca. einem Fünftel Standardabweichung. Weiters wird für Schulen, die einen Selektionsmechanismus erfüllen bzw. nicht erfüllen, in jedem Plot der Median der ICCs über alle Erhebungen angegeben.

6 Ergebnisse

Die Analysen werden entlang der Selektionskriterien (leistungsheterogene vs. leistungshomogene Klassenbildung, Migration, Religion, Zubringerschule) im Folgenden getrennt für den Schultyp (APS, AHS) und die Region (Stadt, Land) berichtet.

6.1 Anzahl an Schulen mit jeweiligem Selektionskriterium

Tabelle 1 enthält die absoluten und relativen Häufigkeiten der Schulen, deren Schulleitungen in mindestens zwei von drei Bildungsstandarterhebungen angaben, das jeweilige Selektionskriterium anzuwenden. Demnach wurden am häufigsten (64 %) Schülerinnen und Schüler aufgrund des Kriteriums „Zubringerschule“ einer Schulklasse zugewiesen, wobei dieses

Kriterium in den AHS (78 %) nochmals deutlich häufiger angewandt wurde als in den APS (58 %). Vergleichsweise selten (5 % bzw. 7 %) kamen die Selektionskriterien Migration und Homogenität zum Einsatz.

Tabelle 1: Prozentsatz an Schulen in der Stichprobe mit Selektion in mindestens 2 von 3 Erhebungen

Selektionskriterium	Gesamt	AHS	APS
Leistungsheterogenität	36 % [275]	17 %	45 %
Leistungshomogenität	7 % [46]	5 %	8 %
Migration	5 % [37]	5 %	5 %
Religion	34 % [250]	58 %	25 %
Zubringerschule	64 % [500]	78 %	59 %

Anmerkungen: AHS = Allgemeinbildende höhere Schule, APS = Allgemeinbildende Pflichtschule. Der Prozentsatz der Schulen bezieht sich immer auf die Schulen in der Stichprobe, also Schulen, in denen mehr als eine Klasse existiert und die Unterrichtsgruppe der Klasse entspricht, und ist gemittelt über alle vier Erhebungen M812, E813, M817 und E819. Bei der Gesamtangabe des Prozentsatzes ist jeweils die über alle vier Erhebungen gemittelte absolute Anzahl an Schulen gegeben, die einen Selektionsmechanismus regelmäßig anwendet.

6.2 Zentrale Befunde zu Effekten der Selektion über Schultypen hinweg

Als zentraler Befund kann festgehalten werden, dass auf Basis der Analysen kein konsistentes Muster über die Selektionskriterien, Erhebungen, den Schultyp und die Region hinweg beobachtet werden kann. Vielmehr zeigen die Befunde, dass die Anwendung von Selektionsmaßnahmen zur Klassenbildung einmal mit – im Vergleich zu Schulen, die das Selektionskriterium nicht anwenden – positiveren und einmal negativeren Abweichungen vom Erwartungswert einhergehen.

Tabelle 2: Abweichung zwischen Schulen, die ein Selektionskriterium anwenden und Schulen, die das Kriterium nicht anwenden, gemittelt über die Erhebungszeitpunkte M812, E813, M817 und E819

Selektionskriterium	AHS		APS	
	Stadt	Land	Stadt	Land
Leistungsheterogenität	-3,01	9,09	0,97	0,28
Leistungshomogenität	5,82	5,93	-7,6	-2,79
Migration	7,15	-6,58	-2,54	5,13
Religion	-0,18	-4,05	-3,52	-0,09
Zubringerschule	-0,67	-2,48	1,24	0,25

Anmerkungen: AHS = Allgemeinbildende höhere Schule, APS = Allgemeinbildende Pflichtschule.

Tabelle 2 enthält die über alle vier Studien gemittelte Abweichung zwischen Schulen, die ein Selektionskriterium anwenden, und Schulen, die das Kriterium nicht anwenden. Positive/negative Werte bedeuten, dass Schulen mit angewendetem Selektionskriterium im Mittel besser/schlechter abschneiden (weiter über/unter dem Erwartungswert liegen) als Schulen, die das Kriterium nicht zur Klasseneinteilung nutzen. Die Tabelle spiegelt das inkonsistente Muster wider. Einzig die Selektion nach Religionszugehörigkeit und Zubringerschule ist in keinem Schultyp und weder in der Stadt noch auf dem Land im Mittel für die Schulen erfolgreicher, als keine Selektion nach Religion vorzunehmen. Dagegen scheint eine Selektion nach Leistungshomogenität in den AHS mit einem Unterschied von über 5 Punkten gegenüber AHS, die nicht nach Homogenität selektieren, im Mittel eher erfolgreich zu sein. Die folgenden Abschnitte zeigen allerdings, wie unterschiedlich die Ergebnisse über die Erhebungszeitpunkte hinweg ausfallen.

6.3 Effekte der Selektion in den allgemeinbildenden höheren Schulen (AHS)

Betrachtet man die AHS-Schulen, so ist der augenscheinlich stärkste Befund hinsichtlich des Selektionskriteriums *Leistungsheterogenität* beobachtbar (Abbildung 1); und zwar im ländlichen Bereich.

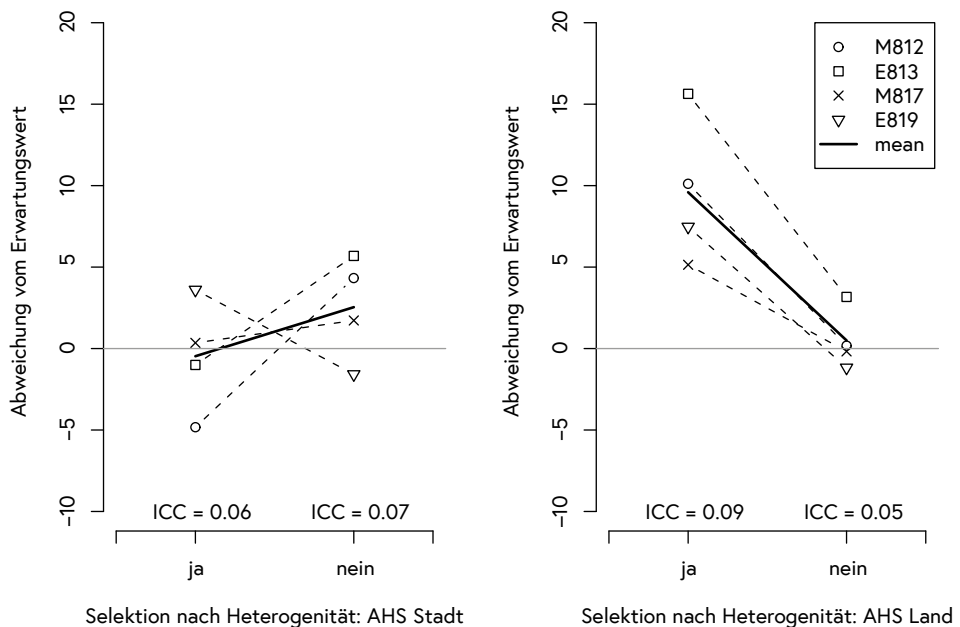


Abbildung 1: Abweichung vom Erwartungswert der Leistung bei Selektion nach Heterogenität in der AHS mit Median der ICC je Selektionsgruppe

Hier zeigt sich über alle Messpunkte hinweg eine positive Abweichung von erwarteten Werten, wenn das Selektionskriterium der Heterogenität angewendet wird. Konkret erzielen

jene Schulen, die das Selektionskriterium anwenden, im Durchschnitt Schülerleistungen, die deutlich stärker über dem Erwartungswert liegen als Schulen, die ihre Schulklassen nicht heterogen zusammensetzen. Diese Differenz liegt zwischen rund 5 und 12 Punkten. Somit entspricht der durchschnittliche Unterschied von rund 9 Punkten über alle Erhebungen hinweg (fette Linie) rund zwei Fünftel Standardabweichungen. Ein ähnlich konsistentes Muster ist für die AHS-Stadtschulen nicht beobachtbar. Die ICC-Analyse verweist darauf, dass die Selektionsmaßnahme im ländlichen Bereich allerdings damit einhergeht, dass die Klassenleistungen stärker zwischen den Klassen variieren, wenn die Selektionsmaßnahme angewandt wird. Wobei jedoch anzumerken ist, dass die ICC (Median 0,09) zwischen den Erhebungen zwischen 0,07 und 0,11 variieren.

Für das Kriterium *Homogenität* zeigen sich sowohl für städtische als auch ländliche AHS-Schulen leichte Vorteile der Selektion (Abbildung 2). Dies gilt allerdings nur im Durchschnitt über alle Erhebungen hinweg betrachtet. Gerade im städtischen Bereich sind auch Erhebungen mit negativen Abweichungen vom Erwartungswert beobachtbar, wenn das Selektionskriterium der Homogenität angewandt wird. Die ICC-Analyse zeigt, dass die durchschnittlichen Klassenleistungen in städtischen AHS-Schulen ohne Selektionsmaßnahme etwas stärker variieren.

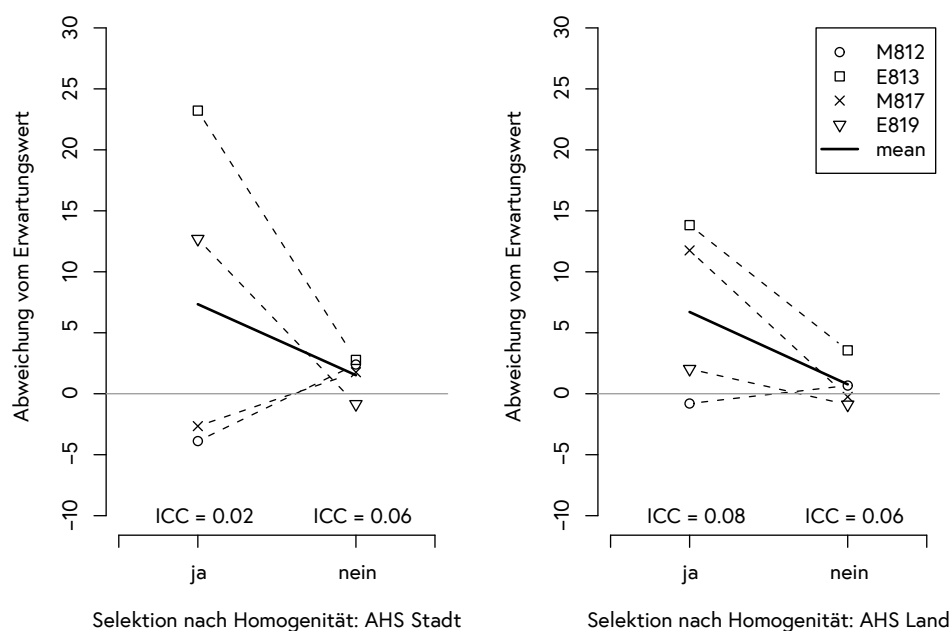


Abbildung 2: Abweichung vom Erwartungswert der Leistung bei Selektion nach Homogenität in der AHS mit Median der ICC je Selektionsgruppe

Auch die Selektion nach *Migration* in AHS-Schulen im städtischen Bereich scheint minimale Vorteile aufzuweisen, während die Selektion nach Migration auf dem Land eher nachteilig zu sein scheint, wobei hier die einzelnen Erhebungszeitpunkte zu ganz unterschiedlichen Befunden führen (Abbildung 3). Die ICC-Analyse zeigt, dass die durchschnittlichen Klassenleistungen in städtischen AHS-Schulen ohne Selektionsmaßnahme stärker variieren.

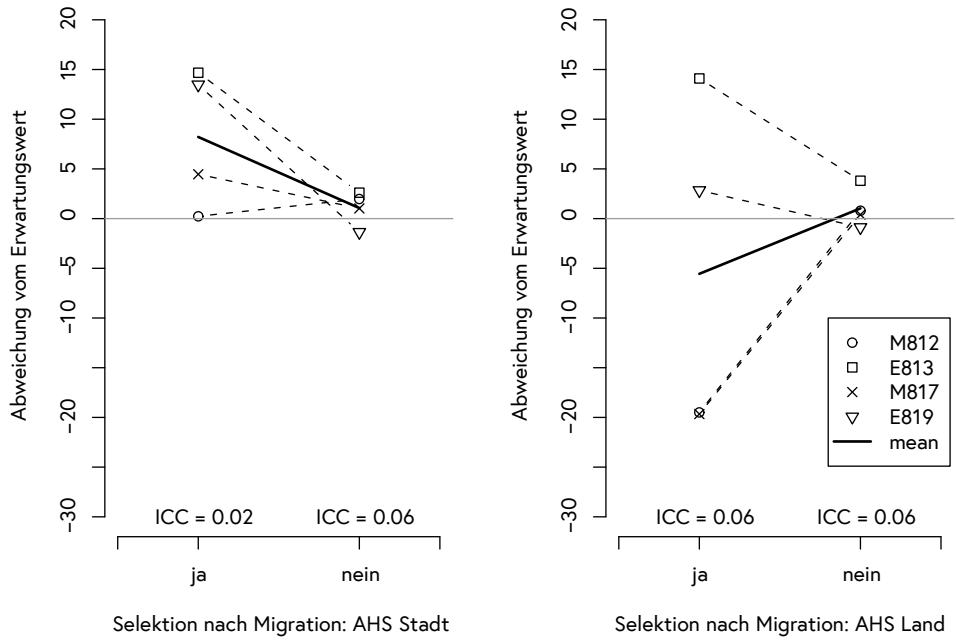


Abbildung 3: Abweichung vom Erwartungswert der Leistung bei Selektion nach Migration in der AHS mit Median der ICC je Selektionsgruppe

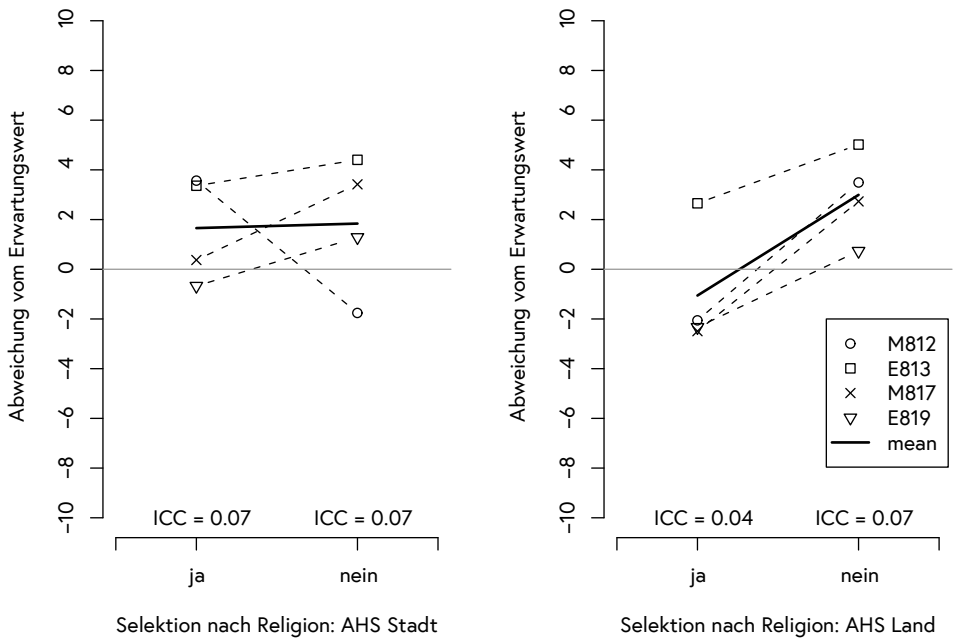


Abbildung 4: Abweichung vom Erwartungswert der Leistung bei Selektion nach Religionszugehörigkeit in der AHS mit Median der ICC je Selektionsgruppe

Das Selektionskriterium *Religion* (Abbildung 4) scheint auf dem Land durchgängig mit negativen Abweichungen vom Erwartungswert einherzugehen; hier schneiden in allen Erhebungen Schulen ohne Selektion besser als erwartet ab, während jene mit Selektion (bis auf eine Erhebung) schlechter als erwartet abschneiden. In der Stadt ist die Klassenbildung nach dem Kriterium der Religionszugehörigkeit in Abhängigkeit der Erhebung einmal mit einem positiven, einmal mit negativen Abweichungen verbunden.

Die Selektion entlang des Kriteriums *Zubringerschule* ist für AHS-Schulen im städtischen Bereich im Durchschnitt der Erhebungen ohne wesentliche Abweichungen vom Erwartungswert (Abbildung 5). Allerdings weisen die verschiedenen Erhebungen auf eine sehr hohe Heterogenität in den Befunden hin. Im ländlichen Bereich führen die Analysen zu eher konsistenten Befunden über Erhebungen hinweg. Es scheint, dass die Selektion nach der Zubringerschule mit leicht negativen Abweichungen zu den Erwartungswerten einhergeht.

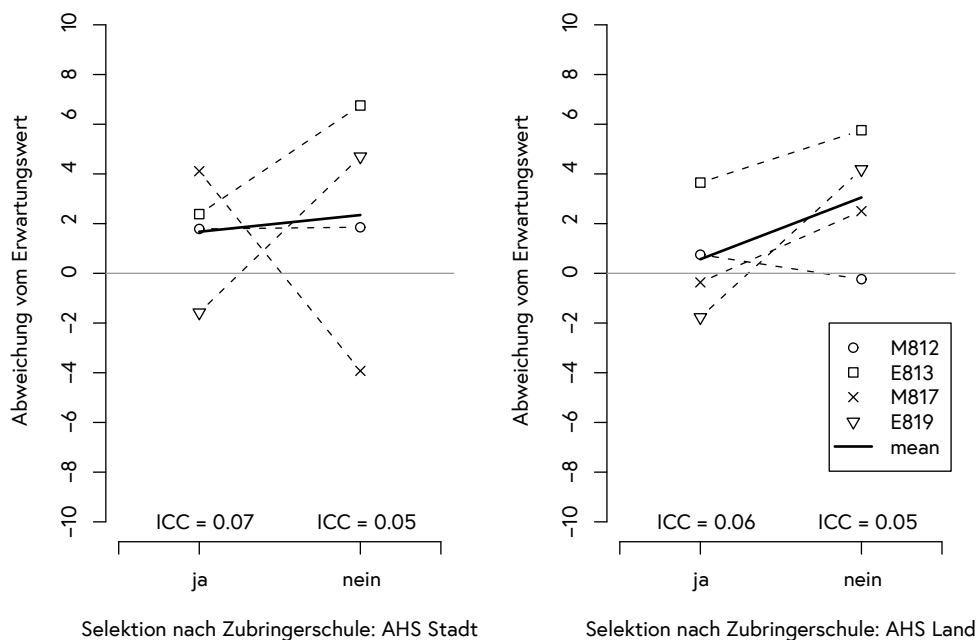


Abbildung 5: Abweichung vom Erwartungswert der Leistung bei Selektion nach Zubringerschule in der AHS mit Median der ICC je Selektionsgruppe

6.4 Effekte der Selektion in den allgemeinbildenden Pflichtschulen (APS)

Betrachtet man die Abweichungen vom Erwartungswert für Pflichtschulen getrennt nach Schulen, die Selektionsmaßnahmen anwenden, und jenen, die das nicht tun, so zeigen sich mit Ausnahme des Kriteriums der Homogenität für keines der Selektionskriterien relevante Abweichungen vom Erwartungswert, die sowohl im städtischen als auch im ländlichen Bereich in dieselbe Richtung gehen.

Das Selektionskriterium *Heterogenität* weist sowohl in Pflichtschulen auf dem Land als auch im städtischen Bereich kaum Abweichungen vom Erwartungswert auf (Abbildung 6).

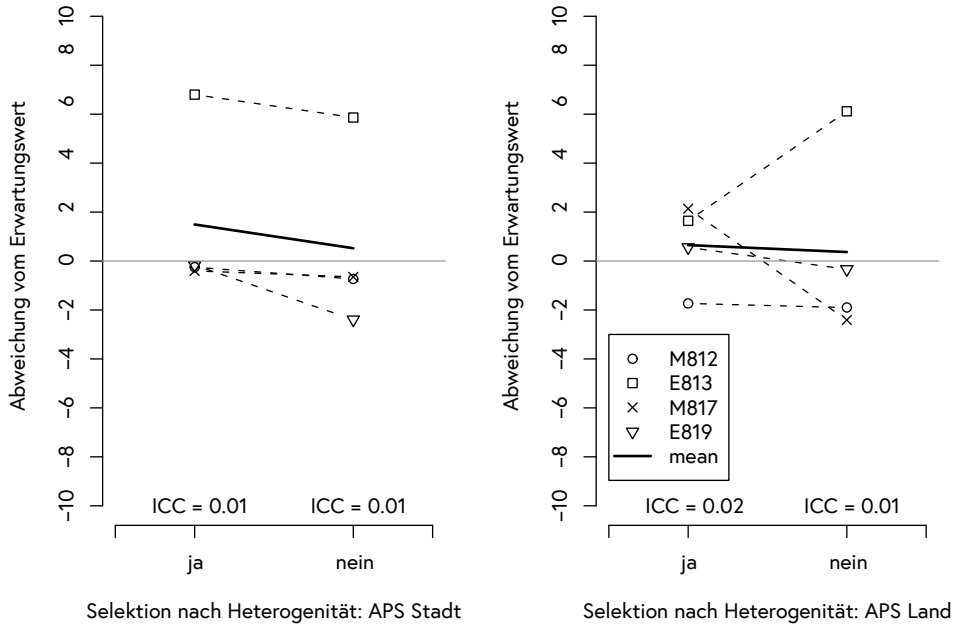


Abbildung 6: Abweichung vom Erwartungswert der Leistung bei Selektion nach Heterogenität in der APS mit Median der ICC je Selektionsgruppe

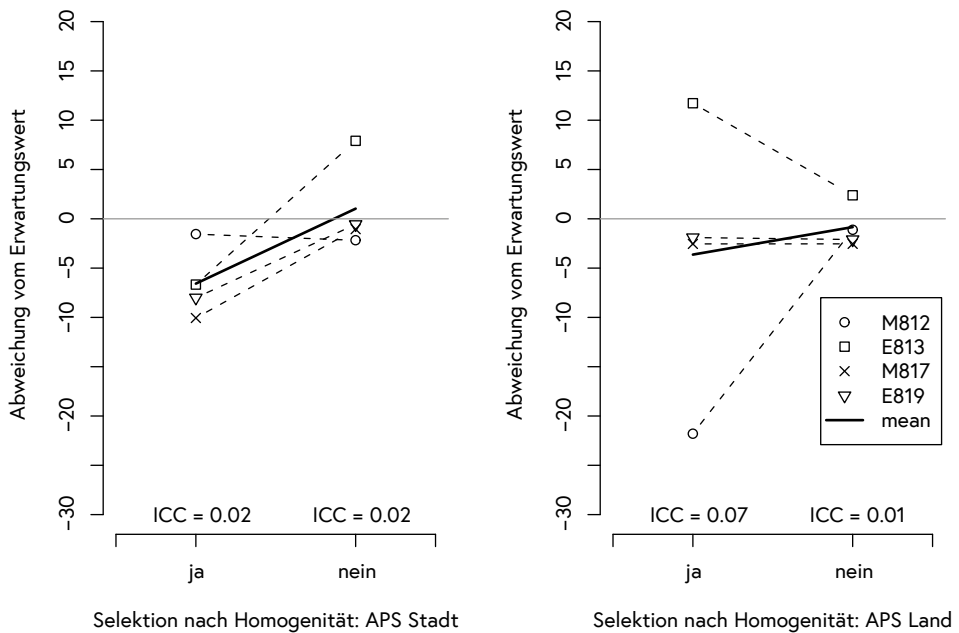


Abbildung 7: Abweichung vom Erwartungswert der Leistung bei Selektion nach Homogenität in der APS mit Median der ICC je Selektionsgruppe

Für das Kriterium der *Homogenität* (Abbildung 7) ist zu beobachten, dass Pflichtschulen, die das Kriterium nicht anwenden, weniger stark negativ vom Erwartungswert abweichen. Dies gilt insbesondere für den städtischen Bereich. Im ländlichen Bereich sind die Abweichungen sehr schwach ausgeprägt und zwischen den Erhebungen ist eine deutliche Varianz in den Befunden beobachtbar. Hier zeigt die ICC-Analyse auch, dass die durchschnittliche Klassenleistung deutlich stärker in Schulen variiert, die das Selektionskriterium anwenden.

Ähnlich wie für das Kriterium der Heterogenität verhält es sich mit dem Kriterium *Migration*, das im Durchschnitt im städtischen Pflichtschulbereich ohne relevante Abweichungen vom Erwartungswert bleibt (Abbildung 8). Im ländlichen Bereich scheint die Selektion zu leichten, positiven Abweichungen vom Erwartungswert zu führen, und zwar für nahezu alle Erhebungen.

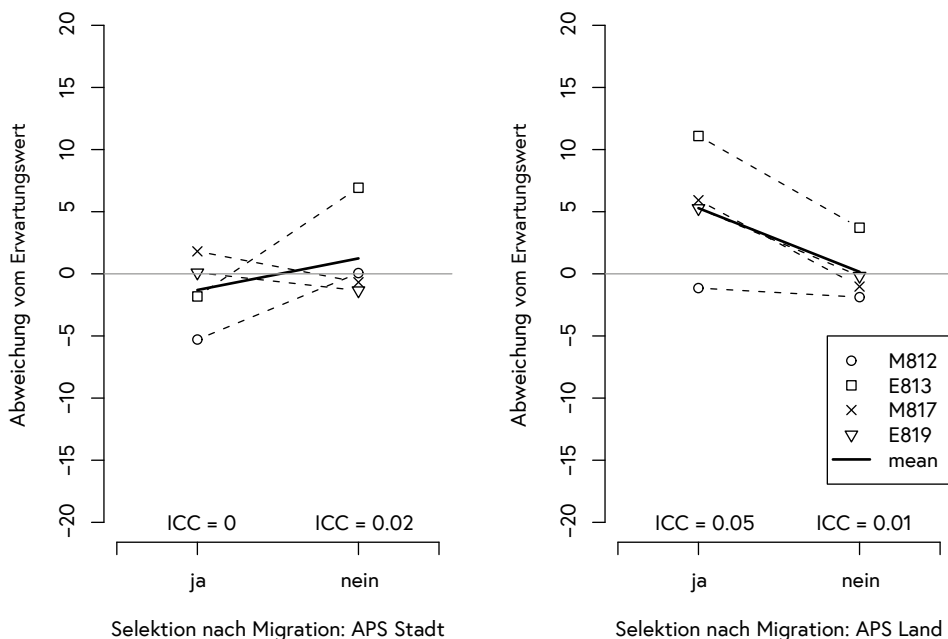


Abbildung 8: Abweichung vom Erwartungswert der Leistung bei Selektion nach Migration in der APS mit Median der ICC je Selektionsgruppe

Die Klassenbildung nach dem Kriterium der *Religionszugehörigkeit* (Abbildung 9) scheint im städtischen Bereich im Mittel negative Effekte zu haben, während eine solche Selektion im ländlichen Bereich im Mittel keinen deutlichen Unterschied zu den Klassen ohne Selektion aufweist.

Die *Zubringerschule* als Selektionskriterium weist im Durchschnitt der Erhebungen weder im städtischen noch im ländlichen Bereich relevante Abweichungen vom Erwartungswert auf (Abbildung 10).

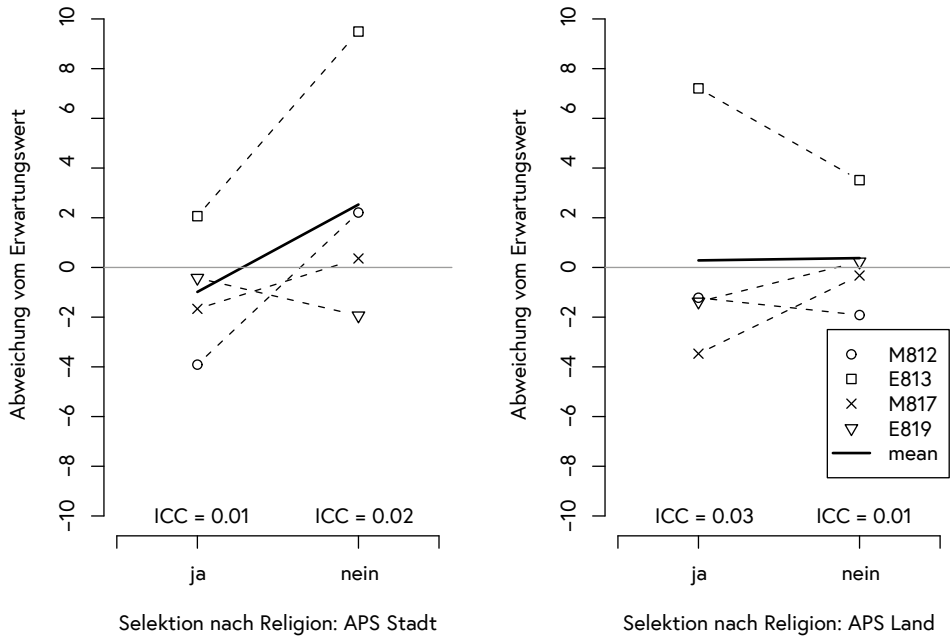


Abbildung 9: Abweichung vom Erwartungswert der Leistung bei Selektion nach Religionszugehörigkeit in der APS mit Median der ICC je Selektionsgruppe

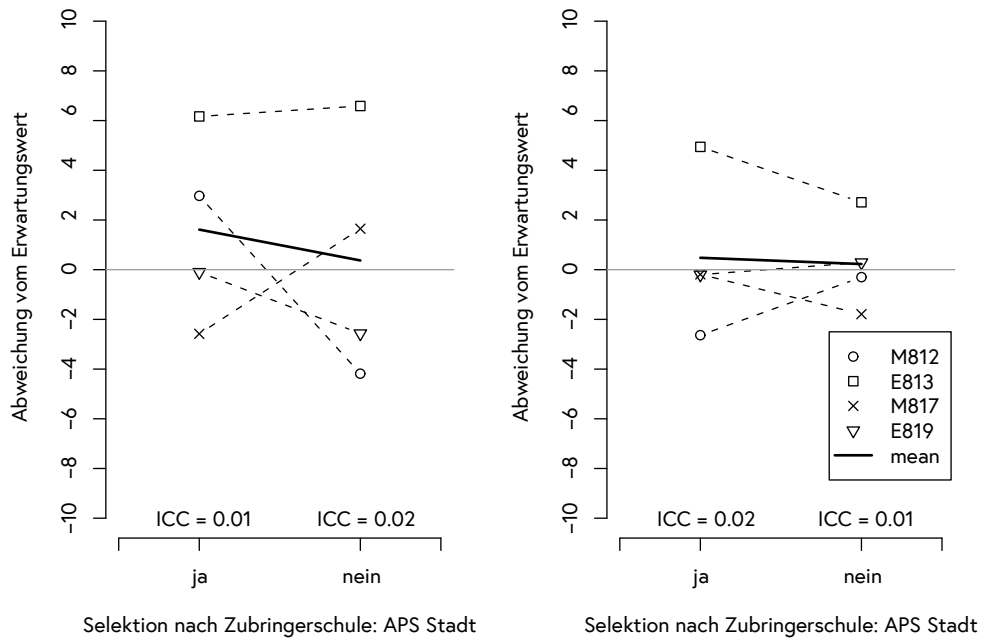


Abbildung 10: Abweichung vom Erwartungswert der Leistung bei Selektion nach Zubringerschule in der APS mit Median der ICC je Selektionsgruppe

7 Diskussion

Qualitative Forschungsbefunde aus Interviews mit österreichischen Schulleitungen (Sertl & Leditzky, 2016; Weber et al., 2019), die Schulleiterbefragungen im Rahmen der Bildungsstandardüberprüfungen (Weber et al., 2019) sowie die rechtliche Möglichkeit, dauerhaft Lerngruppen nach Leistungsniveaus zu bilden (BMBWF, 2020), lassen schließen, dass an Österreichs Schulen Schülerinnen und Schüler systematisch nach gewissen Selektionskriterien in Klassen eingeteilt werden. Internationale Forschungsbefunde legen nahe, dass durch die systematische Klassenzuordnung auch das Lernsetting (z. B. Unterrichtsqualität) sowie das Lernen (z. B. Lernemotionen) und die Outcomes der Schülerinnen und Schüler (z. B. Leistungen) beeinflusst werden, insbesondere, wenn es zum sogenannten Ability Grouping kommt, d. h., homogene, leistungsstarke bzw. -schwache Klassen gebildet werden (z. B. Slavin, 1990).

Vor dem Hintergrund dieser Befundlage wurde in der vorliegenden Untersuchung – auf Basis von auf Schulebene im mehrjährigen Längsschnitt erfassten Schulleistungs- und Kontextdaten – die Frage nach dem Zusammenhang von Kriterien der Klassenzusammensetzung und der Schulleistung österreichischer Schulen analysiert. Empirisch belastbare Antworten auf diese Frage sind sowohl wissenschaftlich als auch praktisch relevant, da sie Einblicke in die Gestaltung effektiver Lernumgebungen bieten. Die vorliegende Studie zeigt allerdings, dass nicht nur eine Vielzahl von verschiedenen Selektionskriterien im unterschiedlichen Ausmaß an Österreichs Schulen angewandt wird, sondern auch, dass die Zusammenhänge der Klassenzusammensetzung bzw. angewandten Selektionskriterien und der Abweichungen von den erwarteten Schulleistungen überwiegend gering und in ihrer Richtung inkonsistent sind. Besonders an ländlichen AHS wurde eine positive Abweichung von der erwarteten Schulleistung beobachtet, wenn das Selektionskriterium der Heterogenität angewandt wurde; dagegen eine negative Abweichung, wenn das Kriterium der Religion Anwendung fand. Für die anderen untersuchten Selektionskriterien waren die Ergebnisse weniger eindeutig. Die heterogenen Befunde deuten auf eine komplexe Beziehung zwischen Selektionsmechanismen und Schulleistung hin. Sie stehen aber auch im Einklang mit der internationalen Befundlage (siehe die Metaanalyse von Steenbergen-Hu et al., 2016) sowie Studien aus Deutschland (siehe Becker et al., 2017, für eine Übersicht).

Mögliche Ursachen für die heterogenen Ergebnisse der vorliegenden Studien sind vielfältig. Erstens: Die Annahme, dass homogene Lerngruppen zu einer höheren Unterrichtsqualität und damit zu einem höheren Lernerfolg oder zu höherer Schulleistung führen, ist schon deshalb zu hinterfragen, da es schwierig ist, Unterrichtsqualität allgemeingültig für unterschiedliche Settings zu definieren. So verweisen Gamoran, Nystrand, Berends und LePore (1995) darauf, dass aus Sicht der Bildungswissenschaft und -forschung noch nicht ausreichend geklärt ist, was Unterrichtsqualität ausmacht, wie diese mit Lernen in unterschiedlichen Kontexten zusammenhängt und wie der Unterrichtsstil in Abhängigkeit von den Charakteristiken und Reaktionen der Schülerinnen und Schüler variiert. Hier schließt die aktuelle, kritische Diskussion von Praetorius, Klieme, Herbert und Pinger (2018) und Begrich et al. (2023) an die drei Basisdimensionen der Unterrichtsqualität an. Damit ist

auch für Lehrkräfte unklar, wie sie ihre Unterrichtsmethoden für unterschiedliche Lerngruppen variieren sollen. Daher kommen Gamoran et al. (1995) zum Schluss, dass, auch wenn AG auf den ersten Blick als eine vielversprechende Anwendung organisatorischer Differenzierung erscheint, sie in der Praxis nicht wie geplant funktioniert und zu nicht intendierten Konsequenzen führen kann (z. B. zu vermehrter Bildungsungleichheit, bspw. Maaz, Trautwein, Lüdtke & Baumert, 2008).

Zweitens: Neben der Theorie, dass homogene Lerngruppen zu förderlicheren Lernumgebungen führen, gibt es konkurrierende Theorien wie die Differentiation-Polarisation-Theorie und die Labelling-Theorie, die von negativen Effekten ausgehen (z. B. Statusverlust, Stereotypisierung, Stigmatisierung). Zudem existieren Argumente, die darauf verweisen, dass die Potenziale homogener Gruppen (z. B. einfacher Schülerzentrierung) von Lehrkräften nicht genutzt werden können. Die durch die Theorien vorhergesagten Effekte könnten sich gegenseitig „neutralisieren“, sodass die Nettoeffekte gering ausfallen; oder aber je nach Kontext unterschiedlich stark auf die Schulleistung „durchschlagen“.

Drittens: Es zeigt sich, dass homogene Lerngruppen eher für leistungsstarke Schülerinnen und Schüler förderlich sind, während sie für leistungsschwache Schülerinnen und Schüler weniger vorteilhaft sein können, da die intellektuelle Kapazität der Mitschülerinnen und -schüler fehlt (z. B. Dar & Resh, 1986). Da die vorliegende Studie keine Individualleistungen in den Blick nimmt, ist zu erwarten, dass sich diese differenziellen Effekte auf Schulebene austarieren (ähnlich wie beim BFLPE; Marsh et al., 2007).

Viertens gehen die Effekte des expliziten AG (Tracking) im österreichischen Schulsystem, d. h. der Trennung in APS und AHS, nicht nur mit starken Unterschieden in der Schulleistung, sondern auch in der Unterrichtsqualität bzw. dem Lehrerverhalten einher (z. B. Helm et al., 2016), sodass die Klassenzusammensetzung eine nur noch untergeordnete Rolle innerhalb dieser beiden Lernmilieus aufweist. Zudem ist auch in Österreich beobachtbar, dass in homogenen Schulklassen tendenziell weniger qualifizierte Lehrkräfte eingesetzt werden (z. B. Weber et al., 2019), was positive Effekte unterminieren könnte.

Fünftens sind mit der verfügbaren Datenlage auch Einschränkungen der (internen) Validität des Untersuchungsdesigns verbunden. Es wurden selbstberichtete Angaben von Schulleitungen zur Klassenzusammensetzung herangezogen, wobei die Zusammensetzung vier Jahre vor der Befragung erfolgte. Sofern die Art der Klassenzusammensetzung nicht Teil einer persistierenden Schulkultur ist, ist es fraglich, ob die Schulleiterangaben die tatsächliche Praxis an den Schulen widerspiegeln. Dieser Problematik wurde in dem Beitrag damit begegnet, dass nur Schulen einem Selektionskriterium zugewiesen wurden, wenn es über die Zeit hinweg zumindest zweimal als bedeutsam erachtet wurde. Dennoch kann das Problem dadurch nur abgeschwächt werden. Gleichmaßen können neben den oben genannten konfundierenden Variablen (Unterrichtsqualität ...) auch zwischenzeitliche Ereignisse (Veränderung der Klassenzusammensetzung durch Wiederholerinnen und Wiederholer, Schulwechslerinnen und Schulwechsler) dazu beitragen, dass die eingangs (z. B. homogen) gebildete Klasse im Laufe der Sekundarstufe 1 an Heterogenität gewinnt.

Schlussendlich können sich auch Zuteilungskriterien – an den Schulen werden in der Regel mehrere Kriterien berücksichtigt – mit unterschiedlichen Effekten überlagern bzw. aufheben (z. B. Leistungsheterogenität und Zubringerschulen).

Dennoch trägt die vorliegende Studie zum vorhandenen Wissen bei, indem sie spezifische Einblicke in die Dynamik der Klassenzusammensetzung in Österreich bietet und somit die Diskussion über Schulqualität, Bildungsgerechtigkeit sowie pädagogische Praktiken bereichert. Die theoretischen und praktischen Implikationen der Ergebnisse weisen auf eine differenzierte Betrachtung der Klassenzusammensetzung hin. Theoretisch erweitern die Befunde das Verständnis von Selektionsmechanismen im Rahmen der Klassenzusammensetzung und deren Einfluss auf die Schulleistung – die bisher kaum beforscht sind (Weber et al., 2019) –, indem sie aufzeigen, dass die beobachteten Zusammenhänge heterogen und kontextabhängig sind, insbesondere von Schultyp und von der Region (Stadt/Land). Folgestudien sind nötig, um zu zeigen, wie die Befunde zur Entwicklung gezielter pädagogischer Strategien beitragen, die die spezifischen Bedingungen verschiedener Schulformen und -standorte berücksichtigen, um so Bildungschancen zu optimieren. Weiterhin regen die Befunde zu weiterführender Forschung an, um die zugrunde liegenden Mechanismen genauer zu verstehen. Zukünftige Forschungsarbeiten könnten sich darauf konzentrieren, die dynamischen Wechselwirkungen zwischen Klassenzusammensetzung, Lehrmethoden und Schülerleistung in verschiedenen Bildungskontexten genauer zu untersuchen. Es besteht Bedarf an Studien, die longitudinale Daten nutzen, um die Langzeiteffekte der Klassenzusammensetzung auf Schülerebene zu erfassen. Ebenso wäre von Interesse, die Rolle individueller Schülermerkmale und deren Interaktion mit der Klassenzusammensetzung detaillierter zu betrachten, um differenziertere pädagogische Ansätze zu entwickeln.

8 Fazit

Die heterogene Befundlage der vorliegenden Analysen sowie die Ergebnisse aus internationalen und nationalen Studien (Steenbergen-Hu et al., 2016; Weber & Kemethofer, 2024) implizieren, dass weder die Bildung homogener noch heterogener Lerngruppen einen bedeutsamen positiven Effekt auf die fachlichen Leistungen von Schülerinnen und Schülern erwarten lässt. Die Hoffnung, dass eine ohne großen Aufwand und finanzielle Mittel leicht realisierbare Maßnahme wie die explizite Zuweisung von Schülerinnen und Schülern zu Klassen entsprechend ihrem Leistungsniveau (bzw. auch anderer Merkmale) Vorteile mit sich bringen würde, wird somit nicht bestätigt. Betrachtet man die Metaanalyse von Steenbergen-Hu et al. (2016), so zeigen sich nur bei flexiblen Strategien geringe positive Effekte. Auch das Kapitel 10 in diesem Band findet keine Effekte: „Eine Analyse der Zusammenhänge für Österreich zwischen dem Ausmaß an Unterricht mit der gesamten Klasse, an Binnendifferenzierung und an Individualisierung mit jeweils Leseleistung ergab keine nennenswerten Korrelationen. Dieses Ergebnis ist laut Forschungsstand (s. Abschnitt 3) durchaus erwartungskonform.“ Kritisch anzumerken ist, dass die nationale und internationale Forschungslage zu Leistungsgruppierungen großteils auf Studien mit einer eher geringen Validität basiert und somit berichtete Befunde durch unterschied-

liche konfundierende Variablen überlagert sein können (siehe auch Diskussion oben). Vor diesem Hintergrund gilt es für die Bildungsforschung, etwaige Effekte von Homogenisierung vs. Heterogenisierung mit hochvaliden Forschungsdesigns (randomisiert kontrollierte Studien ...) zu untersuchen, um hier schließlich eine gesicherte Evidenzbasis für die Frage, wie Klassen idealerweise gebildet werden sollten, zu schaffen. Solange keine ausreichende Evidenzbasis vorliegt, sollten die immer wieder auftauchenden Diskussionen zur Klassenbildung, insbesondere in Bezug auf Leistungsgruppen, vermieden werden und die (mediale) Aufmerksamkeit auf Aspekte gelenkt werden, deren Bedeutung für den Lernerfolg der Schülerinnen und Schüler bereits belegt ist. Abschließend sollte neben dem Fokus auf dem Lernerfolg der Schülerinnen und Schüler auch bedacht werden, dass die Bildung von homogenen bzw. heterogenen Lerngruppen möglicherweise Einfluss auf Aspekte haben kann, die hier nicht untersucht werden, wie das Selbstkonzept der Schülerinnen und Schüler (vgl. Kapitel 9 in diesem Band) oder die Herausforderungen, denen Lehrkräfte im Umgang mit heterogenen Schulklassen gegenüberstehen (vgl. die aktuelle Diskussion zum hohen Anteil an Schülerinnen und Schülern mit Migrationshintergrund und unterschiedlichen Religionsbekenntnissen in städtischen Pflichtschulen).

Literatur

- Abraham, J. (1989). Testing Hargreaves' and Lacey's Differentiation-Polarisation Theory in a Setted Comprehensive. *The British Journal of Sociology*, 40 (1), 46–81. doi:10.2307/590290
- Allport, G. W. (2003). *The nature of prejudice* (25th anniversary ed. [31. print], unabridged.). Perseus Books.
- Altrichter, H., Heinrich, M. & Soukup-Altrichter, K. (2014). School decentralization as a process of differentiation, hierarchization and selection. *Journal of Education Policy*, 29 (5), 675–699. doi:10.1080/02680939.2013.873954
- Baumert, J., Kunter, M., Brunner, M., Krauss, S., Blum, W. & Neubrand, M. (2004). Mathematikunterricht aus Sicht der PISA-Schülerinnen und -Schüler und ihrer Lehrkräfte. In J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost & U. Schiefele (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland* (S. 314–354). Münster: Waxmann.
- Becker, H. (1963). *Outsiders*. Free Press.
- Becker, M., Lüdtke, O., Trautwein, U., Köller, O. & Baumert, J. (2012). The differential effects of school tracking on psychometric intelligence: Do academic-track schools make students smarter? *Journal of Educational Psychology*, 104 (3), 682–699. doi:10.1037/a0027608
- Becker, M., Neumann, M. & Dumont, H. (2017). Recent developments in school tracking practices in Germany: An overview and outlook on future trends. *Orbis Scholae*, 10 (3), 9–25. doi:10.14712/23363177.2017.8

- Begrich, L., Praetorius, A.-K., Decristan, J., Fauth, B., Göllner, R., Herrmann, C. ... Kunter, M. (2023). Was tun? Perspektiven für eine Unterrichtsqualitätsforschung der Zukunft. *Unterrichtswissenschaft*, 51 (1), 63–97. doi:10.1007/s42010-023-00163-4
- BMBWF. (2019). *Schulleitungsprofil: Eine praxisbezogene Orientierung für effektives Schulleitungshandeln*. Wien: Bundesministerium für Bildung, Wissenschaft und Forschung.
- BMBWF. (2020). *Die Mittelschule. Änderungen ab dem Schuljahr 2020/21 im Überblick*. Verfügbar unter: https://www.bmbwf.gv.at/dam/jcr:7b6de1bc-36c1-4b54-88f0-7683120238d0/mittelschule_2020.pdf
- Braunsteiner, M.-L., Fischer, C., Kernbichler, G., Prengel, A. & Wohllhart, D. (2019). *Erfolgreich lernen und unterrichten in Klassen mit hoher Heterogenität*. doi:10.17888/nbb2018-2-1
- Bruneforth, M., Vogtenhuber, S., Lassnigg, L., Oberwimmer, K., Gumpoldsberger, H., Feyerer, E. ... Herzog-Punzenberger, B. (2016). *Indikatoren C: Prozessfaktoren*. doi:10.17888/nbb2015-1-C.3
- Clausen, M. (2002). *Unterrichtsqualität: eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstrukt- und Kriteriumsvalidität*. Waxmann.
- Dar, Y. & Resh, N. (1986). Classroom intellectual composition and academic achievement. *American Educational Research Journal*, 23 (3), 357–374. doi:10.2307/1163053
- Desimone, L. M., Hochberg, E. D. & McMaken, J. (2016). Teacher knowledge and instructional quality of beginning teachers: Growth and linkages. *Teachers College Record: The Voice of Scholarship in Education*, 118 (5), 1–54. doi:10.1177/016146811611800507
- Eder, F. (2002). Fördern Leistungsgruppen das Lernen? Der Förderanspruch von Leistungsgruppen im Lichte von PISA und TIMSS. *Erziehung und Unterricht*, 152(7–8), 979–1000.
- Eder, F. (2011). Wie gut sind Musikhauptschulen? In H. Altrichter, M. Heinrich & K. Soukup-Altrichter (Hrsg.), *Schulentwicklung durch Schulprofilierung?* (S. 165–193). Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/978-3-531-92825-8_5
- Eder, F. & Svecnik, E. (2015). Konzept und methodische Anlage der NMS-Evaluierung. In F. Eder, H. Altrichter, F. Hofmann & C. Weber (Hrsg.), *Evaluation der Neuen Mittelschule (NMS). Befunde aus den Anfangskohorten* (S. 77–94). Graz: Leykam.
- Field, A. P. (2005). Intraclass Correlation. In B. S. Everitt & D. C. Howell (Hrsg.), *Encyclopedia of Statistics in Behavioral Science*. Hoboken: Wiley. doi:10.1002/0470013192.bsa313
- Gamoran, A., Nystrand, M., Berends, M. & LePore, P. C. (1995). An organizational analysis of the effects of ability grouping. *American Educational Research Journal*, 32 (4), 687–715. doi:10.2307/1163331
- George, A. C. & Robitzsch, A. (2015). „Fairer Vergleich“: Technische Dokumentation – BIST-Ü Mathematik, 4. Schulstufe, 2013. Salzburg: BIFIE. <https://www.bifie.at/node/2766>
- Gruehn, S. (2000). *Unterricht und schulisches Lernen. Schüler als Quellen der Unterrichtsbeschreibung*. Münster: Waxmann.
- Hargreaves, D. H. (1967). *Social relations in secondary school*. London: Routledge and Kegan Paul.
- Harker, R. & Tymms, P. (2004). The effects of student composition on school outcomes. *School Effectiveness and School Improvement*, 15 (2), 177–199. doi:10.1076/sesi.15.2.177.30432
- Hattie, J. (2009). *Visible Learning*. London: Routledge.
- Helm, C., Kemethofer, D., Moosbrugger, R., Bröderbauer, S. & Luthe, S. (2016). Effekte der aus Schulleitersicht wahrgenommenen Belastungen des schulischen Lernens auf

- die Unterrichtsgestaltung und die Schülerkompetenzen in Mathematik sowie auf die Schulzufriedenheit. *Zeitschrift für Bildungsforschung*, 6 (3), 245–264. doi:10.1007/s35834-016-0165-0
- Herzog-Punzenberger, B. (2018). *Policy Brief #6*. AK Wien: Wien. Verfügbar unter: <https://wien.arbeiterkammer.at/migration-und-mehrsprachigkeit>
- Hosenfeld, I., Köller, O. & Baumert, J. (1999). Why sex differences in mathematics achievement disappear in German secondary schools. A reanalysis of the German TIMSS-data. *Studies in Educational Evaluation*, 25 (2), 143–161.
- Klieme, E. & Baumert, J. (2001). *TIMSS – Impulse für Schule und Unterricht: Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumentation*. Bonn: Bundesministerium für Bildung und Forschung.
- Kocaj, A., Kuhl, P., Kroth, A. J., Pant, H. A. & Stanat, P. (2014). Wo lernen Kinder mit sonderpädagogischem Förderbedarf besser? Ein Vergleich schulischer Kompetenzen zwischen Regel- und Förderschulen in der Primarstufe. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 66 (2), 165–191. doi:10.1007/s11577-014-0253-x
- Köller, O. & Baumert, J. (2001). Leistungsgruppierungen in der Sekundarstufe I. Ihre Konsequenzen für die Mathematikleistung und das mathematische Selbstkonzept der Begabung. *Zeitschrift für Pädagogische Psychologie*, 15 (2), 99–110. doi:10.1024//1010-0652.15.2.99
- Kuger, S., Kluczniok, K., Kaplan, D. & Rossbach, H.-G. (2016). Stability and patterns of classroom quality in German early childhood education and care. *School Effectiveness and School Improvement*, 27 (3), 418–440. doi:10.1080/09243453.2015.1112815
- Kulik, J. A. & Kulik, C.-L. C. (1992). Meta-analytic findings on grouping programs. *Gifted Child Quarterly*, 36 (2), 73–77. doi:10.1177/001698629203600204
- Kunter, M., Brunner, M., Baumert, J., Klusmann, U., Krauss, S., Blum ... Neubrand, M. (2005). Der Mathematikunterricht der PISA-Schülerinnen und -Schüler. Schulformunterschiede in der Unterrichtsqualität. *Zeitschrift für Erziehungswissenschaft*, 8 (4), 502–520. doi:10.1007/s11618-005-0156-8
- Lavrijsen, J., Dockx, J., Struyf, E. & Verschueren, K. (2022). Class composition, student achievement, and the role of the learning environment. *Journal of Educational Psychology*, 114 (3), 498–512. doi:10.1037/edu0000709
- Lehmann, R. & Lenkeit, J. (2008). *ELEMENT. Erhebung zum Lese- und Mathematikverständnis. Entwicklung in den Jahrgangsstufen 4 bis 6 in Berlin*. Humboldt Universität.
- Liu, W. C., Wang, C. K. J. & Parkins, E. J. (2005). A longitudinal study of students' academic self-concept in a streamed setting: the Singapore context. *The British Journal of Educational Psychology*, 75 (Pt 4), 567–586. doi:10.1348/000709905X42239
- Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B. & d'Apollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research*, 66 (4), 423. doi:10.2307/1170650
- Luschei, T. F. & Jeong, D. W. (2018). Is teacher sorting a global phenomenon? Cross-national evidence on the nature and correlates of teacher quality opportunity gaps. *Educational Researcher*, 47 (9), 556–576. doi:10.3102/0013189X18794401
- Maaz, K., Trautwein, U., Lüdtke, O. & Baumert, J. (2008). Educational transitions and differential learning environments: How explicit between-school tracking contributes

- to social inequality in educational outcomes. *Child Development Perspectives*, 2 (2), 99–106. doi:10.1111/j.1750-8606.2008.00048.x
- Marsh, H. W., Trautwein, U., Lüdtke, O., Baumert, J. & Köller, O. (2007). The Big-Fish-Little-Pond Effect: Persistent negative effects of selective high schools on self-concept after graduation. *American Educational Research Journal*, 44 (3), 631–669. doi:10.3102/0002831207306728
- Maulana, R., Opdenakker, M.-C. & Bosker, R. (2014). Teacher-student interpersonal relationships do change and affect academic motivation: a multilevel growth curve modelling. *The British Journal of Educational Psychology*, 84 (Pt 3), 459–482. doi:10.1111/bjep.12031
- Oakes, J. (1985). *Keeping track: How schools structure inequality*. Yale University Press.
- Polikoff, M. S. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education*, 121 (2), 183–212. doi:10.1086/679390
- Praetorius, A.-K., Klieme, E., Herbert, B. & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of three basic dimensions. *ZDM*, 50 (3), 407–426. doi:10.1007/s11858-018-0918-4
- Reindl, M. (2021). Peer group embeddedness and academic motivation: A developmental perspective. *Frontiers in Psychology*, 12, 701600. doi:10.3389/fpsyg.2021.701600
- Renkl, A. (1997). *Lernen durch Lehren: Zentrale Wirkmechanismen beim kooperativen Lernen*. Wiesbaden: Deutscher Universitätsverlag.
- Retelsdorf, J., Becker, M., Köller, O. & Möller, J. (2012). Reading development in a tracked school system: A longitudinal study over 3 years using propensity score matching. *The British Journal of Educational Psychology*, 82 (Pt 4), 647–671. doi:10.1111/j.2044-8279.2011.02051.x
- Richer, S. (1976). Reference-group theory and ability grouping: A convergence of sociological theory and educational research. *Sociology of Education*, 49 (1), 65. doi:10.2307/2112394
- Rosenbaum, J. E. (1976). *Making Inequality: The Hidden Curriculum of High School Tracking*. New York: Wiley.
- Rowley, J. F. S., Phillips, S. F. & Ferguson, R. F. (2019). The stability of student ratings of teacher instructional practice: examining the one-year stability of the 7Cs composite. *School Effectiveness and School Improvement*, 30 (4), 549–562. doi:10.1080/09243453.2019.1620293
- Scharenberg, K. (2012). *Leistungsheterogenität und Kompetenzentwicklung: Zur Relevanz klassenbezogener Kompositionsmerkmale im Rahmen der KESS-Studie*. Münster: Waxmann. Zugl.: Dortmund, Techn. Univ., Diss., 2011.
- Schneider, W., Stumpf, E., Preckel, F. & Ziegler, A. (2012). *Projekt zur Evaluation der Begabtenklassen in Bayern und Baden-Württemberg*. Verfügbar unter: <http://www.begabungsberatungsstelle.uni-wuerzburg.de/fileadmin/99000016/>
- Sertl, M. & Leditzky, C. (2016). Schulpflicht, Leistung und Wettbewerb: Was bei der Zusammensetzung einer ersten Klasse Volksschule eine Rolle spielt. *Erziehung & Unterricht* (5–6), 387–398.
- Sharplin, E. D. (2014). Reconceptualising out-of-field teaching: Experiences of rural teachers in Western Australia. *Educational Research*, 56 (1), 97–110. doi:10.1080/00131881.2013.874160

- Sheppard, K., Padwa, L., Kelly, A. M. & Krakehl, R. (2020). Out-of-field teaching in chemistry and physics: An empirical census study. *Journal of Science Teacher Education*, 31 (7), 746–767. doi:10.1080/1046560X.2019.1702268
- Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research*, 60 (3), 471–499. doi:10.3102/00346543060003471
- Specht, W. (2011). Restschulen und Restklassen. Ein vernachlässigtes Phänomen im Gefolge neuer Steuerungsformen. In H. Altrichter, M. Heinrich & K. Soukup-Altrichter (Hrsg.), *Schulentwicklung durch Schulprofilierung? Zur Veränderung von Koordinationsmechanismen im Schulsystem* (S. 141–163). Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/978-3-531-92825-8_4
- Steenbergen-Hu, S., Makel, M. C. & Olszewski-Kubilius, P. (2016). What one hundred years of research says about the effects of ability grouping and acceleration on K–12 students' academic achievement. *Review of Educational Research*, 86 (4), 849–899. doi:10.3102/0034654316675417
- Tajfel, H. & Turner, J. C. (2004). The social identity theory of intergroup behavior. In J. T. Jost & J. Sidanius (Hrsg.), *Political Psychology* (S. 276–293). New York: Psychology Press. doi:10.4324/9780203505984-16
- Tan, C. Y. & Dimmock, C. (2022). The relationships among between-class ability grouping, teaching practices, and mathematics achievement: A large-scale empirical analysis. *Educational Studies*, 48 (4), 471–489. doi:10.1080/03055698.2020.1780416
- Taylor, J., Banilower, E. & Clayton, G. (2020). National trends in the formal content preparation of US science teachers: Implications of out-of-field teaching for student outcomes. *Journal of Science Teacher Education*, 31 (7), 768–779. doi:10.1080/1046560X.2020.1762992
- Van Houtte, M. (2006). School type and academic culture: evidence for the differentiation-polarization theory. *Journal of Curriculum Studies*, 38 (3), 273–292. doi:10.1080/00220270500363661
- van Ophuysen, S. & Wendt, H. (2010). Zur Veränderung der Mathematikleistung von Klasse 4 bis 6. Welchen Einfluss haben Kompositionen und Unterrichtsmerkmale? In J. Baumert, K. Maaz & U. Trautwein (Hrsg.), *Bildungsentscheidungen* (S. 302–327). Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/978-3-531-92216-4_13
- Vanfossen, B. E., Jones, J. D. & Spade, J. Z. (1987). Curriculum tracking and status maintenance. *Sociology of Education*, 60 (2), 104. doi:10.2307/2112586
- Vogtenhuber, S., Lassnigg, L. & Bruneforth, M. (2012). Indikatoren A: Kontext des Schul- und Bildungswesens. In M. Bruneforth (Hrsg.), *Nationaler Bildungsbericht Österreich 2012* (S. 15–30). Graz: Leykam.
- Weber, C., Bacher, J., Altrichter, H. & Leitgöb, H. (2015). Sind Leistungsverbesserungen an den NMS-Modellschulen zu beobachten? In F. Eder, H. Altrichter, F. Hofmann & C. Weber (Hrsg.), *Evaluation der Neuen Mittelschule (NMS). Befunde aus den Anfangskohorten* (S. 241–264). Graz: Leykam.
- Weber, C. & Kemethofer, D. (2024). Der Verzicht auf Leistungsgruppen in der NMS – Welche Effekte wären dadurch zu erwarten? In L. Jesacher-Rößler & D. Kemethofer (Hrsg.), *10 Jahre Regelschule – die (Neue) Mittelschule* (S. 186–199). Münster: Waxmann.

- Weber, C., Moosbrugger, R., Hasengruber, K., Altrichter, H. & Schrod, H. (2019). *Wer unterrichtet wen? Die Zusammensetzung von Klassen und Schulen und die Zuteilung von Lehrkräften*. doi:10.17888/nbb2018-2-4
- Weldon, P. R. (2016). *Out-of-field teaching in Australian secondary schools*. Australian Council for Educational Research (ACER).
- Wu, Y.-J. & Becker, M. (2023). Association between school contexts and the development of subjective well-being during adolescence: A context-sensitive longitudinal study of life satisfaction and school satisfaction. *Journal of youth and adolescence*, 52 (5), 1039–1057. doi:10.1007/s10964-022-01727-w
- Ziegler, C. & Richter, D. (2017). Der Einfluss fachfremden Unterrichts auf die Schülerleistung: Können Unterschiede in der Klassenzusammensetzung zur Erklärung beitragen? *Unterrichtswissenschaft*, 45 (2), 136–155.

Klassenebene

Der Big-Fish-Little-Pond-Effekt im Zeitvergleich

1 Der Big-Fish-Little-Pond-Effekt im Zeitvergleich

Das akademische Selbstkonzept, definiert als die Einschätzung der eigenen fachbezogenen Fähigkeiten (Shavelson, Hubner & Stanton, 1976), ist eine zentrale Prädiktorvariable für die (intrinsische) Motivation (Guo, Marsh, Parker, Morin & Dicke, 2017), die Leistung in unterschiedlichen Fachdomänen (Arens, Schmidt & Preckel, 2019) und den Bildungsabschluss (Guay, Larose & Boivin, 2004). Eine wesentliche Quelle für das akademische Selbstkonzept von Kindern und Jugendlichen stellt neben der individuellen Leistung (Calsyn & Kenny, 1977; Guay, Marsh & Boivin, 2003) die Leistung der Klasse dar (Dijkstra, Kuyper, van der Werf, Buunk & van der Zee, 2008).

Die Orientierung an der durchschnittlichen Klassenleistung wird in der Literatur als Big-Fish-Little-Pond-Effekt (BFLPE; Marsh, 1987) bezeichnet. Dieser beschreibt den Einfluss sozialer Vergleiche auf individuelle Fähigkeitseinschätzungen. Demnach führt der soziale Vergleich mit einer leistungsstarken Klasse zu einer Reduktion des akademischen Selbstkonzepts des Individuums, wohingegen der Vergleich mit einer leistungsschwachen Klasse zu einer Erhöhung des akademischen Selbstkonzepts führt (Thijs, Verkuyten & Helmond, 2010).

Die Kalibrierung der eigenen Fähigkeitseinschätzungen über soziale Vergleichsprozesse kann als eine altersbedingte Entwicklungsaufgabe zur Identitätsformierung (Eccles & Roeser, 2009; Havighurst, 1972) verstanden werden, deren Lösung insbesondere im Jugendalter relevant wird. Soziale Vergleiche mit Gleichaltrigen werden daher im Jugendalter besonders wichtig für die Entwicklung des akademischen Selbstkonzepts (Brown & Larson, 2009). Daher ist davon auszugehen, dass der BFLPE im Jugendalter stärker ausfällt als in der Kindheit.

Basierend auf Daten der österreichischen Bildungsstandardüberprüfung wurden in der vorliegenden Trendstudie altersbedingte Unterschiede im BFLPE zwischen der vierten und achten Schulstufe untersucht. Hierbei wurde der Einfluss der individuellen Leistung und der Klassenleistung auf das akademische Selbstkonzept der Schülerinnen und Schüler analysiert.

1.1 Akademisches Selbstkonzept

Das akademische Selbstkonzept wird definiert als die Wahrnehmung und Überzeugung der eigenen fachbezogenen Fähigkeiten (Shavelson et al., 1976). Basierend auf den Annahmen des Reziproken-Effekt-Modells von Marsh (1990) wird vor allem in der Kindheit und Jugend die individuelle Leistung als zentrale Voraussetzung für die Entwicklung des akademischen Selbstkonzepts diskutiert. Demnach schätzen Schülerinnen und Schüler die eigenen fachspezifischen Fähigkeiten auf Basis ihrer individuellen fachspezifischen Leistung ein. Schülerinnen und Schüler bewerten ihre fachspezifischen Fähigkeiten sowohl durch Vergleiche innerhalb eines Fachs als auch durch Vergleiche zwischen verschiedenen Fächern. Dies bedeutet, dass sie ihre mathematischen Fähigkeiten nicht nur anhand ihrer Mathematikleistung einschätzen, sondern auch im Verhältnis zu ihren Leistungen in anderen Fächern wie beispielsweise Deutsch (I/E-Modell; Marsh, 1986). Allerdings zeigte sich, dass vor allem im Fach Mathematik der Zusammenhang zwischen der individuellen Leistung und dem akademischen Selbstkonzept stärker ist als in verbalen Fächern (Möller, Pohlmann, Köller & Marsh, 2009). Eine Erklärung hierfür ist, dass Einschätzungen im Fach Mathematik überwiegend auf schulischen Erfahrungen beruhen, wohingegen für die Einschätzung verbaler Fähigkeiten mehr außerschulische Quellen (z. B. das Lesen von Büchern, Fernsehen oder Onlinespiele) genutzt werden (Wu, Guo, Yang, Zhao & Guo, 2021).

Allerdings hängt die Wahrnehmung der eigenen Fähigkeiten mit steigendem Alter nicht mehr nur von den eigenen Leistungen ab (Calsyn & Kenny, 1977; Guay et al., 2003), sondern wird ebenfalls durch den sozialen Vergleich mit der Leistung einer oder mehrerer Referenzgruppen geprägt (Drücke-Noe, Gniewosz & Paasch, 2022; Mussweiler & Rüter, 2003). Diese sozialen Vergleichsprozesse (Shavelson et al., 1976) werden entscheidend für die Einschätzung der eigenen Fähigkeiten.

1.2 Soziale Vergleiche und der Big-Fish-Little-Pond Effekt

Vor dem Hintergrund der Theorie sozialer Vergleiche (Festinger, 1954) bewerten Individuen ihre eigenen Fähigkeiten in Abhängigkeit ihrer sozialen Umgebung. Die Schule dient hierbei nicht nur der Vermittlung von Wissen und Normen, sondern bietet Kindern und Jugendlichen den idealen Ort, sich in Leistung und Begabung mit anderen Schülerinnen und Schülern zu vergleichen (Hummrich & Kramer, 2017). Dabei orientieren sich Schülerinnen und Schüler vor allem an jenen Personen, die zum einen der eigenen Person ähnlich sind (z. B. Alter, Leistung) und zum anderen an Personen, mit denen sie regelmäßig interagieren (Crusius, Corcoran & Mussweiler, 2022; Mussweiler & Rüter, 2003). Die Schulklasse im österreichischen Schulsystem entspricht nicht nur dem Aspekt der Ähnlichkeit aufgrund selektiver Kriterien der Klassenzusammensetzung (z. B. Altershomogenität; Jahr, 2022), sondern ermöglicht auch regelmäßige Interaktionen mit anderen Schülerinnen und Schülern innerhalb der Schulklasse. Durch den regelmäßigen Austausch untereinander erhalten Schülerinnen und Schüler die relevanten Informationen (z. B. Testergebnisse,

Schulnoten) für die Bewertung der eigenen schulischen Fähigkeiten (Rubin, Bukowski & Bowker, 2015).

Auf der theoretischen Prämisse sozialer Vergleiche basiert auch der BFLPE (Marsh, 1987). Dieser beschreibt, dass Schülerinnen und Schüler ihre eigenen Leistungen mit den Leistungen ihrer Klassenkameradinnen und Klassenkameraden vergleichen und als Grundlage für die Bildung ihres akademischen Selbstkonzepts nutzen. Studien zum BFLPE berücksichtigen dabei die durchschnittliche Klassenleistung als zentrale Bezugsgröße. Die Durchschnittsleistung der Klasse dient als Maßstab, anhand dessen Schülerinnen und Schüler ihre eigene Leistung einordnen und bewerten. Das akademische Selbstkonzept eines Individuums wird demnach nicht nur von den eigenen schulischen Leistungen im Sinne des Reziproken-Effekt-Modells (Marsh, 1990) beeinflusst, sondern ebenfalls durch die klassenspezifische Leistungsstärke. Allerdings unterscheidet sich die Wirkrichtung beider Einflüsse. Während die individuelle Leistung positiv mit dem akademischen Selbstkonzept des Individuums korreliert (Jaiswal & Choudhuri, 2017), wirkt sich die durchschnittliche Klassenleistung negativ auf das akademische Selbstkonzept der Schülerinnen und Schüler aus. Das bedeutet, soziale Vergleiche mit einer leistungsstärkeren Klasse (Aufwärtsvergleiche; „kleiner Fisch im großen Teich“) führen zu einer Verringerung des akademischen Selbstkonzepts. Hingegen führen soziale Vergleiche mit einer leistungsschwächeren Klasse (Abwärtsvergleiche; „großer Fisch im kleinen Teich“) zu einer Erhöhung des akademischen Selbstkonzepts (Hoferichter, Lätsch, Lazarides & Raufelder, 2018). Infolgedessen können Schülerinnen und Schüler mit vergleichbaren akademischen Leistungen durch die jeweilige Leistungsstärke der Klasse zu unterschiedlichen Selbsteinschätzungen ihrer akademischen Fähigkeiten gelangen. Diese Mechanismen unterstreichen die Bedeutung des sozialen Kontexts und sozialer Vergleichsprozesse in der Bildung des akademischen Selbstkonzepts.

Studienergebnisse zeigten, dass der BFLPE ein robustes Phänomen darstellt, welches über unterschiedliche Länder und Schultypen hinweg belegt wurde (Fang, Huang, Zhang, Huang, Li & Yuan, 2018). Dies gilt gleichermaßen für verschiedene Altersgruppen, wie für die Primarstufe (Kavanagh, 2020; Televantou, Marsh, Dicke & Nicolaides, 2021) als auch die Sekundarstufe (Huguet et al., 2009). Allerdings gibt es auch Hinweise darauf, dass die Stärke des Effekts von der Kindheit bis zum Jugendalter variiert (Stockus & Zell, 2023).

1.3 Altersbedingte Unterschiede im Big-Fish-Little-Pond-Effekt

Erklärungen für die Variation des BFLPE über die Altersgruppen hinweg finden sich in entwicklungsbedingten Unterschieden des akademischen Selbstkonzepts und der Bedeutung sozialer Vergleiche hierfür. Mit zunehmendem Alter kommt es zu einer realistischeren Selbsteinschätzung der eigenen Fähigkeiten (Marsh & Martin, 2011). Aufgrund wiederholter Erfahrungen in einem Fach wird die eigene Einschätzung realistischer und stabiler. Gleichzeitig gewinnen soziale Vergleiche in diesem Kalibrierungsprozess immer mehr an Bedeutung (Eccles & Roeser, 2009; Muenks, Wigfield & Eccles, 2018). Mitverantwortlich

ist hierfür die steigende Peerorientierung im Jugendalter (Brown & Larson, 2009; Eccles & Midgley, 1989), wodurch auch Klassenkameradinnen und Klassenkameraden zu einer immer wichtigeren Referenzgruppe im Jugendalter werden, verglichen zur Kindheit. Das akademische Selbstkonzept von Kindern hängt kaum mit externalen Indikatoren (Fähigkeiten, Erfolge oder Leistungen von anderen Personen) zusammen (Marsh et al., 2015) und ist oftmals (zu) sehr positiv ausgeprägt. Jugendliche schätzen ihre eigenen Leistungen hingegen realistischer ein (Stockus & Zell, 2023) und entwickeln ein differenziertes akademisches Selbstkonzept (Marsh et al., 2015), das stärker durch andere Personen und die sozialen Vergleiche mit diesen geprägt ist.

Es gibt bereits erste Hinweise, dass der BFLPE im Jugendalter stärker ausfällt als in der Kindheit. In einer Metaanalyse von Fang et al. (2018) wurde das Alter der Schülerinnen und Schüler als potenzieller Moderator für den BFLPE im Fach Mathematik untersucht. Es zeigte sich, dass die Stärke des BFLPE auf der Primarstufe (Primary School), die alle Volksschülerinnen und -schüler umfasst, am geringsten ist und mit zunehmendem Alter (Middle und High School) ansteigt. Ähnliche Ergebnisse zeigten sich auch in Studien, die Daten der „Trends in International Mathematics and Science Study“ (TIMSS) nutzten. Diese ermöglichen die Untersuchung altersbedingter Unterschiede im BFLPE zwischen der vierten und achten Schulstufe unter Einbezug der an der Studie teilnehmenden Staaten. In der Studie von Liou (2014), die Schülerinnen und Schüler aus Taiwan untersuchte, zeigte sich ein stärkerer BFLPE auf der achten Schulstufe als auf der vierten Schulstufe. Allerdings wurde hier auf die gesamte Schule als Referenzgruppe fokussiert. Marsh et al. (2015) untersuchten die entwicklungsbezogene und interkulturelle Generalisierbarkeit des BFLPE zwischen Schülerinnen und Schülern der vierten und achten Schulstufe aus 13 verschiedenen Ländern (Asien, Australien, Vereinigte Staaten, Europa und Mittlerer Osten). Über alle Länder hinweg zeigte sich ein signifikant stärkerer Effekt der aggregierten Klassenleistung auf das akademische Selbstkonzept auf der achten im Vergleich zur vierten Schulstufe (Marsh et al., 2015). Obwohl bereits erste Studien hinsichtlich altersbedingter Unterschiede im BFLPE vorliegen, gibt es für Österreich noch keine vergleichbaren Ergebnisse.

2 Die vorliegende Studie: Big-Fish-Little-Pond-Effekt im Zeitvergleich

Das Ziel des vorliegenden Beitrages ist es, Unterschiede in der Stärke des BFLPE zwischen der vierten und achten Schulstufe im Fach Mathematik erstmalig anhand von Daten aus der Bildungsstandardüberprüfung in Österreich zu untersuchen. Dazu wird sowohl auf der vierten als auch auf der achten Schulstufe das akademische Selbstkonzept der Schülerinnen und Schüler in Mathematik durch die individuelle Leistung und durch die aggregierte Klassenleistung im Fach Mathematik vorhergesagt.

Basierend auf den Annahmen des BFLPE (Marsh, 1987) wird von einem positiven Effekt der individuellen Leistung auf das akademische Selbstkonzept der Schülerinnen und Schüler sowohl auf der vierten als auch auf der achten Schulstufe ausgegangen. Dagegen wird

ein negativer Effekt der aggregierten Klassenleistung auf das akademische Selbstkonzept des Individuums, bekannt als der BFLPE, erwartet. Vor dem Hintergrund entwicklungsbedingter Unterschiede (Muenks et al., 2018) wird zudem davon ausgegangen, dass ...

1. die Einschätzung der eigenen akademischen Fähigkeiten auf der achten Schulstufe akkurater ausfällt als auf der vierten Schulstufe. Das bedeutet, dass der Effekt der individuellen Leistung auf das akademische Selbstkonzept über die Zeit hinweg ansteigen sollte.
2. der erwartete BFLPE für Schülerinnen und Schüler der achten Schulstufe stärker ist als für Schülerinnen und Schüler der vierten Schulstufe.

Zudem wird der Frage nachgegangen, ob die Effekte der individuellen Leistung (Akkuratheit), der aggregierten Klassenleistung (BFLPE) und die erwarteten Altersunterschiede unter Kontrolle von einbezogenen Hintergrundvariablen (Geschlecht, Sozialstatus und Migrationshintergrund) variieren.

Die vorliegende Studie geht daher in mehreren Punkten über die bestehende Befundlage hinaus. Erstens gibt es kaum Untersuchungen, die Unterschiede in der Stärke des BFLPE anhand von Populationsdaten untersuchen. Der Vorteil der verwendeten Datenquelle aus der Bildungsstandardüberprüfung ist, dass hierbei Stichprobenfehler nicht zu einer Verzerrung der Ergebnisse führen können. Des Weiteren werden in den Analysen zum BFLPE selten Kontrollvariablen berücksichtigt (Dai & Rinn, 2008). Dies ist allerdings notwendig, da sowohl die Einschätzung der eigenen Leistungsfähigkeit als auch soziale Vergleichsprozesse durch soziale und persönliche Hintergrundvariablen wie Geschlecht (Guimond & Chatard, 2014), Sozialstatus (Seaton, Marsh & Craven, 2010) und Migrationshintergrund mitbestimmt werden (Dai & Rinn, 2008) und dadurch den BFLPE beeinflussen können.

3 Methode

3.1 Datengrundlage

Die Analysen beziehen sich auf Daten der Bildungsstandardüberprüfung (nähere Details siehe George, Schreiner, Wiesner, Pointinger & Pacher, 2019) im Fach Mathematik. Diese umfassen 73.655 Schülerinnen und Schüler der vierten Schulstufe (Volksschule) im Jahr 2013 (49,1 % weiblich, 18,7 % mit Migrationshintergrund) und 72.704 Schülerinnen und Schüler der achten Schulstufe im Jahr 2017 (49,2 % weiblich, 20,5 % mit Migrationshintergrund). Die Schülerinnen und Schüler der achten Schulstufe befinden sich im österreichischen Schulsystem auf der Sekundarstufe I. Diese umfasst Volksschuloberstufe, Hauptschule/ Neue Mittelschule (NMS) und allgemeinbildende höhere Schule (AHS-Unterstufe; siehe Schreiner, Breit, Pointinger, Pacher, Neubacher & Wiesner, 2018). Die Daten stellen eine Vollerhebung der österreichischen Schülerpopulation dar.

3.2 Instrumente

3.2.1 Akademisches Selbstkonzept

Das akademische Selbstkonzept in Mathematik wurde sowohl auf der vierten Schulstufe als auch auf der achten Schulstufe mit einem standardisierten Instrument erfasst (Freunberger, Robitzsch & Pham, 2014). Die verwendete Skala besteht in beiden Schulstufen aus vier Items („Normalerweise bin ich gut in Mathematik“, „Mathematik fällt mir schwerer als meinen Mitschülerinnen & Mitschülern“, „Ich lerne schnell in Mathematik“, „Ich bin einfach nicht gut in Mathematik“) mit einem vierstufigen Antwortformat (stimme völlig zu [1] – stimme überhaupt nicht zu [4]). Um eine korrekte Interpretation der Ergebnisse zu sichern, wurden die positiv formulierten Items rekodiert. Ein hoher Skalenwert bedeutet demnach ein hohes akademisches Selbstkonzept.

Die Ergebnisse der konfirmatorischen Faktorenanalysen in Mplus 7 (Muthén & Muthén, 2017) zeigten sowohl auf der vierten Schulstufe ($CFI = .99$, $TLI = .99$, $RMSEA = .04$) als auch auf der achten Schulstufe ($CFI = .99$, $TLI = .98$, $RMSEA = .08$) eine akzeptable bis gute Modellpassung (Hu & Bentler, 1998, 1999; Little, 2013). Die Konstruktreliabilität (Fornell & Larcker, 1981) beträgt auf der vierten Schulstufe $\Omega = .79$ und auf der achten Schulstufe $\Omega = .87$.

3.2.2 Leistung

Die Leistung der Schülerinnen und Schüler wurde mithilfe von standardisierten Tests erhoben (nähere Details zu den Erhebungen siehe George et al., 2019). Diese bieten den Vorteil, dass sie ein objektives und einheitliches Maß für die Leistung aller Schülerinnen und Schüler darstellen. Sie ermöglichen es, Unterschiede in den Leistungen zwischen den Individuen und den Durchschnittsleistungen der Klassen konsistent zu messen (Vogl, Schmidt & Preckel, 2018). Auf der vierten Schulstufe erfolgte die Leistungsüberprüfung in den Teildimensionen Arbeiten mit Zahlen, Operationen, Größen, Ebenen und Raum, Modellieren, Operieren, Kommunizieren und Problemlösen (BIFIE, 2017). Auf der achten Schulstufe umfasste es die Dimensionen Darstellen und Modellbilden, Rechnen und Operieren, Interpretieren, Argumentieren und Begründen, Zahlen und Maß, funktionale Abhängigkeiten, geometrische Figuren und statistische Darstellung (BIFIE, 2018).

Die Leistungsüberprüfungen wurden in Form eines Multi-Matrix-Block-Designs für jede Schulstufe zusammengefasst (Kiefer, Kuhn & Feller, 2016). Das bedeutet, dass die Schülerinnen und Schüler nur eine Auswahl an Testaufgaben bearbeiten mussten. Basierend auf den vorhandenen Antworten erfolgte die Schätzung der fehlenden Werte (siehe Punkt 3.3.; Wiesner, Schreiner & Breit, 2020).

Zur Berechnung der Modelle wurden die Leistungen der Schülerinnen und Schüler anhand der Weighted-Likelihood-Schätzer (WLE; Trendtel, Pham & Yanagida, 2016; Warm, 1989) aus den vorliegenden Daten berechnet und verwendet. Um unverzerrte Standardfehler der Effektschätzer zu erhalten, werden im Large-Scale Assessment Forschungsfragen auf Populationsebene, wie jene in diesem Kapitel, idealerweise unter Einbezug von Plausible Values und multiplen imputierten Kontextdaten diskutiert (Robitzsch, Pham &

Yanagida, 2016). Aufgrund technischer Probleme bei der Berechnung des Modells für die achte Schulstufe wurden in diesem Kapitel die Effekte auf Basis der WLEs der Schülerinnen und Schüler untersucht. Wir erwarten, dass deren Stärke bei Verwendung von Plausible Values mindestens vergleichbar ausfällt.

3.2.3 Kontrollvariablen

Aufgrund möglicher Unterschiede in der Ausprägung des akademischen Selbstkonzepts hinsichtlich des Geschlechts (Goldman & Penner, 2016), des Sozialstatus (HISEI; Seaton et al., 2010) und des Migrationshintergrunds (Dai & Rinn, 2008) wurden in einem zusätzlichen Zwei-Ebenen-Modell, getrennt für die beiden Schulstufen, diese Variablen berücksichtigt.

3.3 Datenanalyse

Die Grundlage der Analysen bildeten die Rohdatensätze der Bildungsstandardüberprüfung der vierten und achten Schulstufe. Die Schätzung der Parameter aufgrund fehlender Werte sowohl durch das Multi-Matrix-Design als auch durch Item-non-response erfolgte über den Weighted-Least-Squares-Mean-and-Variance-Adjusted-Schätzer (WLSMV). Dieser ist analog zur Full-Information-Maximum-Likelihoodschätzung (siehe Asparouhov & Muthén, 2010) zu interpretieren.

Zur Überprüfung der Hypothesen wurde ein zweistufiges Doubly-Latent-Mehr-Ebenen-Modell (Marsh et al., 2009) in Mplus 7 (Muthén & Muthén, 2017) berechnet. Hierfür wurden vorab einige modellspezifische Spezifikationen vorgenommen. Erstens wurden sowohl das akademische Selbstkonzept als abhängige Variable als auch die Leistung als Prädiktorvariable auf der Individualebene (Level 1) und auf der Klassenebene (Level 2) latent spezifiziert. Diese latente Spezifikation liefert eine genauere Schätzung und ist valider als selektive manifeste Variablen (Leonhart, Wirtz & Bengel, 2008). Die Faktorladungen des akademischen Selbstkonzepts wurden ebenfalls auf beiden Ebenen gleichgesetzt. Diese Invarianz der levelübergreifenden Faktorladungen gewährleistet, dass die Interpretation des akademischen Selbstkonzepts zwischen den Ebenen gleich ist (Marsh et al., 2009). Die Faktorladungen der Leistungsvariable wurden auf beiden Ebenen auf den Wert eins fixiert und die Residualvarianz auf null gesetzt. Dies ist nach Geiser (2011) erforderlich, um die Metrik des Faktors zu definieren.

Zweitens wurde zur leichteren Interpretierbarkeit der resultierenden Regressionskoeffizienten und im Einklang mit der Modellierung des BFLPE in der Literatur die Leistung am Gesamtmittelwert der Leistung zentriert und standardisiert (Marsh et al., 2012). Nach Enders und Tofghi (2007) ist dies immer dann notwendig, wenn auf Klassenebene (Level 2) der Effekt der aggregierten Leistung von Interesse ist und für die individuelle Leistung auf Level 1 kontrolliert wird. Aufgrund der Zentrierung am Gesamtmittelwert (Marsh et al., 2012) wurde der BFLPE als zusätzlicher Parameter spezifiziert, indem der Regressionskoeffizient auf Klassenebene vom Regressionskoeffizienten auf Individualebene subtrahiert

wurde. Um die Effekte des BFLPE standardisiert darzustellen und vergleichbar zu machen, wurden die Effektgrößen (ES) nach Hox, Moerbeek und van de Schoot (2017) berechnet.

Eine Voraussetzung für die Modellierung eines Mehr-Ebenen-Modells ist eine Intra-Klassen-Korrelation (ICC) > 0 (Tausendpfund, 2020). Mithilfe der ICC kann das Verhältnis der Varianz des akademischen Selbstkonzepts zwischen den Gruppen zur Gesamtvarianz bewertet werden (Castro, 2002; Lüftenegger, Schober, van de Schoot, Wagner, Finsterwald & Spiel, 2012). Das bedeutet, jener Anteil der Varianz des akademischen Selbstkonzepts, der auf die Klassenebene fällt. Die Ergebnisse zeigen eine ICC auf der vierten Schulstufe von ICC = 0.01 und auf der achten Schulstufe von ICC = 0.03.

In den Modellen 1a (vierte Schulstufe) und 1b (achte Schulstufe) werden unter Berücksichtigung von Mess- und Stichprobenfehler die Effekte der Leistung auf das akademische Selbstkonzept auf der Individualebene (Level 1) und der Klassenebene (Level 2) getrennt spezifiziert. Die Differenz zwischen den beiden Effekten (Level 2–Level 1) stellt den kontextuellen Effekt des BFLPE dar.

In den Modellen 2a (vierte Schulstufe) und 2b (achte Schulstufe) werden darüber hinaus die Kontrollvariablen hinzugefügt. Da es sich bei den Kontrollvariablen (Geschlecht, HISEI und Migrationshintergrund) um individuelle Ausprägungen der Schülerinnen und Schüler handelt, wurden diese auf der Individualebene (Level 1) eingefügt. Geschlecht (codiert 1 weiblich, 2 männlich) und Migrationshintergrund (codiert 0 ohne, 1 mit) wurden dabei als kategoriale Variablen und HISEI als metrische Variable definiert.

Tabelle 1: Güte der Modellanpassung

	BFLPE ohne Kontrollvariablen		BFLPE mit Kontrollvariablen	
	Modell 1a	Modell 1b	Modell 2a	Modell 2b
RMSEA	.04	.04	.04	.04
CFI	.98	.98	.98	.98
TLI	.96	.97	.97	.97
χ^2	1261.03***	1330.32***	2013.49***	2041.73***
df	12	12	21	21
p	< .001	< .001	< .001	< .001
n	73.655	72.704	73.655	72.704

Anmerkungen: RMSEA (Root Mean Square Error of Approximation), CFI (Comparative Fit Index), TLI (Tucker-Lewis Index), χ^2 (chi-square), df (degrees of freedom), *** $p < .001$.

Die angeführten Fit-Indizes dienen dazu, die Güte der Modellpassung zu beurteilen und zu interpretieren. Diese Indizes geben Auskunft darüber, wie gut das postulierte Modell die beobachteten Daten beschreibt und ob es angemessen ist, die theoretischen Annahmen des Modells zu akzeptieren. Der Root Mean Square Error of Approximation (RMSEA) ist ein wichtiges Gütemaß in der Strukturgleichungsmodellierung, das die Anpassung eines Modells an die Daten bewertet. RMSEA berücksichtigt die Modellkomplexität und bevorzugt Modelle mit weniger Parametern. CIF (Comparative Fit Index) und TLI (Tucker-Lewis Index) hingegen sind normierte Indizes, die die relative Passung des Modells zu einem Basismodell bewerten, wobei Werte nahe 1 auf eine gute Passung hinweisen. Der χ^2 -Test vergleicht die beobachteten und erwarteten Häufigkeiten und prüft, ob das Modell die Daten gut beschreibt. Die Güte der Modellanpassung (siehe Tabelle 1) zeigte bei allen durchgeführten Modellen eine akzeptable bis sehr gute Passung (Little, 2013).

4 Ergebnisse¹

Deskriptiv zeigten sich signifikant positive Korrelationen zwischen dem akademischen Selbstkonzept und der Leistung sowohl auf der vierten Schulstufe als auch auf der achten Schulstufe. Auch die Kontrollvariablen (Geschlecht, HISEI und Migrationshintergrund) korrelieren in beiden Schulstufen signifikant mit den latenten Konstrukten akademisches Selbstkonzept und Leistung (siehe Tabelle 2).

Tabelle 2: Zwei-Ebenen-Korrelationen der vierten und achten Schulstufe

	Selbstkonzept	Leistung	Geschlecht	HISEI	Migration
Vierte Schulstufe					
Level 1					
Selbstkonzept	1.00				
Leistung	.60***	1.00			
Geschlecht	.26***	.10***	1.00		
HISEI	.23***	.31***	-.00	1.00	
Migration	-.12***	-.29***	-.00	-.39***	1.00
Level 2					
Selbstkonzept	1.00				
Leistung	.39***	1.00			

1 Da sich bei der Analyse von Populationsdaten Signifikanztests erübrigen, haben die angegebenen Irrtumswahrscheinlichkeiten ausschließlich informativen Charakter.

Achte Schulstufe					
Level 1					
Selbstkonzept	1.00				
Leistung	.57***	1.00			
Geschlecht	.22***	.11***	1.00		
HISEI	.14***	.20***	.01*	1.00	
Migration	-.07***	-.23***	-.00	-.38***	1.00
Level 2					
Selbstkonzept	1.00				
Leistung	.54***	1.00			
Anmerkungen: *** $p < .001$, * $p < .05$					

Gemäß den Hypothesen zeigten die Zwei-Ebenen-Modelle (siehe Tabelle 3) ohne Kontrollvariablen (Modelle 1a & 1b) statistisch signifikant negative BFLPE auf das akademische Selbstkonzept der Schülerinnen und Schüler sowohl für die vierte Schulstufe (Modell 1a) $B = -0.37$, $SE = 0.01$, $p < .001$, $ES = -.23$ als auch für die achte Schulstufe (Modell 1b) $B = -0.40$, $SE = 0.01$, $p < .001$, $ES = -.36$. Der Effekt war auf der achten Schulstufe stärker als auf der vierten Schulstufe. Auch unter Berücksichtigung der Kontrollvariablen (Modelle 2a & 2b) zeigten sich statistisch signifikant negative BFLPE sowohl auf der vierten Schulstufe (Modell 2a) $B = -0.35$, $SE = 0.01$, $p < .001$, $ES = -.22$ als auch auf der achten Schulstufe (Modell 2b) $B = -0.39$, $SE = 0.01$, $p < .001$, $ES = -.36$. Der BFLPE fällt ebenfalls wie in den Modellen ohne Kontrollvariablen auf der achten Schulstufe (Modell 2b) stärker aus als auf der vierten Schulstufe (Modell 2a). Im Vergleich zu den Modellen ohne Kontrollvariablen (Modelle 1a & 1b) waren diese Effekte leicht reduziert.

Auf der Individualebene zeigten sich ohne Berücksichtigung der Kontrollvariablen (Modelle 1a & 1b) in beiden Schulstufen statistisch signifikant positive Effekte der individuellen Leistung auf das akademische Selbstkonzept (Modell 1a; $B = 0.43$, $SE = 0.00$, $p < .001$; Modell 1b, $B = 0.50$, $SE = 0.00$, $p < .001$). Der Effekt war auf der achten Schulstufe stärker als auf der vierten Schulstufe. Auch unter Berücksichtigung der Kontrollvariablen (Modelle 2a & 2b) zeigte sich der erwartete positive Effekt der individuellen Leistung auf das akademische Selbstkonzept der Schülerinnen und Schüler sowohl auf der vierten Schulstufe (Modell 2a) $B = 0.42$, $SE = 0.00$, $p < .001$ als auch auf der achten Schulstufe (Modell 2b) $B = 0.49$, $SE = 0.00$, $p < .001$. Im Vergleich zu den Modellen ohne Kontrollvariablen (Modelle 1a & 1b) waren diese Effekte leicht reduziert. Die Prädiktion des akademischen Selbstkonzepts durch die individuelle Leistung fällt ebenfalls wie in den Modellen ohne Kontrollvariablen auf der achten Schulstufe (Modell 2b) stärker aus als auf der vierten Schulstufe (Modell 2a).

Die Kontrollvariable Geschlecht zeigte positive Effekte sowohl auf der vierten Schulstufe als auch auf der achten Schulstufe. Demnach weisen Buben in beiden Schulstufen ein höheres

akademisches Selbstkonzept auf als Mädchen. Schülerinnen und Schüler mit einem höheren Sozialstatus (HISEI) zeigten ebenfalls in beiden Schulstufen ein höheres akademisches Selbstkonzept. Bei der Variable Migrationshintergrund zeigte sich sowohl auf der vierten als auch auf der achten Schulstufe, dass Schülerinnen und Schüler mit Migrationshintergrund ein höheres akademisches Selbstkonzept aufweisen als Schülerinnen und Schüler ohne Migrationshintergrund.

Tabelle 3: Zwei-Ebenen-Modelle zur Überprüfung des BFLPE der vierten (a) und achten (b) Schulstufe

	BFLPE ohne Kontrollvariablen				BFLPE mit Kontrollvariablen			
	Modell 1a		Modell 1b		Modell 2a		Modell 2b	
	<i>B</i>	<i>SE</i>	<i>B</i>	<i>SE</i>	<i>B</i>	<i>SE</i>	<i>B</i>	<i>SE</i>
Level 1								
SL ^a	0.43***	0.00	0.50***	0.00	0.42***	0.00	0.49***	0.00
Sex ^b					0.14***	0.00	0.11***	0.00
HISEI					0.00***	0.00	0.00***	0.00
Mig ^c					0.05***	0.00	0.06***	0.00
Level 2								
SL ^a	0.07***	0.01	0.10***	0.01	0.07***	0.01	0.10***	0.01
BFLPE								
L2-L1	-0.37***	0.01	-0.40***	0.01	-0.35***	0.01	-0.39***	0.01
Anmerkungen: ^a Leistung, ^b Geschlecht (1 weiblich, 2 männlich), ^c Migrationshintergrund (0 ohne, 1 mit), *** $p < .001$.								

5 Diskussion

Das vorliegende Kapitel fokussierte auf die Untersuchung von entwicklungsbedingten Unterschieden im BFLPE zwischen der vierten und achten Schulstufe. Hierfür wurden erstmalig Populationsdaten der österreichischen Bildungsstandardüberprüfung genutzt. Mithilfe von Mehrebenenanalysen konnten kontextuale Effekte (BFLPE) der aggregierten Klassenleistung (Level 2) und auch individuelle Leistungseffekte (Level 1) auf das akademische Selbstkonzept von Schülerinnen und Schülern untersucht werden. Die Ergebnisse geben einen detaillierten Einblick in altersbedingte Unterschiede in der Bedeutung sozialer Vergleichsprozesse von Schülerinnen und Schülern im österreichischen Schulsystem.

Gemäß dem BFLPE zeigten die Ergebnisse sowohl auf der vierten Schulstufe als auch auf der achten Schulstufe einen negativen Effekt der Klassenleistung (Level 2) auf das akademische Selbstkonzept der Schülerinnen und Schüler, unter Kontrolle ihrer individuellen Leistung. Das bedeutet, Schülerinnen und Schüler in einer leistungsstarken (leistungsschwachen) Klasse

entwickeln tendenziell ein niedrigeres (höheres) akademisches Selbstkonzept als jene in einer leistungsschwachen (leistungsstarken) Klasse. Im Zeitvergleich zeigte sich erwartungskonform, dass der BFLPE auf der achten Schulstufe stärker ausfiel als auf der vierten Schulstufe. Unter Berücksichtigung der Kontrollvariablen (Geschlecht, HISEI und Migrationshintergrund) blieb das Effektmuster bestehen, jedoch fielen die Effekte etwas geringer aus.

Im Vergleich zu bisherigen Studien von Fang et al. (2018) und Liou (2014) fallen die entwicklungsbedingten Unterschiede im BFLPE geringer aus. Dies könnte damit zusammenhängen, dass in der Metaanalyse von Fang et al. (2018) Volksschülerinnen und -schüler als homogene Gruppe betrachtet wurden, während in der vorliegenden Untersuchung ausschließlich Viertklässlerinnen/Viertklässler untersucht wurden. In der Metaanalyse wurden neben Viertklässlerinnen/Viertklässlern auch jüngere Volksschülerinnen und -schüler der ersten, zweiten und dritten Schulstufe berücksichtigt, wodurch die Stärke des BFLPE wohlmöglich aufgrund entwicklungsbedingter Unterschiede (Muenks et al., 2018) geringer ausfiel. Die Viertklässlerinnen/Viertklässler in Österreich entsprechen daher in ihrer Alterszusammensetzung eher der Kategorie Middle School aus der Metaanalyse. Daher replizieren die Unterschiede in unserer Studie weniger die Differenz zwischen Primary School und High School, sondern entsprechen vielmehr den Effekten aus dem Vergleich von Middle School und High School in der Metaanalyse. Auch im Vergleich zu der Studie von Liou (2014), welche Unterschiede im BFLPE zwischen der vierten und achten Schulstufe untersuchten, fallen die Effekte der vorliegenden Studie höher aus. Eine mögliche Erklärung dafür liegt in der jeweils einbezogenen Referenzgruppe. Während Liou (2014) soziale Vergleiche der Schülerinnen und Schüler mit der Schule analysierte, wurde in der vorliegenden Studie die Klasse als Referenzgruppe untersucht. Gemäß dem Lokalen Dominanzeffekt (Zell & Aliche, 2010) hat die dem Individuum nähere (lokale) Referenzgruppe einen größeren Einfluss bei sozialen Vergleichen als eine entferntere (allgemeinere) Referenzgruppe. Da die Klasse für das Individuum näher ist als die gesamte Schule, hat der soziale Vergleich mit den Klassenkameradinnen und Klassenkameraden einen stärkeren Einfluss auf das akademische Selbstkonzept der Schülerinnen und Schüler als der Vergleich mit der gesamten Schule. Vergleichbare Ergebnisse zeigten sich dagegen in der Studie von Marsh et al. (2015), welche die sozialen Vergleiche mit der durchschnittlichen Klassenleistung auf das akademische Selbstkonzept von Schülerinnen und Schülern der vierten und achten Schulstufe untersuchten. Diese BFLPE fallen im Schnitt ähnlich zu den vorliegenden Ergebnissen aus.

Zusammenfassend deuten die Ergebnisse der vorliegenden Studie darauf hin, dass die theoretischen Annahmen zur stärkeren Tendenz zu sozialen Vergleichen im Jugendalter verglichen mit dem Kindesalter zutreffen. Ein Grund hierfür kann in der Orientierung an Gleichaltrigen liegen, die im Jugendalter stärker ausgeprägt ist als in der Kindheit (Brown & Larson, 2009; Eccles & Midgley, 1989). Folglich erhöht sich mit zunehmendem Alter ebenfalls die Wahrscheinlichkeit, Gleichaltrige als Referenz für die Einschätzung der eigenen Leistungen heranzuziehen und sich mit ihnen zu vergleichen. Eine weitere Erklärung bieten entwicklungsbedingte Unterschiede in den kognitiven Fähigkeiten. Achtklässlerinnen/Achtklässler verfügen gegenüber Viertklässlerinnen/Viertklässlern bereits über höhere kognitive Fähigkeiten (Harter, 2003), was zu einer besseren Übereinstimmung zwischen der eigenen

Fähigkeitsbewertung und externalen Indikatoren führt (Guay et al., 2003). Das kann ebenfalls dazu führen, dass soziale Vergleichsprozesse als eine Quelle externer Informationen über die eigene Leistungsfähigkeit verstärkt genutzt werden können. Jugendliche orientieren sich in ihrer Fähigkeitsbewertung demnach mehr an den anderen als jüngere Schülerinnen und Schüler und weisen dadurch eine weniger egozentrische Bewertung ihrer schulischen Fähigkeiten auf (Guay et al., 2003).

Auf individueller Ebene (Level 1) zeigte sich, dass die individuelle Leistung der Schülerinnen und Schüler in beiden Schulstufen einen statistisch signifikanten Einfluss auf deren akademisches Selbstkonzept hat. Dies repliziert die in der Forschung bekannten Effekte der individuellen Leistung auf die Selbsteinschätzung der eigenen Fähigkeiten (Calsyn & Kenny, 1977; Guay et al., 2003; Marsh, 1990). Demzufolge nutzen Schülerinnen und Schüler Indikatoren aus dem Unterrichtsalltag, um ihre eigene Leistungsfähigkeit zu bewerten. Im Vergleich der vierten mit der achten Schulstufe zeigte sich, dass dieser Zusammenhang bei älteren Schülerinnen und Schülern stärker ist als bei den Jüngeren. Dies bestätigt unsere Erwartung, dass mit steigendem Alter und zunehmenden kognitiven Fähigkeiten die Selbsteinschätzungen akkurater werden und sich folglich stärker an dem tatsächlichen Leistungsniveau orientieren (Guay et al., 2003; Harter, 2003; Stockus & Zell, 2023). Eine weitere Erklärung für diesen Effekt kann darin liegen, dass Schülerinnen und Schüler auf der Sekundarstufe mehr Leistungsrückmeldungen erhalten als Volksschülerinnen und -schüler, wodurch sie eine bessere Datenbasis zur Selbsteinschätzung nutzen können (Parker, Dicke, Guo, Basarkod & Marsh, 2021).

Hinsichtlich der Kontrollvariablen (Geschlecht, HISEI und Migrationshintergrund) zeigten sich zwar direkte Effekte auf das akademische Selbstkonzept der Schülerinnen und Schüler. Diese hatten allerdings kaum Einfluss auf die Stärke des Zusammenhangs zwischen der individuellen Leistung und dem akademischen Selbstkonzept. Das bedeutet, dass diese drei Variablen keine konfundierenden Variablen des Zusammenhangs zwischen Leistung und akademischem Selbstkonzept sind. Die in bisherigen Studien bekannten Geschlechtsunterschiede hinsichtlich des mathematischen Selbstkonzepts (Goldman & Penner, 2016) konnten in der vorliegenden Studie repliziert werden. Auch unter Berücksichtigung der eigenen Leistungsfähigkeit schätzen sich Jungen in Mathematik kompetenter ein als Mädchen. Diese Unterschiede in der Selbsteinschätzung sind also nicht auf tatsächliche Kompetenzunterschiede zurückzuführen. Interessanterweise zeigte sich im Vergleich zwischen der vierten und der achten Schulstufe ein etwas geringerer Geschlechtsunterschied bei älteren Schülerinnen und Schülern. Dieses Ergebnis könnte auch als Hinweis auf eine genauere Selbstwahrnehmung des Leistungsniveaus bei älteren Schülerinnen und Schülern interpretiert werden. Die Beziehungen zwischen den beiden anderen Kontrollvariablen (HISEI und Migrationshintergrund) und dem akademischen Selbstkonzept sind aufgrund ihrer geringen Größe vernachlässigbar. Zudem unterscheiden sich diese Effekte nicht zwischen der vierten und der achten Schulstufe.

Zusammengefasst zeigte sich über alle Modelle hinweg, dass auf der achten Schulstufe sowohl der BFLPE stärker als auch die akademische Fähigkeitseinschätzung der Schülerinnen

und Schüler akkurater ausfiel als auf der vierten Schulstufe. Die stärkere Ausprägung des BFLPE auf der achten Schulstufe könnte somit als Indikator für die zunehmende Bedeutung sozialer Vergleichsprozesse interpretiert werden. Schülerinnen und Schüler in dieser Altersgruppe sind möglicherweise besser in der Lage, ihre Leistungen im Kontext ihrer Klassenkameradinnen und Klassenkameraden zu bewerten, was zu einer differenzierteren Selbstwahrnehmung führt. Diese Erkenntnisse legen nahe, dass die schulische Umgebung und die sozialen Vergleichsprozesse nicht nur das Selbstkonzept beeinflussen, sondern auch dessen Entwicklung über die Schuljahre hinweg.

5.1 Implikationen

Die Ergebnisse dieser Studie lassen einige Implikationen für bildungspolitische Entscheidungen und praktische Interventionen zu, um sowohl das akademische Selbstkonzept als auch die schulische Leistung der Schülerinnen und Schüler zu verbessern. Zunächst betonen sie die Relevanz des Klassenkontexts für die Bewertung des akademischen Selbstkonzepts von Schülerinnen und Schülern. Die komplexen Zusammenhänge zwischen Klassenumgebung, sozialen Vergleichen und dem akademischen Selbstkonzept wurden auch in der Studie von Fleischmann, Hübner, Marsh, Guo, Trautwein und Nagengast (2022) bestätigt. Diese Ergebnisse zeigten, dass der BFLPE in Mathematikklassen ausgeprägter ist als auf der Schulebene, was auf die saliente Rolle der Klassenumgebung hinweist. In Anbetracht dieser Erkenntnisse sollte die Bildungspolitik Interventionen priorisieren, die eine positive Klassenumgebung fördern und negative Auswirkungen sozialer Vergleiche auf das akademische Selbstkonzept der Schülerinnen und Schüler insbesondere im Mathematikunterricht minimieren.

Gezielte Maßnahmen zur Förderung von kooperativem Lernen und positiver Peer-Interaktion könnten im Mathematikunterricht implementiert werden, um ein unterstützendes Umfeld zu schaffen, das das akademische Selbstkonzept stärkt. Die Implementierung von Peer-Mentoring-Programmen basierend auf den Ansätzen der Peer-Education (Heyer, 2010) könnte darauf abzielen, ein unterstützendes und kooperatives Klima in der Klasse zu fördern und den Schülerinnen und Schülern dabei zu helfen, das Vertrauen in ihre eigenen Fähigkeiten zu stärken. Dies könnte unter anderem durch den Einsatz von klassenübergreifenden Gruppenübungen erreicht werden. Schülerinnen und Schüler vergleichen sich dadurch mit unterschiedlichen Peers und nicht nur mit denselben Personen aus der eigenen Klasse. Dies kann dazu führen, dass Schülerinnen und Schüler unterschiedliche Erfahrungen durch soziale Vergleiche machen und ein differenzierteres Selbstkonzept entwickeln (Drüke-Noe et al., 2022). Die Mentorinnen und Mentoren könnten daher auch ältere Schülerinnen und Schüler sein, die als positive Vorbilder fungieren und den jüngeren Schülerinnen und Schülern dabei helfen, ihre akademischen Ziele zu erreichen und ein positives Selbstkonzept zu entwickeln. Die Ergebnisse von Fleischmann, Hübner, Nagengast und Trautwein (2023) zeigten zudem, dass Detracking das akademische Selbstkonzept von Schülerinnen und Schülern mit durchschnittlichen und niedrigen schulischen Leistungen negativ beeinflussen kann. Heterogene Gruppen, in denen Schülerinnen und Schüler unterschiedlicher

Leistungsniveaus zusammenarbeiten, können jedoch auch positive Auswirkungen auf das akademische Selbstkonzept haben. Unterschiedliche leistungsstarke Schülerinnen und Schüler können durch die Zusammenarbeit voneinander profitieren, indem sie sich gegenseitig herausfordern und gleichzeitig unterstützen (Drüke-Noe et al., 2022). Das kooperative Zusammenarbeiten mit leistungsstärkeren Mitschülerinnen und Mitschülern kann dabei zu einer Verbesserung der Selbsteinschätzung führen. Schülerinnen und Schüler, die sich mit diesen Vorbildern vergleichen, könnten motiviert sein, ihre eigenen Anstrengungen zu steigern, um ähnliche Erfolge zu erzielen, was letztlich ihre eigene Leistung verbessert. Gleichzeitig kann durch das Zusammenarbeiten mit leistungsschwächeren Mitschülerinnen und Mitschülern das Selbstwertgefühl der Schülerinnen und Schüler gestärkt werden, indem sie ihre eigenen Fähigkeiten in einem besseren Licht sehen. Diese positive Selbstwahrnehmung kann dazu beitragen, ihre Leistung zu stabilisieren oder sogar zu verbessern, da das Vertrauen in die eigenen Fähigkeiten gestärkt wird (Crusius et al., 2022).

Darüber hinaus könnte die Implementierung von differenzierten Lernumgebungen eine wirksame Intervention sein, um den negativen Auswirkungen sozialer Vergleiche entgegenzuwirken. Differenzierte Lernumgebungen würden es den Schülerinnen und Schülern ermöglichen, in ihrem eigenen Tempo zu lernen und ihre individuellen Stärken und Schwächen zu berücksichtigen, anstatt sich ständig mit anderen zu vergleichen. Eine Metaanalyse von Hattie (2009) zeigte, dass differenzierte Lernumgebungen signifikante positive Effekte auf das Lernen und die Leistung der Schülerinnen und Schüler haben können. Durch die Schaffung von differenzierten Lernumgebungen können Schulen dazu beitragen, das akademische Selbstkonzept der Schülerinnen und Schüler zu stärken, indem sie ihnen die Möglichkeit geben, Erfolge zu erleben und sich als kompetent und fähig wahrzunehmen (Drüke-Noe et al., 2022). Dadurch kann eine Resilienz gegenüber negativen sozialen Vergleichsprozessen aufgebaut werden.

Des Weiteren legen die vorliegenden Befunde nahe, dass entwicklungsbedingte Unterschiede in der Bedeutung sozialer Vergleiche das akademische Selbstkonzept über verschiedene Schulstufen hinweg beeinflussen können. Bildungspolitische Entscheidungsträgerinnen und -träger sowie Pädagoginnen und Pädagogen sollten sich dieser Entwicklungsdynamiken bewusst sein und ihre Interventionen entsprechend gestalten. Zum Beispiel könnten Programme zur Förderung der Selbstwahrnehmung und Selbstreflexion entwickelt werden, die den Entwicklungsstand der Schülerinnen und Schüler berücksichtigen und individuelle Lernbedürfnisse unterstützen. Gerade nach dem Sekundarschulübertritt steigt in der Schule eine Feedbackkultur, die soziale Vergleiche befördert (Büchner & Koch, 2001). Dies ist gerade bei einer gesteigerten Peer-Orientierung ein Risikofaktor für eine negative Selbstkonzept- und Leistungsentwicklung.

5.2 Limitationen

Bei der Interpretation der Ergebnisse sind einige Einschränkungen zu berücksichtigen. Ein Teil unserer theoretischen Argumentation bezog sich auf entwicklungsbedingte Ver-

änderungen im Jugendalter. Obwohl Trendstudien wie diese gute Hinweise liefern können, sind für eine sichere Überprüfung dieser Mechanismen Längsschnittstudien notwendig (Becker & Neumann, 2018).

Eine weitere Herausforderung besteht darin, dass auf der vierten und achten Schulstufe zwar vergleichbare, aber dennoch nicht identische Tests der Mathematikkompetenz verwendet wurden. Dies liegt in der Natur der Sache, da Schülerinnen und Schüler der vierten Schulstufe nicht zu den gleichen Inhalten der achten Schulstufe getestet werden können. Andernfalls wären die Tests nicht mehr inhaltsvalide (Moosbrugger & Kelava, 2008). Aus messtheoretischer Sicht wäre es jedoch notwendig, die exakt gleichen Tests auf der vierten und achten Schulstufe zu verwenden, um altersbedingte Unterschiede in den Effekten auf die akademischen Selbstkonzepte zu untersuchen.

Die letzte Einschränkung bezieht sich auf die sparsame Operationalisierung des akademischen Selbstkonzepts. Bei der Reanalyse des Datensatzes standen nur vier inhaltsvalide Items zur Erfassung des akademischen Selbstkonzepts zur Verfügung. Daher kann das akademische Selbstkonzept in der Mathematikdomäne nicht in seiner vollen Breite abgedeckt werden. Allerdings ist die verfügbare Itembasis gut an die vorliegende Literatur zum akademischen Selbstkonzept anschließbar (Liou, 2014; Marsh et al., 2015).

5.3 Zukünftige Forschungsfragen

Diese Studie eröffnet weitere Forschungsfragen im Bereich sozialer Vergleichsprozesse und des BFLPE. Erstens wäre es interessant, zusätzliche Moderatoren des BFLPE zu untersuchen, wie von Seaton et al. (2010) angeregt. Basierend auf den vorliegenden Ergebnissen könnten zukünftige Forschungen personenspezifische Effekte und deren Einfluss auf den BFLPE genauer analysieren. Wie aus bisherigen Studien bereits bekannt ist, gibt es geschlechtsspezifische Unterschiede in der Präferenz der Referenzgruppen (Guimond & Chatard, 2014). Dies kann zu Unterschieden in der Größe des BFLPE zwischen Buben und Mädchen führen (Plieninger & Dickhäuser, 2015). Mithilfe personenzentrierter Methoden könnten diese intraindividuellen Unterschiede im BFLPE sichtbar gemacht werden (Bogat, von Eye, & Bergman, 2016). Personenzentrierte Forschungen sind von entscheidender Bedeutung, um optimale Kontext-Personen-Konstellationen zu identifizieren und daraus detaillierte praktische Implikationen abzuleiten.

Zweitens könnten künftige Forschungsarbeiten auch die Schulebene zusätzlich einbeziehen, um nicht nur den Einfluss des BFLPE zu untersuchen, sondern auch potenzielle Assimilationseffekte genauer zu beleuchten. Insbesondere auf Schulebene spielen Assimilationseffekte, wie der „basking-in-reflected-glory effect“ (BIRGE; Marsh, Kong & Hau, 2000), eine wichtige Rolle (von Keyserlingk, Becker & Jansen, 2019).

Die Verwendung von standardisierten Tests bietet zwar den Vorteil hinsichtlich der Vergleichbarkeit und Objektivität, allerdings erfassen diese nicht vollständig die subjektiven

Wahrnehmungen und sozialen Vergleichsprozesse der Schülerinnen und Schüler. Das bedeutet, wie Schülerinnen und Schüler ihre eigene Leistung im Vergleich zu ihren Klassenkameradinnen und Klassenkameraden wahrnehmen. Schülerinnen und Schüler nehmen nicht nur die durchschnittliche Leistung ihrer Klasse wahr, sondern auch die Verteilung der Leistungen ihrer Mitschülerinnen und -schüler. Diese Wahrnehmung kann zusätzliche Einblicke in die Dynamiken des BFLPE bieten und benötigt weitere Forschungsarbeiten. Darüber hinaus werden in sozialen Vergleichsprozessen oftmals die Noten der Klassenkameradinnen und Klassenkameraden herangezogen. Noten, die durch Lehrkräfte vergeben und im Kontext von Klassenarbeiten oder mündlichen Beiträgen ermittelt werden, bieten ein direkteres Feedback an die Schülerinnen und Schüler. Dieses Feedback ist häufig stark subjektiv und kann von der Beziehung zwischen Lehrkraft und Schülerin/Schüler sowie von der individuellen Interpretation der Leistung beeinflusst sein. Noten spiegeln somit nicht nur die tatsächliche Leistung wider, sondern auch den sozialen Kontext, in dem diese Leistung erbracht wurde. Dieser Prozess wird durch das Phänomen des „grading-on-a-curve“ (Marsh, Kuyper, Morin, Parker & Seaton, 2014, S. 51) verstärkt, bei dem die Leistung der Schülerinnen und Schüler relativ zu der Leistung der Mitschülerinnen und -schüler bewertet wird. Dies verstärkt den sozialen Vergleich, da Schülerinnen und Schüler nicht nur ihre absoluten Leistungen betrachten, sondern auch ihre relative Position innerhalb der Klasse. Zukünftige Untersuchungen könnten sich möglichen Positionseffekten gezielter widmen, um die Rolle individueller Positionen innerhalb des Klassenverbunds genauer zu erforschen.

Zudem sind die Mitschülerinnen und -schüler in der Klasse nicht die einzigen relevanten Bezugsgrößen zur Einschätzung der eigenen Leistungsfähigkeit. Gute Freunde und Freundinnen oder Cliquen innerhalb und außerhalb der Klasse sind gerade für soziale Vergleiche im Jugendalter bedeutsam. Erste Studien deuten bereits darauf hin, dass diese Peergruppen ebenfalls relevante Referenzgruppen darstellen (Altermatt & Pomerantz, 2005; Jansen, Boda & Lorenz, 2022; Auer, Reindl & Gniewosz, 2022).

5.4 Schlussfolgerung

Obwohl die berichteten Unterschiede relativ geringe Effekte aufweisen, ermöglicht der verwendete Datensatz dennoch interessante Einblicke in altersbedingte Veränderungen sozialer Vergleichsprozesse von Schülerinnen und Schülern bezüglich der akademischen Selbsteinschätzung im Fach Mathematik. Mit der Zeit scheint die Klasse als soziale Referenzgruppe für diese Einschätzungen an Bedeutung zu gewinnen, da der BFLPE auf der achten Schulstufe stärker ausfällt als auf der vierten Schulstufe. Gleichzeitig wurden die Selbsteinschätzungen der Schülerinnen und Schüler auf der achten Schulstufe akkurater. Dieser Zuwachs an Genauigkeit bei den Selbsteinschätzungen könnte möglicherweise auf eine stärkere Orientierung an den sozialen Referenzgruppen innerhalb der Klasse zurückzuführen sein.

Literatur

- Altermatt, E. R. & Pomerantz, E. M. (2005). The implications of having high-achieving versus low-achieving friends: A longitudinal analysis. *Social Development*, 14 (1), 61–81. doi:10.1111/j.1467-9507.2005.00291.x
- Arens, A. K., Schmidt, I. & Preckel, F. (2019). Longitudinal relations among self-concept, intrinsic value, and attainment value across secondary school years in three academic domains. *Journal of Educational Psychology*, 111 (4), 663–684. doi:10.25656/01:17475
- Asparouhov, T. & Muthén, B. (2010). *Weighted least squares estimation with missing data. Mplus technical appendix, 2010 (1–10)*, 5. Verfügbar unter: <http://www.statmodel.com/download/GstrucMissingRevision.pdf>
- Auer, T., Reindl, M. & Gniewosz, B. (2022, 20.–23. September). *Big-Fish-Little-Pond Effekt: soziale Vergleichsprozesse mit Clique und Klasse und deren Einfluss auf die Entwicklung des akademischen Selbstkonzepts in Mathematik innerhalb eines Schuljahres. Eine Mehr-Ebenen-Analyse* [Konferenzbeitrag]. Kongress der Österreichischen Gesellschaft für Forschung und Entwicklung im Bildungswesen (ÖFEB), Graz, Österreich.
- Becker, M. & Neumann, M. (2018). Longitudinal big-fish-little-pond effects on academic self-concept development during the transition from elementary to secondary schooling. *Journal of Educational Psychology*, 110 (6), 882–897. doi:10.1037/edu0000233
- BIFIE (2017). *Almanach zur Standardüberprüfung 2013 Mathematik, 4. Schulstufe. Schüler-ebene, M413I. Forschungsdatenbibliothek (FDB). Nicht-imputierter Datensatz, v2.0-E*. Verfügbar unter: <https://www.bifie.at/fdb/M413>
- BIFIE (2018). *Almanach zur Standardüberprüfung 2017 Mathematik, 8. Schulstufe. Schüler-ebene, M817I. Forschungsdatenbibliothek (FDB). Nicht-imputierter Datensatz, v1.0-E*. Verfügbar unter: <https://www.bifie.at/fdb/m817/>
- Bogat, G. A., von Eye, A. & Bergman, L. R. (2016). Person-oriented approaches. In D. Cicchetti & D. J. Cohen (Hrsg.), *Developmental Psychopathology* (S. 701–749). doi:10.1002/9781119125556.devpsy118
- Brown, B. B. & Larson, J. (2009). Peer relationships in adolescence. In R. M. Lerner & L. Steinberg (Hrsg.), *Handbook of adolescent psychology: Contextual influences on adolescent development* (S. 74–104). New Jersey: John Wiley & Sons, Inc. doi:10.1002/9780470479193.adlpsy002004
- Büchner, P. & Koch, K. (2001). Von der Grundschule in die Sekundarschule. Der Übergang aus Kinder- und Elternsicht. Opladen: Leske u. Budrich. doi: 10.25656/01:27512
- Calsyn, R. & Kenny, D. (1977). Self-concept of ability and perceived evaluations by others: Cause or effect of academic achievement? *Journal of Educational Psychology*, 69 (2), 136–145. doi:10.1037/0022-0663.69.2.136
- Castro, S. L. (2002). Data analytic methods for the analysis of multilevel questions: A comparison of intraclass correlation coefficients hierarchical linear modeling, within- and between-analysis, and random group resampling. *The Leadership Quarterly*, 13 (1), 69–93. doi:10.1016/S1048-9843(01)00105-9
- Crusius, J., Corcoran, K., & Mussweiler, T. (2022). Social Comparison. In D. Chadee (Hrsg), *Theories in Social Psychology* (S. 165–187). New Jersey: John Wiley & Sons, Inc. doi:10.1002/9781394266616.ch7

- Dai, D. Y. & Rinn, A. N. (2008). The big-fish-little-pond effect: What do we know and where do we go from here? *Educational Psychology Review*, 20, 283–317. doi:10.1007/s10648-008-9071-x
- Dijkstra, P., Kuyper, H., Van der Werf, G., Buunk, A. P. & van der Zee, Y. G. (2008). Social comparison in the classroom: A review. *Review of educational research*, 78 (4), 828–879. doi:10.3102/0034654308321210
- Drücke-Noe, C., Gniewosz, B. & Paasch, D. (2022). Mathematisches Selbstkonzept und Schülerleistungen – Zusammenhänge für den Unterricht nutzbar machen. In A. C. George, S. Götz, M. Illetschko & E. Süss-Stepancik (Hrsg.), *Empirische Befunde zu Kompetenzen im Mathematikunterricht der Sekundarstufe 1 und Folgerungen für die Praxis – Ergänzende Analysen zu den Bildungsstandardüberprüfungen* (S. 212–232). Münster: Waxmann.
- Eccles, J. S. & Midgley, C. (1989). Stage/environment fit: Developmentally appropriate classrooms for early adolescents. In R. Ames & C. Ames (Hrsg.), *Research on motivation in education* (Vol. 3, S. 139–186). New York: Academic Press.
- Eccles, J. S. & Roeser, R. W. (2009). Schools, academic motivation, and stage-environment fit. In R. M. Lerner & L. Steinberg (Hrsg.), *Handbook of adolescent psychology: Individual bases of adolescent development* (S. 404–434). New Jersey: John Wiley & Sons, Inc. doi:10.1002/9780470479193.adlpsy001013
- Enders, C. K. & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12 (2), 121–138. doi:10.1037/1082-989X.12.2.121
- Fang, J., Huang, X., Zhang, M., Huang, F., Li, Z. & Yuan, Q. (2018). The big-fish-little-pond effect on academic self-concept: A meta-analysis. *Frontiers in Psychology*, 9, 1569. doi:10.3389/fpsyg.2018.01569
- Festinger, L. (1954). A theory of social communication processes. *Human Relations*, 7 (2), 117–140. doi:10.1177/001872675400700202
- Fleischmann, M., Hübner, N., Marsh, H. W., Guo, J., Trautwein, U. & Nagengast, B. (2022). Which class matters? Juxtaposing multiple class environments as frames-of-reference for academic self-concept formation. *Journal of Educational Psychology*, 114 (1), 127–143. doi:10.1037/edu0000491
- Fleischmann, M., Hübner, N., Nagengast, B. & Trautwein, U. (2023). The dark side of detracking: Mixed-ability classrooms negatively affect the academic self-concept of students with low academic achievement. *Learning and Instruction*, 86, 101753. doi:10.1016/j.learninstruc.2023.101753
- Fornell, C. & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of marketing research*, 18 (1), 39–50. doi:10.1177/002224378101800104
- Freunberger, R., Robitzsch, A., & Pham, G. (2014). *Hintergrundvariablen und spezielle Analysen in der BIST-Ü-M4 2013. Technischer Bericht*. Salzburg: BIFIE.
- Geiser, C. (2011). *Datenanalyse mit Mplus. Eine anwendungsorientierte Einführung* (2. Aufl.) Wiesbaden: VS. doi:10.1007/978-3-531-93192-0_1

- George, A. C., Schreiner, C., Wiesner, C., Pointinger, M. & Pacher, K. (2019). *Fünf Jahre flächendeckende Bildungsstandardüberprüfungen in Österreich: Vertiefende Analysen zum Zyklus 2012 bis 2016*. Münster: Waxmann.
- Goldman, A. D. & Penner, A. M. (2016). Exploring international gender differences in mathematics self-concept. *International Journal of Adolescence and Youth*, 21 (4), 403–418. doi:10.1080/02673843.2013.847850
- Guay, F., Larose, S. & Boivin, M. (2004). Academic self-concept and educational attainment level: A ten-year longitudinal study. *Self and Identity*, 3 (1), 53–68. doi: 10.1080/13576500342000040
- Guay, F., Marsh, H. W. & Boivin, M. (2003). Academic self-concept and academic achievement: Developmental perspectives on their causal ordering. *Journal of Educational Psychology*, 95 (1), 124–136. doi:10.1037/0022-0663.95.1.124
- Guimond, S. & Chatard, A. (2014). Basic principles of social comparison: Does gender matter? In F. X. Gibbons & Z. Križan (Hrsg.), *Communal Functions of Social Comparison* (S. 205–229). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139035583.013
- Guo, J., Marsh, H. W., Parker, P. D., Morin, A. J. & Dicke, T. (2017). Extending expectancy-value theory predictions of achievement and aspirations in science: Dimensional comparison processes and expectancy-by-value interactions. *Learning and Instruction*, 49, 81–91. doi:10.1016/j.learninstruc.2016.12.007
- Harter, S. (2003). The development of self-representations during childhood and adolescence. In M. R. Leary & J. P. Tangney (Hrsg.), *Handbook of Self and Identity* (S. 610–642). New York: Guilford Press.
- Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.
- Havighurst, R. J. (1972). *Developmental tasks and education* (3. Aufl.). New York: David McKay.
- Heyer, R. (2010). Peer-Education – Ziele, Möglichkeiten und Grenzen. In M. Harring, O. Böhm-Kasper, C. Rohlf & C. Palentien (Hrsg.), *Freundschaften, Cliquen und Jugendkulturen. Peers als Bildungs- und Sozialisationsinstanzen* (S. 407–421). Wiesbaden: Springer VS.
- Hoferichter, F., Lätsch, A., Lazarides, R. & Raufelder, D. (2018). The big-fish-little-pond effect on the four facets of academic self-concept. *Frontiers in Psychology*, 9, 1247. doi: 10.3389/fpsyg.2018.01247
- Hu, L.-T. & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3 (4), 424–453. doi:10.1037/1082-989X.3.4.424
- Hu, L.-T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling*, 6 (1), 1–55. doi:10.1080/10705519909540118
- Huguet, P., Dumas, F., Marsh, H., Régner, I., Wheeler, L., Suls, J., Seaton, M. & Nezlek, J. (2009). Clarifying the role of social comparison in the big-fish-little-pond effect (BFLPE): An integrative study. *Journal of personality and social psychology*, 97 (1), 156–170. doi:10.1037/a0015558

- Hummrich, M. & Kramer, R. T. (2017). Was ist schulische Sozialisation? In *Schulische Sozialisation. Basiswissen Sozialisation* (S. 7–27). Wiesbaden: Springer VS. doi: 10.1007/978-3-531-19057-0_2
- Hox, J., Moerbeek, M. & van de Schoot, R. (2017). *Multilevel Analysis: Techniques and Applications* (3. Aufl.). New York: Routledge. doi:10.4324/9781315650982
- Jahr, D. (2022). *Die Politik der Schulklasse. Dokumentarische Videoanalysen unterrichtlicher Praktiken zwischen Integration und Destruktion*. Wiesbaden: Springer VS. doi: 10.1007/978-3-658-38222-3
- Jaiswal, S. K. & Choudhuri, R. (2017). Academic self concept and academic achievement of secondary school students. *American Journal of Educational Research*, 5 (10), 1108–1113. doi:10.12691/education-5-10-13
- Jansen, M., Boda, Z. & Lorenz, G. (2022). Social comparison effects on academic self-concepts – Which peers matter most? *Developmental Psychology*, 58 (8), 1541–1556. doi:10.1037/dev0001368
- Kavanagh, L. (2020). Academic self-concept formation: testing the internal/external frame of reference model, big-fish-little-pond model, and an integrated model at the end of primary school. *European journal of Psychology of Education*, 35 (1), 93–109. doi: 10.1007/s10212-019-00416-w
- Kiefer, T., Kuhn, J.-T. & Fellingner, R. (2016). Testdesign. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (S. 149–184). Wien: facultas. doi:10.31244/9783830991489
- Leonhart, R., Wirtz, M. & Bengel, J. (2008). Measuring effect sizes using manifest versus latent variables: Consequences and implications for research. *International Journal of Rehabilitation Research*, 31 (3), 207–216. doi:10.1097/MRR.0b013e3282fb7872
- Little, T. D. (2013). *Longitudinal structural equation modeling*. New York: Guilford Press.
- Liou, P. Y. (2014). Investigation of the big-fish-little-pond effect on students' self-concept of learning mathematics and science in Taiwan: Results from TIMSS 2011. *The Asia-Pacific Education Researcher*, 23 (3), 769–778. doi:10.1007/s40299-013-0152-3
- Lüftenegger, M., Schober, B., Van de Schoot, R., Wagner, P., Finsterwald, M. & Spiel, C. (2012). Lifelong learning as a goal – Do autonomy and self-regulation in school result in well prepared pupils?. *Learning and Instruction*, 22 (1), 27–36. doi:10.1016/j.learninstruc.2011.06.001
- Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal*, 23 (1), 129–149.
- Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, 79 (3), 280–295. doi:10.1037/0022-0663.79.3.280
- Marsh, H. W. (1990). The causal ordering of academic self-concept and academic achievement: A multiwave, longitudinal panel analysis. *Journal of Educational Psychology*, 82, 646–656. doi:10.1037/0022-0663.82.4.646
- Marsh, H. W., Abduljabbar, A. S., Morin, A. J., Parker, P., Abdelfattah, F., Nagengast, B. & Abu-Hilal, M. M. (2015). The big-fish-little-pond effect: Generalizability of social comparison processes over two age cohorts from Western, Asian, and Middle Eastern Islamic countries. *Journal of Educational Psychology*, 107 (1), 258–271. doi:10.1037/a0037485

- Marsh, H. W., Kong, C. K. & Hau, K. T. (2000). Longitudinal multilevel models of the big-fish-little-pond effect on academic self-concept: Counterbalancing contrast and reflected-glory effects in Hong Kong schools. *Journal of Personality and Social Psychology*, 78 (2), 337–349. doi:10.1037/0022-3514.78.2.337
- Marsh, H. W., Kuyper, H., Morin, A. J., Parker, P. D. & Seaton, M. (2014). Big-fish-little-pond social comparison and local dominance effects: Integrating new statistical models, methodology, design, theory, and substantive implications. *Learning and Instruction*, 33, 50–66. doi:10.1016/j.learninstruc.2014.04.002
- Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S. & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, 47 (2), 106–124. doi:10.1080/00461520.2012.670488
- Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B. & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, 44 (6), 764–802. doi:10.1080/00273170903333665
- Marsh, H. W. & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology*, 81 (1), 59–77. doi:10.1348/000709910X503501
- Moosbrugger, H. & Kelava, A. (2008). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 7–26). Berlin: Springer. doi:10.1007/978-3-540-71635-8_2
- Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research*, 79 (3), 1129–1167. doi:10.3102/0034654309337522
- Muenks, K., Wigfield, A. & Eccles, J. S. (2018). I can do this! The development and calibration of children's expectations for success and competence beliefs. *Developmental Review*, 48, 24–39. doi:10.1016/j.dr.2018.04.001
- Mussweiler, T. & Rüter, K. (2003). What friends are for! The use of routine standards in social comparison. *Journal of Personality and Social Psychology*, 85 (3), 467–481. doi:10.1037/0022-3514.85.3.467
- Muthén, L. K. & Muthén, B. O. (2017). *Mplus. Statistical analysis with latent variables. User's guide*. Los Angeles: Muthén and Muthén.
- Parker, P., Dicke, T., Guo, J., Basarkod, G. & Marsh, H. (2021). Ability stratification predicts the size of the big-fish-little-pond effect. *Educational Researcher*, 50 (6), 334–344. doi:10.3102/0013189X20986176
- Plieninger, H. & Dickhäuser, O. (2015). The female fish is more responsive: Gender moderates the BFLPE in the domain of science. *Educational Psychology*, 35 (2), 213–227. doi:10.1080/01443410.2013.814197
- Robitzsch, A., Pham, G. & Yanagida, T. (2016). Fehlende Daten und Plausible Values. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R: methodische Grundlagen der österreichischen Bildungsstandard-Überprüfung* (S. 259–293) Wien: facultas.

- Rubin, K. H., Bukowski, W. M. & Bowker, J. C. (2015). Children in peer groups. In M. H. Bornstein, T. Leventhal & R. M. Lerner (Hrsg.), *Handbook of child psychology and developmental science: Ecological settings and processes* (S. 175–222). New Jersey: John Wiley & Sons, Inc.
- Schreiner C., Breit S., Pointinger M., Pacher K., Neubacher M. & Wiesner C. (2018). *Standardüberprüfung 2017. Mathematik, 8. Schulstufe. Bundesergebnisbericht*. Salzburg: Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens. Verfügbar unter: <https://www.iqs.gv.at/downloads/archiv-des-bifie/bildungsstandardueberpruefungen/ergebnisberichte>
- Seaton, M., Marsh, H. W. & Craven, R. G. (2010). Big-fish-little-pond effect: Generalizability and moderation—Two sides of the same coin. *American Educational Research Journal*, 47 (2), 390–433. doi:10.3102/0002831209350493
- Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-Concept: Validation of construct interpretations. *Review of Educational Research*, 46 (3), 407–441. doi:10.3102/00346543046003407
- Stockus, C. A. & Zell, E. (2023). Moderators of the big-fish-little-pond effect in educational settings: A scoping review. *Social and Personality Psychology Compass*, e12860. doi: 10.1111/spc3.12860
- Tausendpfund, M. (2020). Mehrebenenanalyse. In M. Tausendpfund (Hrsg.), *Fortgeschrittene Analyseverfahren in den Sozialwissenschaften: Ein Überblick* (S. 119–164). Wiesbaden: Springer VS. doi:10.1007/978-3-658-30237-5
- Televantou, I., Marsh, H. W., Dicke, T. & Nicolaides, C. (2021). Phantom and big-fish-little-pond-effects on academic self-concept and academic achievement: Evidence from English early primary schools. *Learning and Instruction*, 71, 101399. doi:10.1016/j.learninstruc.2020.101399
- Thijs, J., Verkuyten, M. & Helmond, P. (2010). A further examination of the big-fish-little-pond effect: Perceived position in class, class size, and gender comparisons. *Sociology of Education*, 83 (4), 333–345. doi:10.1177/0038040710383521
- Trendtel, M., Pham, G. & Yanagida, T. (2016). Skalierung und Linking. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (S. 185–224). Wien: facultas.
- Vogl, K., Schmidt, I. & Preckel, F. (2018). The role of academic ability indicators in big-fish-little-pond effect research: A comparison study. *The Journal of Educational Research*, 111 (4), 429–438. doi:10.1080/00220671.2017.1291485
- von Keyserlingk, L., Becker, M. & Jansen, M. (2019). Academic self-concept during the transition to upper secondary school. *Contemporary Educational Psychology*, 56, 152–160. doi:10.1016/j.cedpsych.2019.01.001
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54 (3), 427–450. doi:10.1007/BF02294627
- Wiesner, C., Schreiner, C. & Breit, S. (2020). Modelle und Orientierungsrahmen für Kompetenz- und Leistungsmessungen Grundlage für Konzeption sowie die Beurteilung ihrer Qualität. In U. Greiner, F. Hofmann, C. Schreiner & C. Wiesner (Hrsg.), *Bildungsstandards. Kompetenzorientierung, Aufgabenkultur und Qualitätsentwicklung im Schulsystem* (S. 124–143). Münster: Waxmann. doi:10.31244/9783830991489

- Wu, H., Guo, Y., Yang, Y., Zhao, L. & Guo, C. (2021). A meta-analysis of the longitudinal relationship between academic self-concept and academic achievement. *Educational Psychology Review*, 33 (4), 1749–1778. doi:10.1007/s10648-021-09600-1
- Zell, E. & Alicke, M. D. (2010). The local dominance effect in self-evaluation: Evidence and explanations. *Personality and Social Psychology Review*, 14 (4), 368–384. doi: 10.1177/1088868310366144

Binnendifferenzierung und Individualisierung im Volksschulleseunterricht: Österreich und ausgewählte EU-Vergleichsländer im 15-Jahres-Trend

1 Einleitung

„Differenzierung“ und „Individualisierung“ sind Schlagworte, die seit einigen Jahrzehnten im Zusammenhang mit gelingendem Kompetenzerwerb und passgenauer Förderung von Schülerinnen und Schülern nicht mehr wegzudenken sind.¹ Um insbesondere lesestarke und leeschwache Schülerinnen und Schüler angemessen fördern zu können und den ungleichen Voraussetzungen in Schülergruppen gerecht zu werden, wird oftmals Individualisierung im Unterricht gefordert. Ziel ist es demnach, eine Passung zwischen den bei Schülerinnen und Schülern vorhandenen Lernvoraussetzungen und den jeweiligen Lernangeboten herzustellen (Altrichter, Trautmann, Wischer, Sommerauer & Doppler, 2009, S. 343). Wenn- gleich auch in der Sekundarstufe nicht von homogenen Klassen ausgegangen werden kann (Altrichter et al., 2009, S. 342), ist das Thema heterogener Klassen auf der Primarstufe besonders relevant, da hier noch keine Selektion durch unterschiedliche Schulformen erfolgt ist (Hess & Lipowsky, 2017).

Bei „Differenzierung“ handelt es sich um eine merkmalsbezogene Gruppierung der Schülerinnen und Schüler. Dabei kann eine Zuordnung zu einer Gruppe nach Alter, Leistung, Interesse, Geschlecht etc. erfolgen. Die Ausrichtung des Lernangebots richtet sich nach der einzelnen Gruppe, nicht nach dem Individuum. Es wird häufig in äußere (Schularten, Schulzweige, Klassen, Kursgruppen) und innere Differenzierung bzw. Binnendifferenzierung (Stoffumfang, Komplexitätsgrad, Ausmaß der Selbstständigkeit etc.) unterschieden (Bönsch, 1995, 2009; Klafki, 1994, S. 188 ff.). „Individualisierung“ hat im Gegensatz dazu das einzelne Kind im Blick, erhebt dessen Ausgangslage und stellt entsprechende, individuelle Lernangebote zur Verfügung (Bohl, Batzel & Richey, 2012; von der Groeben & Kaiser, 2011) (s. Abschnitt 2).

1 Zu den Gründen für das gesteigerte Interesse an Individualisierung und Binnendifferenzierung s. Altrichter et al. (2009, S. 346 f.).

Binnendifferenzierung und Individualisierung sind an verschiedenen Stellen im österreichischen Lehrplan der Volksschule 2012 Thema. Es heißt dort beispielsweise: „Diesen Unterschiedlichkeiten der Kinder soll die Lehrerin bzw. der Lehrer durch differenzierende und individualisierende Maßnahmen entsprechen“ (BGBl., I Nr. 303/2012). Im Teil „Allgemeine didaktische Grundsätze für die Grundschule“ ist Abschnitt 7 dem „Individualisieren, Differenzieren und Fördern“ gewidmet. Darin werden nicht zuletzt unterschiedliche Formen der inneren Differenzierung behandelt (BGBl. I Nr. 303/2012).

Die Forschungslage zur Wirksamkeit individualisierten und differenzierten Unterrichts ist indes laut Bohl et al. (2012, S. 56) aufgrund der unterschiedlichen Definitionen und Sichtweisen von Differenzierung und Individualisierung, der Vielfältigkeit der methodischen Ansätze sowie der geringen Anzahl an Längsschnittstudien alles andere als klar.

Wenn man davon ausgeht, dass Binnendifferenzierung und Individualisierung vielversprechende Ansätze zur passgenauen Förderung von Kindern im Leseunterricht sein können, lohnt sich ein Blick auf den Stand und generelle Trends dazu in Österreich im Zeitverlauf. Vor diesem Kontext verfolgt der vorliegende Beitrag das Ziel, vorhandene Daten aus einer nationalen und vier internationalen Erhebungen der letzten 15 Jahre zum Leseunterricht in Volksschulen zusammenzubringen. Aufgrund der uneindeutigen Forschungslage zur Wirksamkeit von Binnendifferenzierung und Individualisierung betrachten wir zudem einen möglichen Zusammenhang von differenziertem bzw. individualisiertem Leseunterricht und der Lesekompetenz.

Daraus ergeben sich die folgenden Forschungsfragen. Sie werden in Abschnitt 2.4 genauer erläutert:

Unterscheidet sich Österreich in Bezug auf Binnendifferenzierung und Individualisierung von anderen Ländern? Hat sich der Einsatz von Binnendifferenzierung und Individualisierung im Laufe der Zeit verändert? Gibt es einen Zusammenhang zwischen jeweils dem Unterricht mit der gesamten Klasse, Binnendifferenzierung bzw. Individualisierung und Lesekompetenz?

Zur Beantwortung dieser Forschungsfragen wird eine Kurzskaala zu Binnendifferenzierung und Individualisierung verwendet, welche im Laufe von 15 Jahren vier Mal bei dem international vergleichenden Lesetest PIRLS (*Progress in International Reading Literacy Study*)² eingesetzt wurde. Für ergänzende Analysen zur Individualisierung werden darüber hinaus einzelne Aussagen, die nur zu bestimmten Zeitpunkten erfasst wurden, verwendet. Schließlich werden Ergebnisse für Österreich im Überblick präsentiert. Dafür werden zusätzlich zu den internationalen Daten die Baseline-Daten der Bildungsstandardüberprüfung in Österreich aus dem Jahr 2010 hinzugezogen (BIFIE [IQS], 2021; Breit, Bruneforth & Schreiner, 2016).

2 In Deutschland ist die Studie unter dem Akronym IGLU (Internationale Grundschul-Lese-Untersuchung) bekannt.

2 Theoretischer Hintergrund

„Differenzierung“ und „Individualisierung“ sind Konzepte, die sich mit dem Begriff des sogenannten *adaptiven Unterrichts* (Fladung, 2022; Helmke, 2013, S. 34; Hess & Lipowsky, 2017, S. 25) überschneiden, bei dem es darum geht, Unterricht an die individuellen Differenzen in den Lernvoraussetzungen der Schülerinnen und Schüler anzupassen.

2.1 Differenzierung

Unter „Differenzierung“ (hier verstanden als „Binnendifferenzierung“) wird in diesem Beitrag die Gruppierung von Schülerinnen und Schülern innerhalb der Klasse nach bestimmten Merkmalen wie Kompetenzniveau oder Interesse verstanden. Dabei sind nicht unbedingt die individuellen Lernvoraussetzungen eines Kindes ausschlaggebend – diese werden im Unterschied zur Individualisierung für die Bildung von Gruppen nicht eigens erfasst und auch das Lernangebot wird nicht individuell angepasst. Vielmehr geht es bei der Binnendifferenzierung darum, mit geringem Aufwand Gruppen aufgrund bestimmter Merkmale (wie Leistung) innerhalb der Klasse zu bilden. Das erhöht im Vergleich zum Unterricht mit der gesamten Klasse ohne Differenzierung die Chance, eine Passung zwischen Lernangebot und Interessen/Lernvoraussetzungen/Kompetenzniveaus etc. der Schülerinnen und Schüler herzustellen (Bohl et al., 2012, S. 47). Denn der Unterricht mit der gesamten Klasse orientiert sich häufig am durchschnittlichen Kompetenzniveau der Klasse (Bönsch, 2012, S. 10). Dabei bearbeitet die Klasse meist „die gleiche Aufgabe auf die gleiche Art“ (Fladung, 2022, S. 85). Im Gegensatz dazu erfolgt bei der Binnendifferenzierung eine Bildung von Gruppen, die nach unterschiedlichen Gesichtspunkten erfolgen kann. Häufig werden entweder „homogene“ Gruppen mit ähnlichem Fähigkeitsniveau oder „heterogene“ Gruppen mit unterschiedlichem Fähigkeitsniveau gebildet. Diese Gruppen bearbeiten dann z. B. je nach Kompetenzniveau unterschiedliche Aufgaben oder sie lösen die gleiche Aufgabe beispielsweise mit unterschiedlichen Methoden, unterschiedlichen Zeitvorgaben oder einem unterschiedlichen Maß an Hilfestellung (Fladung, 2022, S. 85 f.).

2.2 Individualisierung

Unter „Individualisierung“ werden in diesem Beitrag mit Bohl et al. (2012, S. 44 f.) zwei Varianten verstanden: Erstens die lehrerseitige Anpassung des Lernangebots an individuelle Voraussetzungen und Interessen von Schülerinnen und Schülern und zweitens Lernangebote, aus denen Schülerinnen und Schüler nach ihren Voraussetzungen, Kompetenzniveaus und/oder Interessen auswählen können.

Zunächst zu Variante 1. Diese kann den Autorinnen und Autoren zufolge – ähnlich auch Helmke (2013, S. 35) – grob in fünf Bereiche gefasst werden:

1. Die Individuelle Anpassung von Zielen, Inhalten sowie Kompetenzen,
2. die Individuelle Diagnostik der jeweiligen Voraussetzungen,
3. die Anpassung des Lernangebots an die jeweiligen Voraussetzungen,
4. die Individuelle Anpassung der Begleitung beim Lernen und
5. die Individuelle Diagnostik der Ergebnisse.

Die genannten Bereiche bauen aufeinander auf und sind als integrale Bestandteile von Individualisierung zu sehen und sollten daher nicht isoliert, sondern im Zusammenspiel eingesetzt werden. Im Gegensatz zur Binnendifferenzierung stehen nun das Individuum und seine Voraussetzungen, Interessen etc. im Zentrum.

In Variante 2 können mit „Individualisierung“ Lernangebote z. B. mit Projektcharakter gemeint sein, die für verschiedene (Lern-)Voraussetzungen, Kompetenzniveaus und Interessen Spielräume bieten. Während bei der ersten Variante die Steuerung stärker aufseiten der Lehrperson liegt, sind es bei der zweiten Variante stärker die Schülerinnen und Schüler, die das Angebot ihren Voraussetzungen und Interessen entsprechend nutzen sollen.

Zusammenfassend lassen sich aufgrund der Literatur folgende drei Formen des Unterrichts unterscheiden: 1. Unterricht mit der gesamten Klasse, 2. Binnendifferenzierung und 3. Individualisierung. Dabei werden zwei Arten der Individualisierung und zwei ausgewählte Möglichkeiten der Binnendifferenzierung unterschieden: Binnendifferenzierung als Bildung von Schülerinnen- und Schülergruppen mit gleichem sowie mit unterschiedlichem Fähigkeitsniveau und Individualisierung als *lehrerseitige Anpassung* des Lernangebots an die Voraussetzungen der Schülerinnen und Schüler sowie Individualisierung als *schülerseitige Nutzung des Lernangebots* entsprechend ihren Voraussetzungen und Interessen.

Zum besseren Verständnis der einzelnen Formen soll folgende grafische Darstellung dienen:

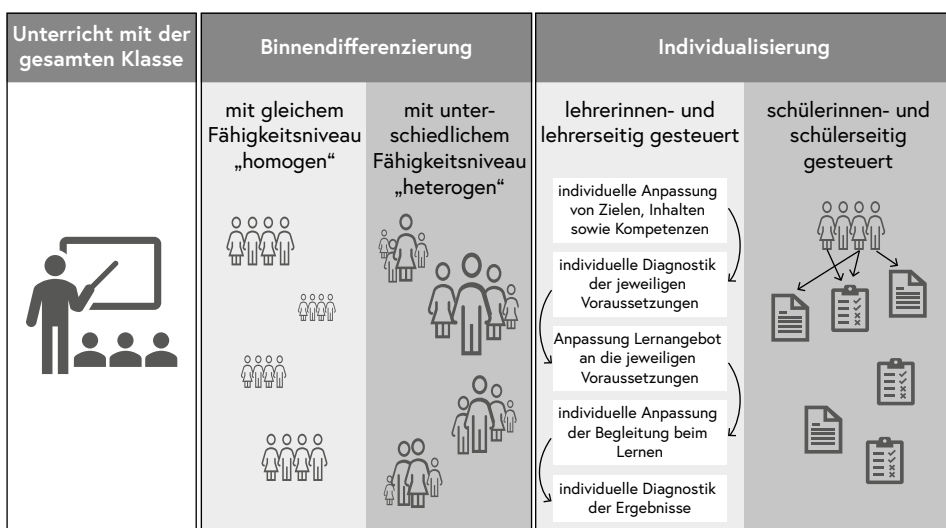


Abbildung 1: Theoriemodell des Beitrags. Quelle: Darstellung durch die Autorinnen.

Da für den vorliegenden Beitrag keine eigenen Fragebögen auf Basis des theoretischen Hintergrundes vorliegen, sondern auf Datenmaterial von bestehenden internationalen Leistungsvergleichsstudien, so genannten *large-scale assessments*, zurückgegriffen wird, können nur ausgewählte Aspekte von „Individualisierung“ und „Differenzierung“ mit den vorhandenen Daten analysiert werden. Diese werden in Abschnitt 3.1 vorgestellt.

2.3 Forschungsstand zur Wirksamkeit von Individualisierung/ Differenzierung

Bezüglich der Wirksamkeit von Individualisierung und Binnendifferenzierung zeichnet die Forschungslage kein eindeutiges Bild.

Individualisierter Unterricht scheint – verglichen mit traditionellem Unterricht – etwas erfolgreicher zu sein. Allerdings liegt der durchschnittliche Effekt deutlich unter der relevanten Grenze von $d = 0.4$ (Hattie, 2009), ab der nach Hattie Effekte als lernwirksam angesehen werden. Konkret gibt Hattie (2009, S. 100) für Binnendifferenzierung eine Effektstärke von $d = 0.16$ und für Individualisierung eine Effektstärke von $d = 0.23$ an. Altrichter et al. (2009) berichten, dass die meisten Studien „keine systematischen Vorteile der jeweiligen Formen individualisierten Unterrichts“ (Altrichter et al., 2009, S. 348) finden konnten.

Für Deutschland konnten Schrader und Helmke (2008) zeigen, dass auf der Primarstufe Differenzierung weder zu besseren noch zu schlechteren Ergebnissen beim Leseverstehen führt. Hess und Lipowsky (2017, S. 25) berichten über Metaanalysen zu verschiedenen Formen der Binnendifferenzierung, in denen sich vor allem für Lernende mittlerer Leistung positive Effekte zwischen $d = 0.12$ und $d = 0.34$ für leistungshomogene im Gegensatz zu leistungsheterogenen Gruppenbildungen zeigen. Für Binnendifferenzierung fasst Fladung (2022, S. 89–91) für den Deutschunterricht die Studienlage folgendermaßen zusammen: Sie kann sich lernförderlich auswirken, vorausgesetzt, die Gruppeneinteilung erfolgt in gezielter Form. Eine Anpassung der Aufgaben an die Gruppen ist aber auch hier besonders förderlich.

In Metastudien, aber auch Einzeluntersuchungen zur Wirksamkeit von Individualisierung und Binnendifferenzierung wurde die Qualität des individualisierten Unterrichts oft nicht erfasst bzw. konnte nicht erfasst werden. Laut Helmke (2015, S. 262, 263) dürfte dieser Aspekt eine entscheidende Rolle bei der Interpretation der Ergebnisse spielen. Helmke (2013, S. 36) erklärt, dass es nicht ausschlaggebend ist, ob differenziert und individualisiert wird, sondern wie dies erfolgt. Die Realisierungsformen von Individualisierung und Binnendifferenzierung können demnach sehr unterschiedlich sein (Altrichter et al., 2009, S. 349). Von ihnen und ihrer konkreten Umsetzung hängt allerdings ab, ob sie ihre Wirksamkeit entfalten können oder nicht. So sind etwa die Einstellungen und Überzeugungen von Lehrpersonen zu Heterogenität, Klassenführung, die didaktischen Materialien, die diagnostische Kompetenz von Lehrpersonen etc. relevante Aspekte für das Gelingen individualisierten Unterrichts oder von Binnendifferenzierung. Dazu kommen noch Faktoren, die nicht un-

bedingt von der Lehrperson beeinflussbar sind, wie Schülervoraussetzungen oder Rahmenbedingungen an Schulen (Bohl et al., 2012, S. 54 f.; Helmke, 2015, S. 257–259).

Insgesamt verdeutlicht die Forschungslage, dass individualisierter, differenzierter oder offener Unterricht keinesfalls per se wirksam ist, sondern im Gegenteil insbesondere mit Blick auf hohe Fachleistungen und mit Blick auf leistungsschwächere Schülerinnen und Schüler intelligent organisiert sein muss. Derartige Konzeptionen sind voraussetzungsreich, komplex, bergen einen hohen Anspruch an die Expertise der Lehrkräfte und sind ressourcenintensiv (Bohl et al., 2012, S. 62).

Individualisierung und Binnendifferenzierung als Prinzip scheint zudem nicht für alle Unterrichtssituationen und Schülerinnen- und Schülergruppen gleich gut einsetzbar zu sein. Empfohlen wird daher oft der gezielte und begründete Einsatz von differenzierenden Phasen wie auch Phasen eines Gesamtunterrichts (Altrichter et al., 2009, S. 350). Je nach Ausgangs- und Lernvoraussetzungen kann zudem derselbe Unterricht auf Schülerinnen und Schüler unterschiedlich wirken. Offener Unterricht kann beispielsweise für Schülerinnen und Schüler mit wenig Vorkenntnissen überfordernd sein, während andere Schülerinnen und Schüler von einer Reduktion an Vorgaben profitieren können (Helmke, 2015, S. 253), und zwar nicht nur hinsichtlich ihrer Leistung, sondern auch mit Blick auf andere Ziele wie eine Stärkung des Gefühls der Mitbestimmung und Partizipation.

2.4 Forschungsfragen

Vor dem skizzierten theoretischen Hintergrund ergeben sich nun folgende Forschungsfragen mit Blick auf Österreich im Zeitverlauf:

1. Unterscheidet sich Österreich in Bezug auf das Ausmaß von Differenzierung bzw. Individualisierung von den in Abschnitt 3.1 erwähnten Vergleichsländern, und zu welchen Zeitpunkten (2006, 2011, 2016, 2021)? (S. Unterabschnitte 4.1 bis 4.3.)
2. Hat sich das Ausmaß an Differenzierung bzw. Individualisierung in Österreich zwischen 2006 und 2021 verändert? (S. Unterabschnitt 4.4.)
3. Gibt es Zusammenhänge zwischen dem Ausmaß an Unterricht mit der gesamten Klasse, an Differenzierung bzw. Individualisierung und der Lesekompetenz? (S. Unterabschnitt 4.5.)

3 Untersuchungsdesign und Fragestellungen

3.1 Verwendete Daten

Die Analysen beruhen auf Daten der PIRLS-Erhebungen (Jahre 2006, 2011, 2016 und 2021) und Daten der Baseline-Testung für die Bildungsstandardüberprüfung (B410I; BIFIE [IQS], 2021; Breit et al., 2016, S. 15) aus dem Jahr 2010. Alle Daten wurden auf der 4. Schulstufe erhoben.

3.1.1 PIRLS

PIRLS (*Progress in International Reading Literacy Study*) ist eine Studie zur Lesekompetenz von 10-jährigen Schülerinnen und Schülern. Sie liefert seit 2001 international vergleichbare Daten, welche Lesekompetenzen und Einstellungen Schülerinnen und Schüler auf der 4. Schulstufe zum Lesen haben. Österreich nimmt erstmals 2006 an der Studie teil, weswegen im hier vorliegenden Kapitel auf die Erhebungszeitpunkte 2006, 2011, 2016 und 2021 Bezug genommen wird. Bei den Stichproben handelt es sich jeweils um ca. 4.500 Schülerinnen und Schüler aus ca. 290 Klassen in 160 Schulen (siehe bspw. bei PIRLS 2021: Mielach, 2023).

Die Analysen beziehen sich auf insgesamt elf wie folgt ausgewählte Länder: Diese sind EU-Länder, haben an den vier PIRLS-Zyklen teilgenommen, an denen auch Österreich teilgenommen hat, haben Daten zu den betreffenden Fragen und a) haben zu einem oder mehreren Erhebungszeitpunkten einen höheren Landesmittelwert in Lesen als Österreich erreicht oder b) sind Nachbarländer Österreichs (s. Abbildung 2).

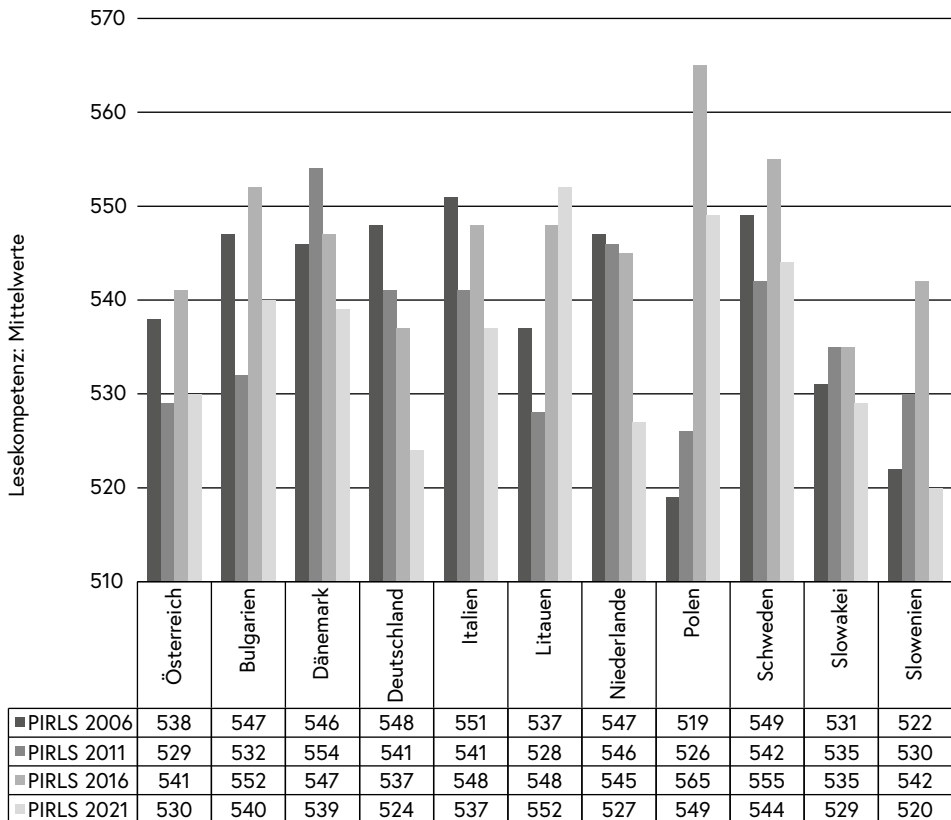


Abbildung 2: Vergleichsländer aus PIRLS. Quelle: PIRLS 2006, 2011, 2016 und 2021

Zu beachten ist, dass die PIRLS-2021-Schülerkohorte von den COVID-19-Beschränkungen der Jahre 2020 und 2021 betroffen ist, die zu weniger regulären Schultagen führten (OECD, 2021). PIRLS 2021 zeigt einen allgemeinen Abwärtstrend, der wahrscheinlich auf die COVID-

19-Pandemie zurückzuführen ist (Mullis et al., 2023, S. 18). Diese Umstände können auch die Möglichkeiten beeinflussen, den Unterricht zu individualisieren bzw. zu differenzieren.

3.1.2 Baseline-Testung 2010: 4. Schulstufe Deutsch/Lesen und Mathematik

Als weitere Datenbasis dient die Baseline-Testung der 4. Schulstufe aus dem Jahr 2010. Breit et al. (2016, S. 15) berichten über die Vorarbeiten für die Standardüberprüfungen, zu denen auch die sogenannte Baseline-Testung zählt. Diese Erhebung ist eine Ausgangsmessung für die späteren Bildungsstandards. Mit dieser Ausgangsmessung werden die späteren Vergleiche (Mathematik: 2015; Deutsch: 2016) ermöglicht. Zusätzlich zu den Kompetenztests werden Fragebögen an Schülerinnen und Schülern und deren Lehrpersonen eingesetzt. Bei der Stichprobe handelt es sich um 9.478 Schülerinnen und Schüler aus 500 Klassen in 267 Schulen (BIFIE [IQS], 2021).

Zu den Bildungsstandards im Allgemeinen in Österreich gibt es bei Schreiner und Wiesner (2019) eine umfassende Darstellung. Zu den Ergebnissen des ersten Überprüfungszyklus von 2012 bis 2016 sei auf Neubacher, Ober, Wimmer und Hartl (2019) verwiesen. Sie greifen auch die Ergebnisse der Baseline-Testung 2010 der 4. Schulstufe in Deutsch Lesen auf. Die Fragebögen zu den Bildungsstandards Deutsch 4. Schulstufe des Jahres 2015 enthalten keine Fragen zum Thema dieses Beitrags, weswegen auf Daten der Baseline-Erhebung zurückgegriffen wird.

3.2 Verwendete Variablen und Einordnung in den theoretischen Hintergrund

Die für dieses Kapitel wichtigsten Fragebogenvariablen der PIRLS-Erhebungen stammen aus dem Lehrerfragebogen mit dem Fragestamm *„Wenn Sie Lesen unterrichten und/oder lesebezogene Aktivitäten mit den Schüler/innen durchführen: Wie oft organisieren Sie diese in folgender Weise?“* und den folgenden fünf Aussagen:

Tabelle 1: Variablen zur Organisation lesebezogener Aktivitäten laut PIRLS-Fragebogen und Einordnung in den theoretischen Hintergrund. Quelle: Lehrerfragebögen PIRLS 2006, 2011, 2016 und 2021.

Aussagen (Variablen)	Einordnung in den theoretischen Hintergrund (s. Abschnitt 2.1 und 2.2)
Ich unterrichte die ganze Klasse gemeinsam in Lesen.	Unterricht mit der gesamten Klasse (keine Individualisierung oder Binnendifferenzierung)
Ich bilde Gruppen von Schülerinnen und Schülern mit gleichem Fähigkeitsniveau.	Binnendifferenzierung („homogen“)
Ich bilde Gruppen von Schülerinnen und Schülern mit unterschiedlichem Fähigkeitsniveau.	Binnendifferenzierung („heterogen“)
Ich gebe individuell abgestimmten Leseunterricht.	Individualisierung als lehrerseitige Anpassung des Lernangebots an die Voraussetzungen der Schülerinnen und Schüler
Die Schülerinnen und Schüler arbeiten selbstständig nach einem festgelegten Plan oder einer Zielvorgabe.	Individualisierung als schülerseitige Nutzung des Lernangebots entsprechend ihren Voraussetzungen und Interessen

Die Antwortoptionen sind 1 = immer oder fast immer, 2 = oft, 3 = manchmal und 4 = nie.

Für zwei PIRLS-Zeitpunkte (2016 und 2021) gibt es weiters den Fragestamm „*Wie oft machen Sie Folgendes während des Leseunterrichts in dieser Klasse?*“ mit drei zum Thema passenden Aussagen, die einzelne Bereiche von Individualisierung als lehrerseitige Anpassung des Lernangebots an die Voraussetzungen der Schülerinnen und Schüler näher bestimmen:

Tabelle 2: Variablen zur Individualisierung als lehrerseitige Anpassung des Leseunterrichts laut PIRLS-Fragebogen und Einordnung in den theoretischen Hintergrund. Quelle: Lehrerfragebögen PIRLS 2016 und 2021

Aussagen (Variablen)	Einordnung in den theoretischen Hintergrund (s. Abschnitt 2.1 und betreffende Unterpunkte)
Bei neuen Inhalten an Vorkenntnisse der Schülerinnen und Schüler anknüpfen.	Individuelle Anpassung von Zielen, Inhalten sowie Kompetenzen
Jedem Schüler/jeder Schülerin individuelle Rückmeldung geben.	Die individuelle Diagnostik der jeweiligen Voraussetzungen bzw. der Ergebnisse
Materialien bereitstellen, die dem Leseniveau einzelner Schülerinnen und Schüler entsprechen.	Anpassung des Lernangebots an die jeweiligen Voraussetzungen

Die Antwortoptionen sind 1 = in jeder oder fast jeder Stunde, 2 = in ungefähr der Hälfte der Stunden, 3 = in einigen Stunden und 4 = nie.

Aus dem Lehrerfragebogen der Baseline-Testung 2010 gehören fünf Aussagen zum Themenbereich Binnendifferenzierung/Individualisierung im Unterricht. Der Fragestamm lautet „*Wie oft kommt Folgendes in Ihrem Unterricht in Deutsch und Mathematik vor?*“. Da sich der Fragestamm und dadurch in der Folge die Lehrerantworten auf Deutsch und Mathematik beziehen, ist die Interpretierbarkeit in Bezug auf Lesen eher eingeschränkt. Folgende fünf Aussagen werden verwendet:

Tabelle 3: Variablen zur Organisation lesebezogener Aktivitäten laut Baseline-Fragebogen und Einordnung in den theoretischen Hintergrund. Quelle: Lehrerfragebogen Baseline 2010 D4/M4.

Aussagen (Variablen)	Einordnung in den theoretischen Hintergrund (s. Abschnitt 2.1 und 2.2)
Ich unterrichte die Schülerinnen und Schüler in der Gesamtgruppe.	Unterricht mit der gesamten Klasse (keine Individualisierung oder Binnendifferenzierung)
Ich bilde Gruppen von Schülerinnen und Schülern mit gleichen Fähigkeitsniveaus und gebe nur Schülerinnen und Schülern in einer Gruppe dieselben Aufgaben.	Binnendifferenzierung („homogen“)
Ich bilde Gruppen von Schülerinnen und Schülern mit unterschiedlichen Fähigkeitsniveaus und gebe allen Gruppen dieselben Aufgaben.	Binnendifferenzierung („heterogen“)
Ich gebe den Schülerinnen und Schülern je nach Können unterschiedliche Aufgaben.	Individualisierung als lehrerseitige Anpassung des Lernangebots an die Voraussetzungen der Schülerinnen und Schüler
Ich erkläre den einzelnen Schülerinnen und Schülern etwas so lange, bis es alle verstanden haben.	Individualisierung als lehrerseitige Anpassung des Lernangebots an die Voraussetzungen der Schülerinnen und Schüler

Die Antwortoptionen sind 1 = in (fast) jeder Stunde, 2 = in den meisten Stunden, 3 = in einigen Stunden und 4 = nie.

Bei der individuellen Rückmeldung an Schülerinnen und Schüler wird in den Fragebögen nicht näher spezifiziert, ob diese aufgrund von Individualdiagnostik oder beispielsweise Beobachtungen durch die Lehrperson erfolgt. Um sich dieser Frage anzunähern, wird ergänzend folgende Fragestellung aus dem PIRLS-Fragebogen für Österreich (nationaler Zusatz) betrachtet:

„Welches Gewicht messen Sie den folgenden Quellen bei, um den Fortschritt der Schüler/innen im Lesen zu beurteilen?“

Tabelle 4: Variablen zur Organisation lesebezogener Aktivitäten laut PIRLS-Fragebogen und Einordnung in den theoretischen Hintergrund. Quelle: Lehrerfragebögen PIRLS 2006, 2011, 2016 und 2021.

Aussagen (Variablen)	Einordnung in den theoretischen Hintergrund (s. Abschnitt 2.1 und betreffende Unterpunkte)
Tests und Schularbeiten (z. B. vom Lehrer/der Lehrerin entworfene oder dem Lehrbuch entnommene Tests).	Individuelle Diagnostik der Ergebnisse
Diagnostische Lesetests (einschließlich Lesefehleranalyse).	Individuelle Diagnostik der Ergebnisse

Die Antwortoptionen sind 1 = großes Gewicht, 2 = einiges Gewicht, 3 = wenig oder kein Gewicht. Zu dieser Frage liegen Daten aus den Erhebungszeitpunkten 2006, 2011 und 2016 vor.

Es wird zudem nicht erfasst, ob Lehrpersonen, die ihren Leseunterricht individualisieren, alle fünf der genannten Bereiche der Individualisierung oder nur einzelne davon berücksichtigen und sie in der von der Literatur intendierten Abfolge und Bezogenheit aufeinander einsetzen. Aus den genannten Gründen kann keine vollumfängliche Analyse der Konzepte erfolgen, sondern nur eine Annäherung anhand ausgewählter Aspekte. Auch dies ist bei der Interpretation der Ergebnisse zu berücksichtigen.

Mit Blick auf Individualisierung als schülerseitige Nutzung des Lernangebots entsprechend ihren Voraussetzungen und Interessen muss einschränkend erwähnt werden, dass die Variable „Die Schüler/innen arbeiten selbstständig nach einem festgelegten Plan oder einer Zielvorgabe.“ Spielraum für mehrere Interpretationen lässt. Es kann nicht ausgeschlossen werden, dass die Lehrpersonen bei der Beantwortung der Frage unter „selbstständigem Arbeiten nach Plan oder Zielvorgabe“ weniger Projekt- bzw. offenen Unterricht vor Augen haben als etwa das selbstständige Bearbeiten von Leseaufgaben und Arbeitsblättern ohne Individualisierung bzw. Differenzierung oder eine lehrergesteuerte Festlegung von Zielen, die dann von den Schülerinnen und Schülern bearbeitet werden. Für diese Variante der Individualisierung wäre allerdings entscheidend, dass die Schülerinnen und Schüler aus

verschiedenen Lernangeboten selbst das passende für sich auswählen oder zumindest Wahlmöglichkeiten (etwa aus Lernangeboten zu unterschiedlichen Fähigkeitsniveaus, Themengebieten etc.) haben.

Zudem sei erwähnt, dass im Bereich der Binnendifferenzierung in den PIRLS-Fragebögen und im Baseline-Fragebogen jeweils leicht unterschiedliche Formen erfasst werden: Während es in den PIRLS-Fragebögen lediglich um die Bildung von Gruppen mit gleichem bzw. unterschiedlichem Fähigkeitsniveau geht, enthält der Baseline-Fragebogen darüber hinaus eine Differenzierung nach Aufgaben. Andere Formen der Binnendifferenzierung (z. B. nach Aufgabenstellungen und -formaten, nach der Bearbeitungszeit oder dem -umfang, nach Themen etc. [Fladung, 2022, S. 86]) wurden hier nicht betrachtet.

3.3 Analysemethoden

Die PIRLS-Lehrerangaben werden auf Schülerinnen- und Schülerebene übertragen. Dieses kombinierte Schüler-Lehrer-Datenfile wird auf Lehrerebene gewichtet (TCHWGT; Fishbein, Liquin & Foy, 2024).

Berechnet werden Prozentwerte von Schülerinnen und Schülern, die eine Lehrperson haben, die das betreffende Merkmal angegeben hat, sowie ein Zusammenhang zwischen diesen Angaben und der Lesekompetenz (ASRREA01-10) für Österreich. Weiters wird berichtet, sofern sich eine statistisch signifikante Veränderung zwischen Zeitpunkten innerhalb von Ländern ergibt, und welche Länder sich zu welchen Erhebungszeitpunkten von Österreich unterscheiden.

Die Baseline-2010-Lehrerangaben werden auf Schülerinnen- und Schülerebene übertragen, das kombinierte Schülerinnen/Schüler-Lehrerinnen/Lehrer-Datenfile auf Klassenebene gewichtet. Die Analysen beziehen sich nur auf Österreich. Berichtet werden Prozentwerte von Schülerinnen und Schülern, die eine Lehrperson haben, die das betreffende Merkmal angegeben hat, sowie ein Zusammenhang zwischen diesen Angaben und der Lesekompetenz (PVD4L) für Österreich.

4 Ergebnisse

Die Ergebnisdarstellung folgt der referenzierten theoretischen Basis (s. Abschnitt 2.1 und 2.2: Bohl et al., 2012) und der zur Orientierung angebotenen grafischen Umsetzung (s. Abbildung 1).

Im Detail werden zur jeweiligen Aussage zuerst Österreichs Ergebnisse berichtet. Danach folgen im Überblick die Ergebnisse der Vergleichsländer. Welche Länder sich im Detail von Österreich zu welchen Zeitpunkten statistisch signifikant unterscheiden, enthält jeweils die unterste Zeile der Abbildungen (* = signifikant unterschiedlich zu Österreich; $p \leq .05$).

4.1 Unterricht der ganzen Klasse in Lesen

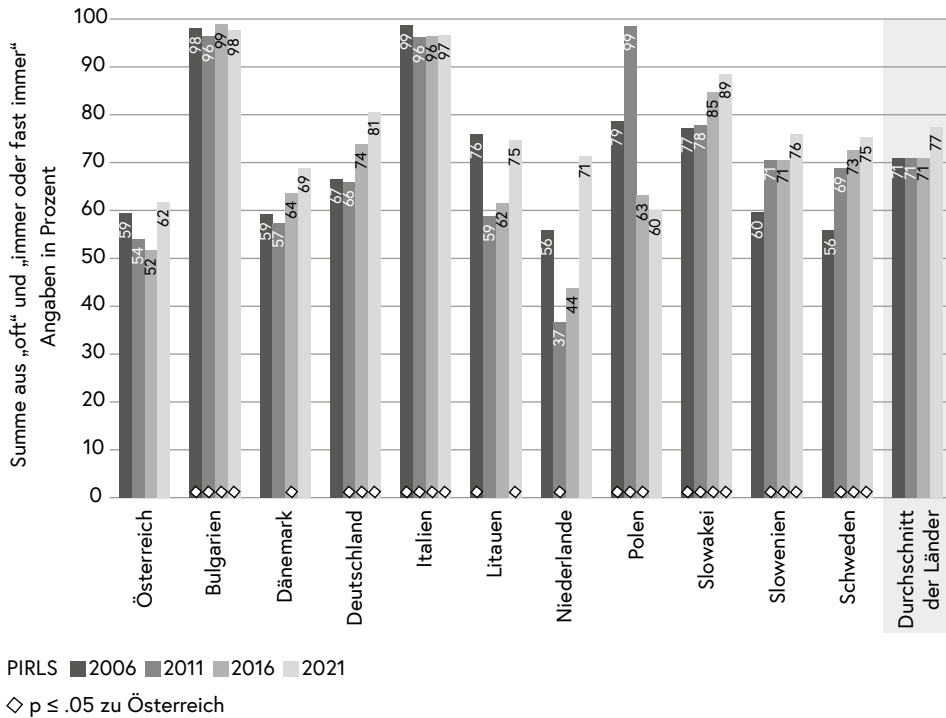


Abbildung 3: Gemeinsamer Leseunterricht im Klassenverband im 15-Jahres-Trend. Quellen: PIRLS 2006, 2011, 2016 und 2021. Berechnungen und Darstellung durch die Autorinnen.

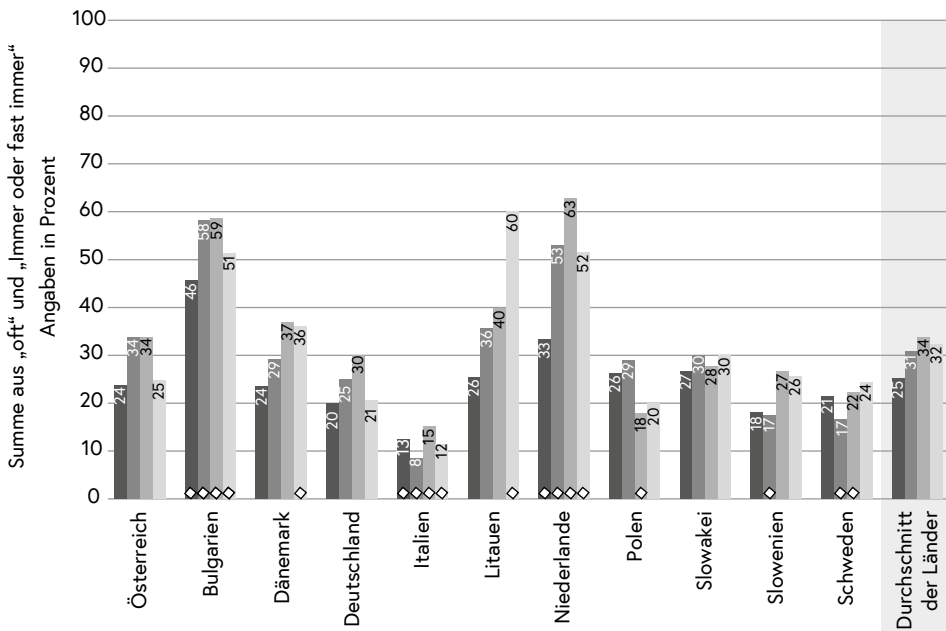
Der gemeinsame Leseunterricht der ganzen Klasse ist als Ausgangspunkt der Analysen zu sehen (Abbildung 3). Österreich fällt als eines der Länder auf, bei dem der Anteil an Kindern, die von ihren Lehrpersonen mindestens „oft“ im Klassenverband in Lesen unterrichtet werden, eher niedrig ist. Zu allen vier Erhebungszeitpunkten liegen die nationalen Ergebnisse unter dem Durchschnitt der Länder, was dafürsprechen könnte, dass in Österreich andere Unterrichtsformen häufiger als in den Vergleichsländern eingesetzt werden. Der Anteil an Kindern, die von einer Lehrperson unterrichtet werden, die angibt, den Klassenverband in Lesen gemeinsam mindestens „oft“ zu unterrichten, ist zwischen 2006 und 2016 rückläufig. Dies deutet auf eine entsprechende vermehrte Verwendung anderer Unterrichtsformen hin. Zwischen 2016 und 2021 gibt es eine statistisch signifikante Steigerung des gemeinsamen Leseunterrichts in Österreich.

Bezüglich der Vergleichsländer sind Bulgarien und Italien mit sehr hohen Anteilen für den gemeinsamen Leseunterricht im Klassenverband nennenswert. Für das Jahr 2021 wiederum zeigt sich für fast alle Länder eine Steigerung (außer insbesondere Polen, aber auch Bulgarien und Italien) für den gemeinsamen Leseunterricht im Klassenverband, so auch für Österreich. Dies ist möglicherweise mit den pandemiebedingten Einschränkungen (Fernunterricht) zu erklären.

Inwieweit die polnische Bildungsreform 2009 (Schuleintritt bereits mit sechs statt mit sieben Jahren) und ihre Rücknahme im Jahr 2016 oder der neue Lehrplan für die Volksschule in Polen aus 2017 zum Rückgang des Leseunterrichts mit der gesamten Klasse beigetragen haben mögen, kann hier nicht geklärt werden (Kazmierczak, 2022). Der Lehrplan 2017 greift Individualisierung jedenfalls explizit auf (Verordnung der Ministerin für nationale Bildung in Polen, Nr. 356/2017), wenngleich er bei seiner Einführung für eine Renaissance des Frontalunterrichts kritisiert wurde (Kellermann, 2017).

4.2 Binnendifferenzierung

Das Konzept der Binnendifferenzierung sieht im Unterschied zur „Individualisierung“ keine individuelle Erfassung der Voraussetzungen und keine individuelle Anpassung des Lernangebots vor. Vielmehr steht bei der Binnendifferenzierung die Gruppierung von Schülerinnen und Schülern nach bestimmten Merkmalen wie Kompetenzniveau, Lernvoraussetzungen oder Interesse im Fokus (s. Abschnitt 2.1).



PIRLS ■ 2006 ■ 2011 ■ 2016 ■ 2021

◇ $p \leq .05$ zu Österreich

Abbildung 4: Binnendifferenzierung durch Gruppen mit Kindern mit gleichem Fähigkeitsniveau im 15-Jahres-Trend. Quellen: PIRLS 2006, 2011, 2016 und 2021. Berechnungen und Darstellung durch die Autorinnen.

In Österreich haben 2021 25 % der Schülerinnen und Schüler eine Lehrperson, die angibt, Schülerinnen- und Schülergruppen mit gleichem Fähigkeitsniveau zu unterrichten.

Damit liegt Österreich unter dem Schnitt der Vergleichsländer (32 %). Der Trend zeigt, dass Österreich 2021 auf einem ähnlichen Niveau wie 2006 liegt, während 2011 und 2016 der Wert um rund 10 % mehr beträgt (Abbildung 4). Der Anstieg von 2006 auf 2011 sowie der Rückgang von 2016 auf 2021 sind statistisch signifikant.

Deutlich niedrigere Werte bei der Binnendifferenzierung mit gleichem Fähigkeitsniveau erreicht Italien, während in Bulgarien, Litauen (mit einem Anstieg von 2016 auf 2021) und den Niederlanden deutlich häufiger Schülerinnen- und Schülergruppen mit gleichem Fähigkeitsniveau gebildet werden.

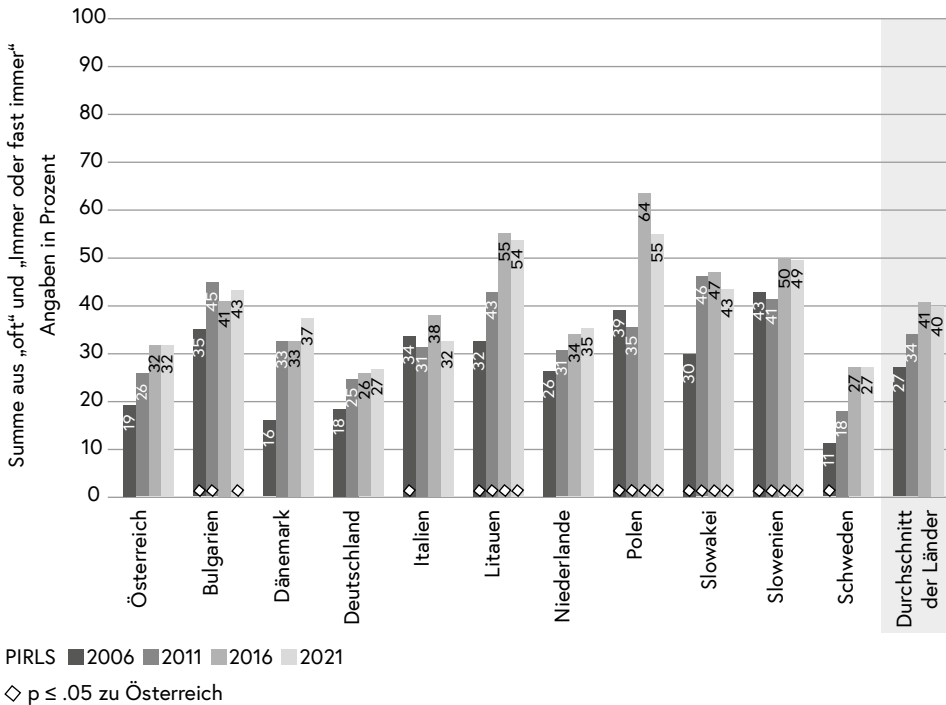


Abbildung 5: Binnendifferenzierung durch Gruppen mit Kindern mit unterschiedlichem Fähigkeitsniveau im 15-Jahres-Trend. Quellen: PIRLS 2006, 2011, 2016 und 2021. Berechnungen und Darstellung durch die Autorinnen.

Im Bereich der Binnendifferenzierung nach unterschiedlichem Fähigkeitsniveau liegt Österreich ebenfalls unter dem Durchschnitt der Vergleichsländer. Allerdings ist zugleich zu beobachten, dass der Wert von 19 % der Schülerinnen und Schüler (2006), deren Lehrperson angibt, Schülerinnen- und Schülergruppen mit unterschiedlichem Fähigkeitsniveau zu bilden, statistisch signifikant auf 32 % (2011) der Schülerinnen und Schüler etwas angestiegen und geblieben (2021) ist (Abbildung 5).

Tendenziell zeigt sich eine Zunahme an Binnendifferenzierung nach unterschiedlichem Fähigkeitsniveau für die Mehrheit der Vergleichsländer. Dabei ist Polen das Land mit dem höchsten Wert innerhalb der Vergleichsgruppe.

4.3 Individualisierung

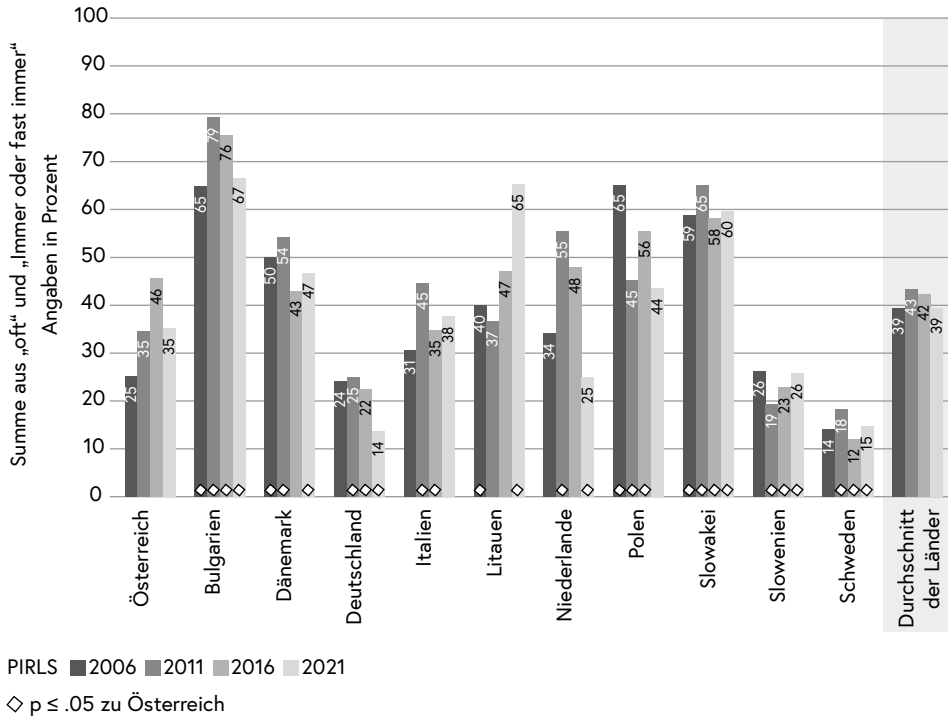


Abbildung 6: Individuell abgestimmter Leseunterricht im 15-Jahres-Trend. Quellen: PIRLS 2006, 2011, 2016 und 2021. Berechnungen und Darstellung durch die Autorinnen.

Im Durchschnitt der vier Erhebungszeitpunkte geben 35 % der österreichischen Lehrpersonen an, dass sie mindestens „oft“ individuell abgestimmten Leseunterricht geben (Abbildung 6). Die nationalen Ergebnisse liegen zu allen vier Zeitpunkten unter dem Durchschnitt der gezeigten Länder. Während es zwischen 2006 und 2016 zu einer Zunahme des individuell abgestimmten Leseunterrichts kommt, ist für 2021 (eventuell pandemiebedingt) ein Rückgang beobachtbar. Zwischen allen vier Erhebungszeitpunkten gibt es statistisch signifikante Veränderungen (2006–2011–2016: signifikante Zunahme; 2016–2021: signifikante Abnahme).

Nennenswert in Bezug auf die Vergleichsländer ist Bulgarien, welches trotz eines hohen Anteils an gemeinsamem Leseunterricht im Klassenverband ebenfalls hohe Angaben bei individuell abgestimmtem Leseunterricht zeigt.

Mavrodieva und Damyanova (2022) beschreiben bezüglich Bulgarien im „country chapter“ der Enzyklopädie zu PIRLS 2021, dass Kinder mit Leseschwierigkeiten anhand eines individualisierten Programms Unterstützung und Förderung erhalten. Das Programm wurde in Bulgarien 2014 eingeführt.

In der Slowakei wird für Kinder mit Leseschwierigkeiten ein „individual study plan“ durch Lehrerinnen und Lehrer erstellt und mit den Eltern der betreffenden Kinder abgestimmt. In regelmäßigen Abständen wird das betreffende Kind getestet, um herauszufinden, ob es dem „individual study plan“ weiterhin folgt oder ob es ausreichend aufgeholt hat (Kopas, 2022).

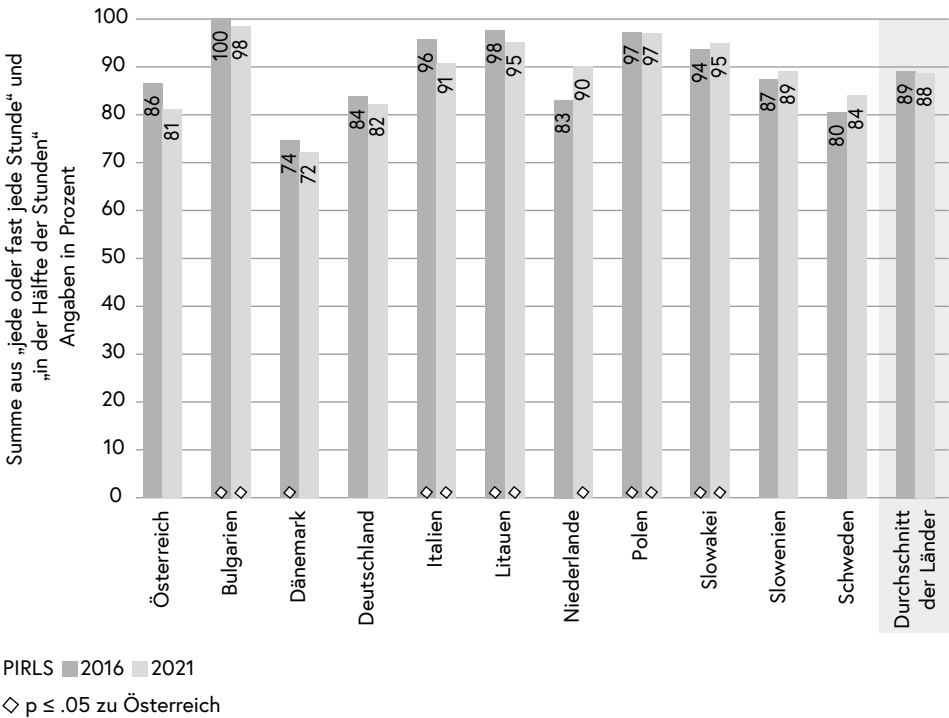


Abbildung 7: Vorkenntnisse und neue Inhalte verknüpfen im 5-Jahres-Trend. Quellen: PIRLS 2016 und 2021. Berechnungen und Darstellung durch die Autorinnen.

Blickt man nun näher auf die in Abschnitt 2.1 dargestellten fünf Bereiche von Individualisierung,³ so zeigt sich Folgendes: Für den ersten Bereich der individuellen Anpassung von Zielen, Inhalten und Kompetenzen ist beobachtbar, dass Österreich beim Verknüpfen von Vorkenntnissen mit neuen Inhalten leicht unter dem Durchschnitt der betrachteten Länder liegt. 2021 haben in Österreich 81 % der Schülerinnen und Schüler eine Lehrkraft, die in mindestens der Hälfte der Stunden neue Inhalte mit Vorkenntnissen verknüpft (Abbildung 7). Hier ist ein leichter Rückgang von 2016 zu 2021 zu beobachten (möglicherweise coronabedingt).

Verglichen mit den anderen Ländern verzeichnet 2021 nur Dänemark mit 72 % niedrigere Werte als Österreich und Bulgarien gemeinsam mit Litauen, Polen und der Slowakei sehr hohe Werte (min. 95 %).

3 Die Daten beziehen sich auf alle befragten Lehrpersonen, nicht nur auf jene, die angeben, individualisierten Unterricht zu geben.

Offen muss an dieser Stelle bleiben, wie die Vorkenntnisse der Schülerinnen und Schüler bestimmt werden, mit denen neue Inhalte verknüpft werden. Denkbar wäre z. B. durch Einschätzung der Lehrperson, durch thematisches Anschließen an bereits im Unterricht behandelte Inhalte oder durch individuelle Diagnostik. Individualisierung – wie in Abschnitt 2.1 beschrieben – würde nach einer individuellen Anpassung von Zielen, Inhalten und Kompetenzen jedenfalls eine individuelle Diagnostik der Voraussetzungen vorsehen. In diesem Zusammenhang sind Einstellungen von Lehrpersonen gegenüber diagnostischen Tests interessant. Fragt man nach dem Gewicht, das Lehrpersonen in Österreich diagnostischen Lesetests beimessen, um den Fortschritt der Schülerinnen und Schüler im Lesen zu beurteilen, so zeigt sich folgendes Bild (zu dieser Frage liegen nur Daten für Österreich vor):

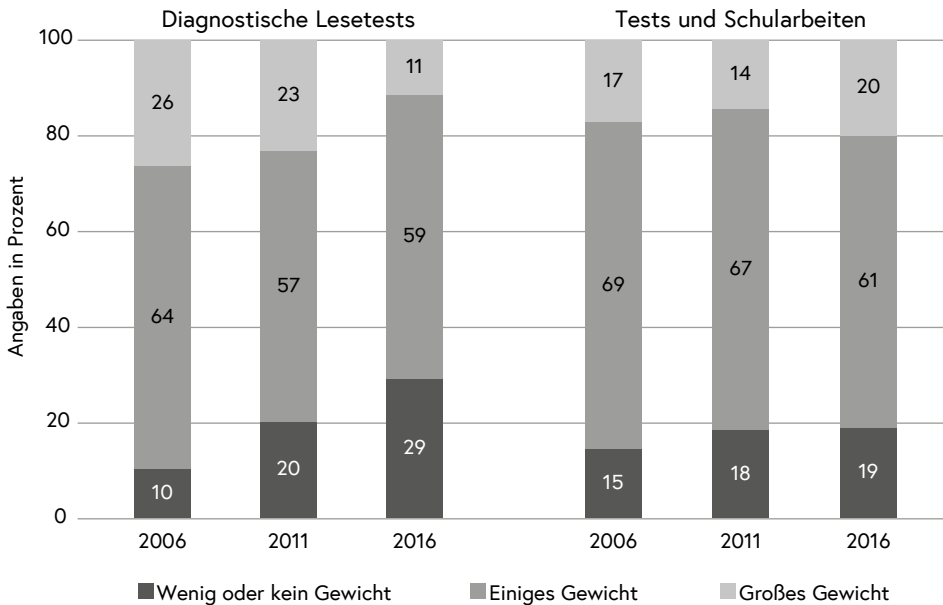


Abbildung 8: Strategien zur Beurteilung der Leseleistung (Bedeutung von Tests/Schularbeiten und diagnostischen Lesetests für die Erfassung des Lesefortschritts) im 10-Jahres-Trend für Österreich. Quellen: PIRLS 2006, 2011 und 2016. Berechnungen und Darstellung durch die Autorinnen.

2006 haben noch 26 % der Schülerinnen und Schüler in Österreich eine Lehrperson, die diagnostischen Lesetests großes Gewicht bei der Beurteilung ihres Lesefortschritts beimisst (Abbildung 8). Dieser Wert sinkt im 10-Jahres-Vergleich auf 11 % der Schülerinnen und Schüler statistisch signifikant. Während also die Bedeutung von diagnostischen Lesetests für Lehrpersonen zwischen 2006 und 2016 deutlich sinkt, verändert sich die Bedeutung von Tests und Schularbeiten zur Beurteilung des Lesefortschritts im 10-Jahres-Vergleich kaum (und statistisch nicht signifikant).

Als nächsten Schritt sieht das Konzept der Individualisierung wie oben beschrieben vor, dass das Lernangebot an die jeweiligen Voraussetzungen angepasst wird, wobei diese

wiederum idealerweise durch Individualdiagnostik festgestellt werden. Wie das Leseniveau der Schülerinnen und Schüler bestimmt wird, war nicht Teil der Fragestellung.

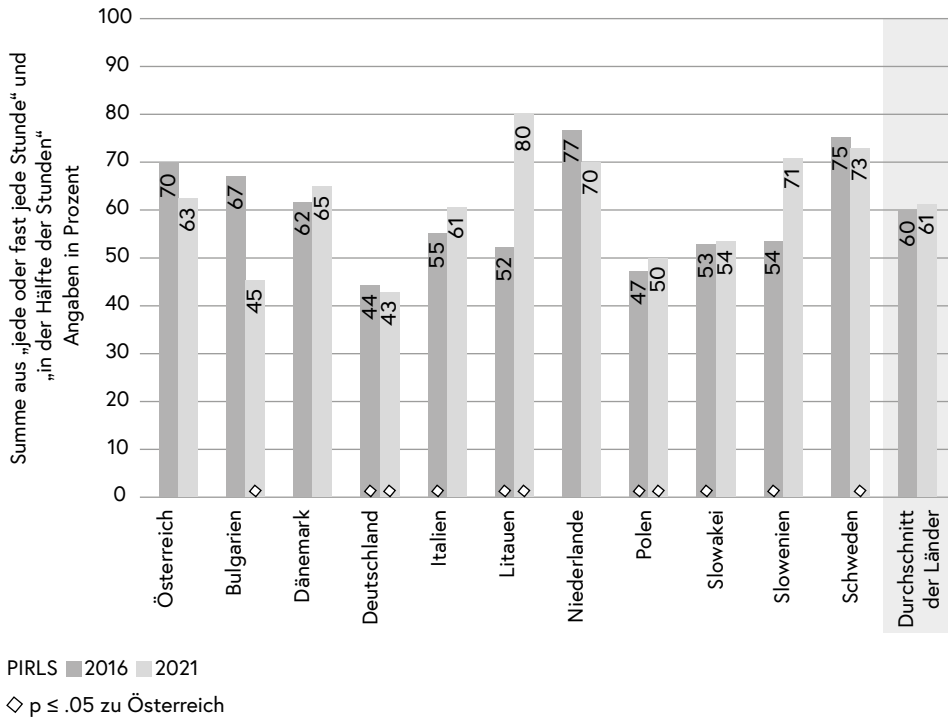


Abbildung 9: An Leseniveau angepasstes Material bereitstellen im 5-Jahres-Trend. Quellen: PIRLS 2016 und 2021. Berechnungen und Darstellung durch die Autorinnen.

Rund zwei Drittel der österreichischen 10-Jährigen werden nach Angaben ihrer Lehrpersonen mindestens „in ungefähr der Hälfte der Stunden“ mit an das Leseniveau angepassten Materialien versorgt (Abbildung 9). Auch dieser Anteil ist zwischen 2016 auf 2021 leicht rückläufig. In Österreich werden an das Leseniveau angepasste Materialien etwas häufiger bereitgestellt als im Durchschnitt der Länder.

Der Blick auf die Länderergebnisse ergibt, dass Schweden generell hohe Anteile aufweist und sich für Litauen und Slowenien deutliche Zuwächse zeigen, während Bulgarien wiederum einen deutlichen Rückgang verzeichnet.

Zum nächsten Bereich – der individuellen Anpassung der Begleitung beim Lernen – finden sich in den begleitenden Fragebögen zu den PIRLS-Erhebungen keine spezifischen Fragen.

Schließlich wird in den PIRLS-Fragebögen erfasst, wie häufig den Schülerinnen und Schülern individuelle Rückmeldung gegeben wird.

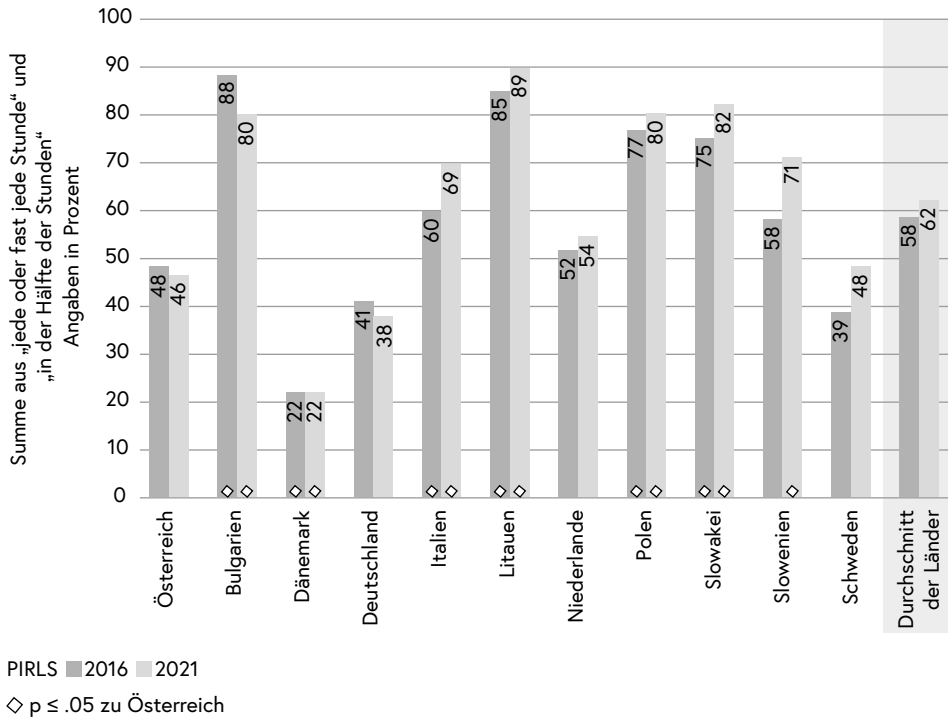


Abbildung 10: Individuelle Rückmeldung geben im 5-Jahres-Trend. Quellen: 2016 und 2021. Berechnungen und Darstellung durch die Autorinnen.

In Österreich hat ca. die Hälfte der Schülerinnen und Schüler eine Lehrkraft, die in mindestens der Hälfte der Stunden individuelle Rückmeldung gibt. Der Wert verändert sich zwischen 2016 und 2021 kaum. Damit liegt Österreich unter dem Schnitt der Vergleichsländer (58 % 2016 bzw. 62 % 2021) (Abbildung 10).

Im Vergleich mit den ausgewählten Ländern fällt auf, dass Bulgarien, Litauen, Polen und die Slowakei häufig individuelle Rückmeldung geben und Dänemark deutlich unterdurchschnittliche Werte aufweist.

Neben dieser ersten, eher von der Lehrperson gesteuerten Variante von Individualisierung wird in Abschnitt 2.1 eine zweite, eher von den Schülerinnen und Schülern ausgehende Variante skizziert. Die Schülerinnen und Schüler wählen dabei aus Lernangeboten mit verschiedenen Kompetenzniveaus, Interessen und Voraussetzungen die für sie passenden Angebote vorwiegend selbst aus.

In Österreich haben ca. 70 % der Schülerinnen und Schüler eine Lehrperson, die angibt, dass die Schülerinnen und Schüler mindestens oft selbstständig nach einem festgelegten Plan oder einer Zielvorgabe beim Lesen arbeiten. Dieser Wert hat sich in den letzten 15

Jahren kaum verändert und ist zuletzt (2021) leicht angestiegen. Österreich liegt damit über dem Durchschnitt der Vergleichsländer (Abbildung 11).

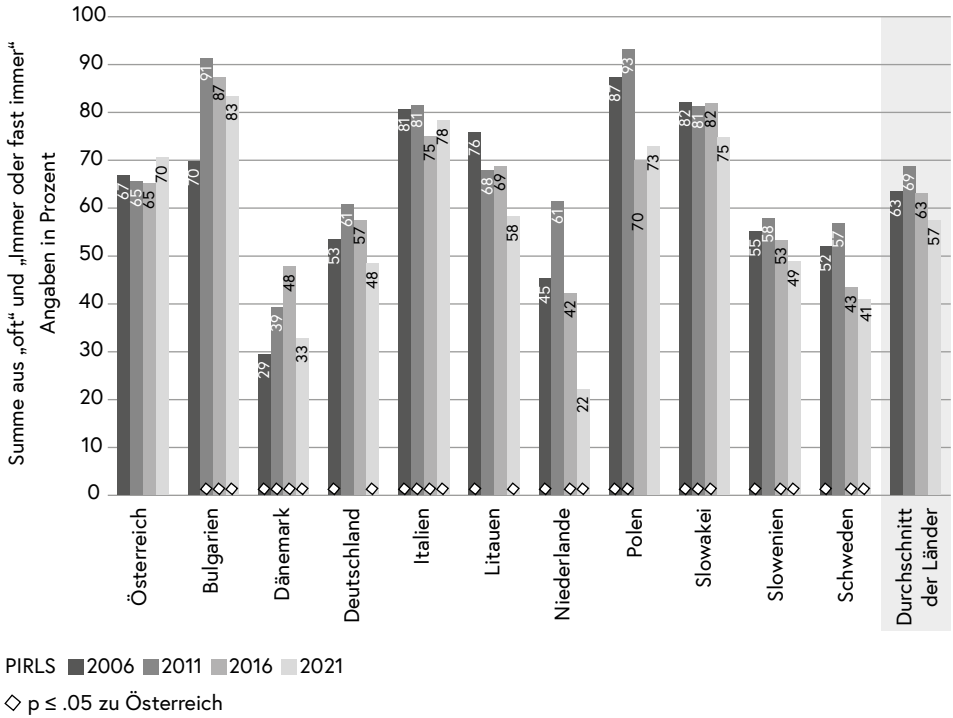


Abbildung 11: Individualisierung nach Plan bzw. einer Zielvorgabe im 15-Jahres-Trend. Quellen: PIRLS 2006, 2011, 2016 und 2021. Berechnungen und Darstellung durch die Autorinnen.

Deutlich höhere Werte als Österreich verzeichnet Bulgarien (statistisch signifikanter Anstieg von 2006 auf 2011). In allen Vergleichsländern bis auf Italien und Österreich ist die Häufigkeit der Individualisierung nach Plan oder Zielvorgabe rückläufig.

4.4 Ergebnisse für Österreich im Überblick

4.4.1 PIRLS 2006–2021

Abbildung 12 enthält zusammengefasst für Österreich die Anteile an 10-Jährigen, deren Lehrpersonen angeben, das jeweilige Lehr-Lern-Szenario mindestens „oft“ im Unterricht einzusetzen. Der Überblick zeigt, dass eine Individualisierung nach Plan bzw. einer Zielvorgabe in Österreichs Leseunterricht in Bezug auf die Häufigkeit führend ist, gefolgt vom gemeinsamen Leseunterricht der ganzen Klasse. An dritter Stelle rangiert der individuell abgestimmte Leseunterricht. Die beiden Arten der Binnendifferenzierung kommen ungefähr gleich häufig vor.

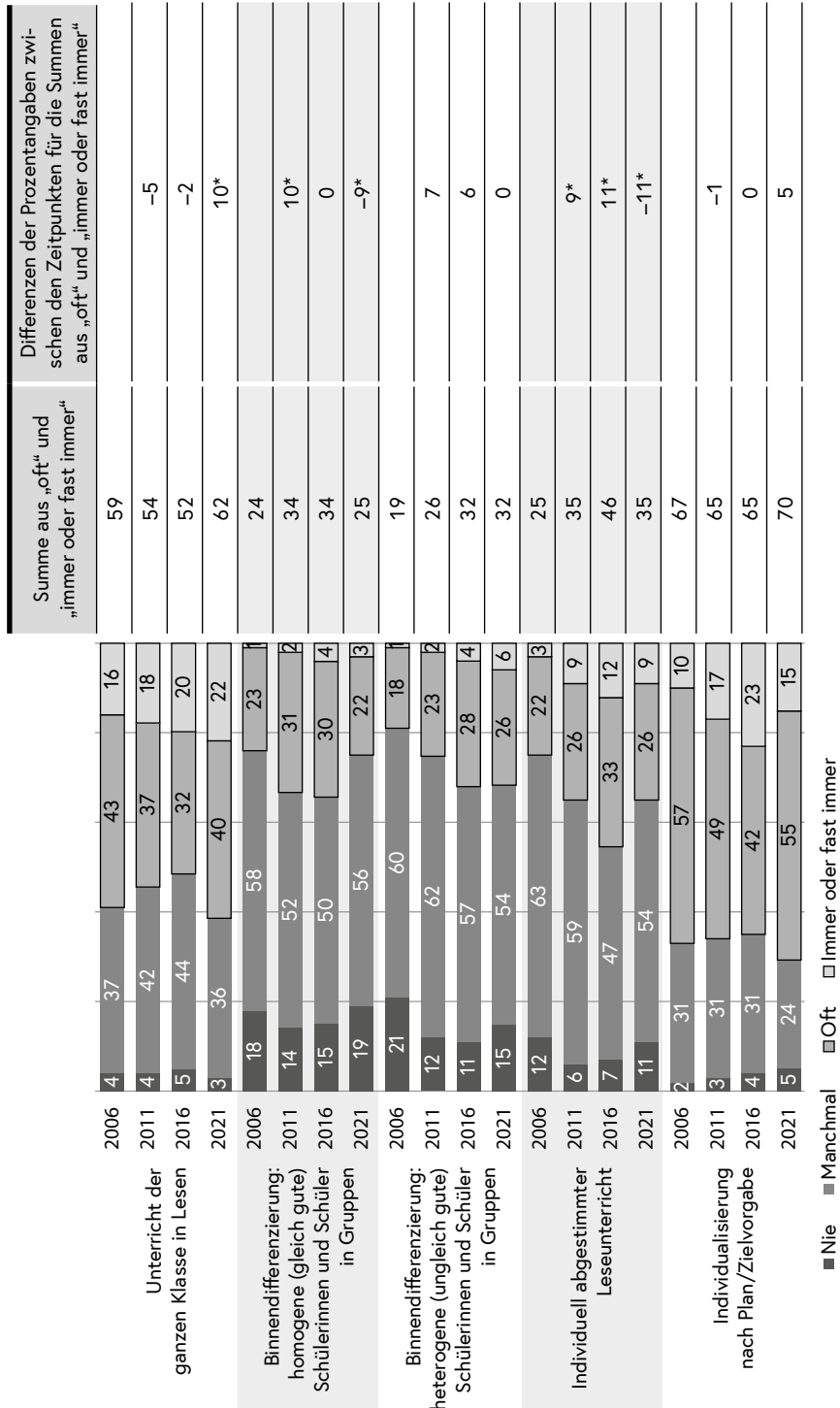


Abbildung 12: Überblick über Individualisierung und Differenzierung in Österreich im 15-Jahres-Trend. Quellen: PIRLS 2006, 2011, 2016 und 2021. Berechnungen und Darstellung durch die Autorinnen. Signifikante Unterschiede sind mit * markiert.

4.4.2 Baseline-2010 D4/M4

Ergänzend zu den PIRLS-Daten folgen nun ähnliche Fragestellungen aus dem Baseline-Fragebogen der Bildungsstandardüberprüfung für die 4. Schulstufe, die im Jahr 2010 erhoben werden.

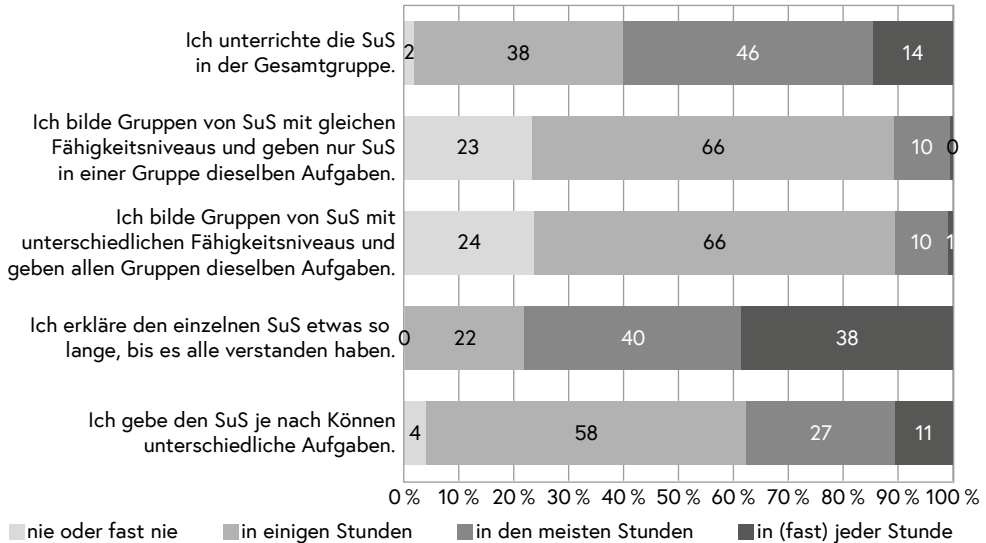


Abbildung 13: Binnendifferenzierung und Individualisierung in Österreich. Quelle: Baseline 2010 D4/M4. Berechnungen und Darstellung durch die Autorinnen.

Auch hier zeigt sich, dass der Unterricht der gesamten Gruppe laut Lehrerangaben eine der am häufigsten eingesetzten Formen ist: 60 % der Schülerinnen und Schüler haben eine Lehrperson, die in den meisten oder in (fast) jeder Stunde die gesamte Gruppe ohne Differenzierung unterrichtet. Auch die Höhe dieses Befunds deckt sich mit den Ergebnissen zu PIRLS. Die beiden Formen der Binnendifferenzierung liegen mit rund je 11 % deutlich hinter den Werten für den Unterricht der Gesamtgruppe und für individualisierte Aufgaben (Abbildung 13). Hier unterscheiden sich die Ergebnisse im Vergleich zu PIRLS, in dem im 15-Jahres-Durchschnitt von 29 % (gleiche Fähigkeitsniveaus) bzw. 27 % (unterschiedliche Fähigkeitsniveaus) berichtet wird. Allerdings ist auch die Fragestellung, wie in Abschnitt 3 erwähnt, inhaltlich nicht gleich. Die Frage nach der Individualisierung nach Plan bzw. einer Zielvorgabe, die sich in den PIRLS-Daten als die häufigste Form des Lesunterrichts zeigt, wird im Baseline-Fragebogen nicht gestellt (s. Abschnitt 3). Mit Blick auf individualisierten Unterricht haben 78 % der Schülerinnen und Schüler eine Lehrkraft, die ihnen in den meisten oder in (fast) jeder Stunde etwas so lange erklärt, bis es alle verstanden haben. Individualisierte Aufgaben im engeren Sinne werden demgegenüber viel seltener eingesetzt: Hier sind es 38 % der Schülerinnen und Schüler, die eine Lehrperson haben, die ihnen je nach Können unterschiedliche Aufgaben in den meisten oder in (fast) jeder Stunde gibt. Zu beachten ist hier, dass die Fragestellung sowohl Deutsch als auch

Mathematik umfasst. Gerade die Aussage, dass Lehrpersonen Schülerinnen und Schüler etwas so lange erklären, bis es alle verstanden haben, scheint eher ein für den Mathematikunterricht passendes Item zu sein als für den Leseunterricht.

4.5 Zusammenhang zwischen den Unterrichtsformen und der Lesekompetenz

Tabelle 5 zeigt die Zusammenhangskoeffizienten zwischen den verschiedenen Unterrichtsformen und der Lesekompetenz für die vier Erhebungszeitpunkte von PIRLS für Österreich. Im Allgemeinen gilt, dass je höher der Zusammenhangskoeffizient ist, desto höher ist der Zusammenhang zwischen der jeweiligen Unterrichtsform und der Lesekompetenz (wobei Zusammenhang nicht Kausalität bedeutet).

Tabelle 5: Korrelationskoeffizienten zwischen Unterrichtsformen und Lesekompetenz für Österreich.
Quellen: PIRLS 2006, 2011, 2016 und 2021. Berechnungen durch die Autorinnen.

PIRLS	2006		2011		2016		2021	
Unterrichtsformen	<i>r</i>	(SE)	<i>r</i>	(SE)	<i>r</i>	(SE)	<i>r</i>	(SE)
Ich unterrichte die ganze Klasse gemeinsam in Lesen.	-.05	(-.03)	-.02	(0,02)	-.05	(0,03)	-.06	(0,03)
Ich bilde Gruppen von Schülerinnen und Schülern mit gleichem Fähigkeitsniveau.	.00	(0,03)	.03	(0,03)	.02	(0,03)	-.01	(0,03)
Ich bilde Gruppen von Schülerinnen und Schülern mit unterschiedlichem Fähigkeitsniveau.	.05	(0,04)	.02	(0,02)	.03	(0,03)	.00	(0,02)
Ich gebe individuell abgestimmten Leseunterricht.	.08	(0,03)	.02	(0,03)	.02	(0,03)	.04	(0,02)
Die Schülerinnen und Schüler arbeiten selbstständig nach einem festgelegten Plan oder einer Zielvorgabe.	.05	(0,03)	.03	(0,02)	.08	(0,03)	.10	(0,03)

Wie bereits eingangs angedeutet, zeigen sich keine nennenswerten Korrelationen ($r \geq .10$; Cohen, 1988) zwischen den Unterrichtsformen und der Lesekompetenz.

Tabelle 6 zeigt die Zusammenhangskoeffizienten zwischen den verschiedenen Unterrichtsformen und der Lesekompetenz für die Baseline 2010 D4/M4.

Tabelle 6: Korrelationskoeffizienten zwischen Unterrichtsformen und Lesekompetenz anhand Baseline D4/M4. Quelle: Baseline 2010 D4/M4. Berechnungen durch die Autorinnen.

Baseline D4/M4	2010	
Unterrichtsformen	<i>r</i>	(SE)
Ich unterrichte die Schülerinnen und Schüler in der Gesamtgruppe.	.00	(0,02)
Ich gebe den Schülerinnen und Schülern je nach Können unterschiedliche Aufgaben.	.04	(0,02)
Ich bilde Gruppen von Schülerinnen und Schülern mit gleichen Fähigkeitsniveaus und gebe nur Schülerinnen und Schülern in einer Gruppe dieselben Aufgaben.	-.01	(0,02)
Ich bilde Gruppen von Schülerinnen und Schülern mit unterschiedlichen Fähigkeitsniveaus und gebe allen Gruppen dieselben Aufgaben.	-.05	(0,02)
Ich erkläre den einzelnen Schülerinnen und Schüler etwas so lange, bis es alle verstanden haben.	.03	(0,02)

Auch für die Baseline-2010-Daten ergeben sich keine nennenswerten Zusammenhänge ($r \geq .10$) zwischen den fünf Unterrichtsformen und der Lesekompetenz.

Da die Frageformulierung der Baseline Deutsch und Mathematik enthält, wird probehalber auch ein Zusammenhang zwischen den anderen drei Kompetenzbereichen (Rechtschreiben, Verfassen von Texten, Textproduktion) und den Unterrichtsformen sowie zwischen Mathematikkompetenz und den Unterrichtsformen geprüft. Keine dieser Analysen ergibt ein nennenswertes Ergebnis in Bezug auf die Höhe des Zusammenhangs (ohne Tabelle).

5 Gesamtdiskussion

5.1 Resümee der Resultate

Der vorliegende Beitrag betrachtet Lehrerangaben (auf Schülerinnen- und Schülerebene zur Individualisierung und Binnendifferenzierung des Leseunterrichts zu insgesamt fünf Erhebungszeitpunkten und von zwei Studien – PIRLS mit vier Erhebungszeitpunkten und die Baseline-Testung in Österreich mit einem Erhebungszeitpunkt. Dadurch können insgesamt 15 Jahre Leseunterricht in den Blick genommen werden. Österreich wird dabei mit ausgewählten EU-Ländern verglichen. Dabei zeigt sich Folgendes:

5.1.1 Forschungsfrage 1

„Unterscheidet sich Österreich in Bezug auf Binnendifferenzierung und Individualisierung von anderen Ländern?“

Österreichische Lehrpersonen unterrichten weniger häufig die ganze Klasse gemeinsam in Lesen als in ausgewählten Vergleichsländern. Dennoch liegt die Häufigkeit von

individualisiertem Leseunterricht sowie von Binnendifferenzierung – sowohl in Gruppen mit gleichem wie auch mit unterschiedlichem Fähigkeitsniveau – laut Lehrerangaben und im Vergleich unter dem Durchschnitt der ausgewählten EU-Länder. Welche Formen des Leseunterrichts an Österreichs Volksschulen angesichts der eher niedrigen Werte beim gemeinsamen Unterricht sowie den erfassten Formen der Individualisierung und Binnendifferenzierung vermehrt eingesetzt werden, muss an dieser Stelle offenbleiben.

Bei der schülerzentrierten Variante von Individualisierung nach einem festgelegten Plan oder einer Zielvorgabe liegt Österreich laut Lehrerangaben über dem Durchschnitt der betrachteten Länder. Diese Form der Individualisierung dürfte in Österreich über die Jahre hinweg weit verbreitet sein. Dieses Bild bestätigt sich auch in den Ergebnissen für Österreich im Überblick, für die sowohl PIRLS- als auch Baseline-Daten vorliegen: Am häufigsten erfolgt Individualisierung nach Plan bzw. einer Zielvorgabe in Österreichs Leseunterricht (basiert nur auf PIRLS-Daten), gefolgt vom Gesamtunterricht mit der ganzen Klasse. Dahinter liegen der individualisierte Unterricht sowie die beiden diskutierten Formen der Binnendifferenzierung. Diese Reihung ist zwischen 2006 und 2021 gleich geblieben.

Ein detaillierterer Blick auf individualisierten Unterricht, der vorwiegend von der Lehrperson gesteuert ist, zeigt, dass Österreich bei der Häufigkeit des Verknüpfens von Vorkenntnissen mit neuen Inhalten sowie dem Geben individueller Rückmeldung unter dem Durchschnitt der Vergleichsländer liegt. An das Leseniveau angepasste Materialien werden in Österreich etwas häufiger bereitgestellt als im Durchschnitt der Länder. Leider geht aus der Fragestellung nicht hervor, wie das Leseniveau vorab bestimmt wurde. Diagnostischen Lesetests zur Erfassung des Lesefortschritts wird in Österreich allerdings wenig Gewicht beigemessen und seit 2006 nimmt dieses Gewicht statistisch signifikant weiter ab (s. dazu auch Österbauer, Illetschko & Mödlhamer, 2023).

5.1.2 Forschungsfrage 2

„Hat sich der Einsatz von Binnendifferenzierung und Individualisierung im Laufe der Zeit verändert?“

In den zehn Jahren zwischen 2006 und 2016 ist der gemeinsame Unterricht der Klasse in Lesen rückläufig, während individuell abgestimmter Leseunterricht sowie die beiden Formen der Binnendifferenzierung zunehmen. Für das Jahr 2021 zeigt sich mehr als zuvor, dass Schülerinnen und Schüler gemeinsam in der Klasse in Lesen unterrichtet werden, während individuell abgestimmter Leseunterricht sowie Binnendifferenzierung in gleich guten Gruppen weniger häufig als davor praktiziert werden. Der Rückgang der Binnendifferenzierung und des individualisierten Unterrichts im Jahr 2021 kann eventuell auf notwendige Veränderungen in der Schulorganisation aufgrund der Pandemie zurückgeführt werden.

5.1.3 Forschungsfrage 3

„Gibt es einen Zusammenhang zwischen jeweils Unterricht mit der gesamten Klasse, Binnendifferenzierung bzw. Individualisierung und Lesekompetenz?“

Eine Analyse der Zusammenhänge für Österreich zwischen dem Ausmaß an Unterricht mit der gesamten Klasse, an Binnendifferenzierung und an Individualisierung mit jeweils Lesekompetenz ergibt keine nennenswerten Korrelationen. Dieses Ergebnis ist laut Forschungsstand (s. Abschnitt 3) durchaus erwartungskonform.

5.2 Methodische Limitationen

Laut Schulz-Heidorf (2017) ist in der Theorie, der Empirie und der Praxis das Verständnis dessen, was unter individueller Förderung und der Individualisierung des Unterrichts zu verstehen ist, äußerst uneinheitlich. Dies kann für die Erfassung des Konstrukts vor allem für *large-scale assessments* Folgen haben: Eventuell interpretieren die befragten Lehrpersonen ihre angewendete Art der Individualisierung und deren Häufigkeit nicht als Individualisierung und unterschätzen systematisch die anhand von Einzelitems erfasste Häufigkeit von Individualisierung (Schulz-Heidorf, 2017, S. 21).

Die diesem Beitrag zugrunde liegenden Fragebögen werden wie erwähnt nicht eigens für diese Untersuchung erstellt, sondern es wird auf bestehende Fragebögen aus nationalen und internationalen Erhebungen zurückgegriffen. Das ermöglicht einerseits die Betrachtung im Zeitverlauf, hat allerdings auch den Nachteil, dass erstens nur Ausschnitte der drei in Abschnitt 2.1 und 2.2 dargestellten Formen von Binnendifferenzierung bzw. Individualisierung erfasst werden können und zweitens bei manchen Fragestellungen der Interpretationsspielraum groß ist und nicht mit Sicherheit festgestellt werden kann, wie diese von den befragten Lehrpersonen verstanden worden sind. Zunächst zum ersten Punkt: Es wird in den Fragebögen beispielsweise nicht näher spezifiziert, welche Form individuellen Unterrichts gegeben ist, welche Bereiche dieser umfasst und wie dieser organisiert ist. Blickt man weiters auf die Einzelvariablen in Tabelle 2, so fällt auf, dass der gesamte Bereich der „individuellen Anpassung der Begleitung beim Lernen“ (ad 4. in Abschnitt 2) fehlt. Aus dem ersten Bereich „individuelle Anpassung von Zielen, Inhalten sowie Kompetenzen“ kann nur das Anknüpfen an Vorkenntnisse erfasst werden. Die Anpassung an Ziele oder Inhalte im engeren Sinne ist damit nicht abgedeckt. Beim Anknüpfen an Vorkenntnisse wird zudem nicht explizit angegeben, dass es um die individuellen Vorkenntnisse gehen soll. Die Lehrpersonen könnten die Frage daher auch auf den bisherigen Lernstand der Klasse laut Unterricht mit der gesamten Klasse beziehen. Dies ist einschränkend zu berücksichtigen. Zum zweiten Punkt ist besonders anzumerken, dass gerade bei der für Österreich am häufigsten genannten schülerzentrierten Variante von Individualisierung – dem selbstständigen Arbeiten nach einem festgelegten Plan oder einer Zielvorgabe – nicht eindeutig aus der Fragestellung hervorgeht, dass die Schülerinnen und Schüler aus verschiedenen Lernangeboten das passende für sich auswählen. Auch liegen keine Daten dazu vor, wie die Lehrpersonen zu einer Bestimmung des Leseniveaus kommen.

Die Befundlage zur Wirksamkeit von Individualisierung und Binnendifferenzierung ist, wie in Abschnitt 2.3 diskutiert, nicht eindeutig. Studien wie PIRLS oder die Baseline der Bildungsstandardüberprüfung können keine Gelingensbedingungen für den Ein-

satz von Maßnahmen zur Individualisierung und Differenzierung erfassen. Mit den vorliegenden Daten und Fragebögen ist daher die Qualität von Individualisierung und Binnendifferenzierung im Leseunterricht nicht erfassbar. Die Frage, wie Individualisierung im Leseunterricht konkret erfolgt, kann anhand der vorliegenden Daten nicht beantwortet werden.

Für die Baseline-Daten sei abschließend noch einmal darauf hingewiesen, dass sich die Daten auf Deutsch und Mathematik gleichermaßen beziehen und dadurch eine Unschärfe aufweisen. Zu beachten ist, dass die Art und das Ausmaß von Individualisierung und Binnendifferenzierung nach Fächern unterschiedlich gestaltet sein kann.

5.3 Diskussion und Implikationen

Folgende Potenziale für die Individualisierung des Leseunterrichts auf der vierten Schulstufe in Österreichs Volksschulen können durch einen Langzeitblick auf mehrere Erhebungszeitpunkte aus insgesamt zwei Studien identifiziert werden:

Aus Längsschnittstudien (z. B. Landerl & Wimmer, 2008) ist bekannt, dass Leseprobleme von leistungsschwachen Kindern am Ende der ersten Schulstufe über die Primarstufe hinaus eher stabil bestehen bleiben. Daher müssen leistungsschwache Schülerinnen und Schüler so früh wie möglich identifiziert und mit individuell angepassten Hilfsmaßnahmen unterstützt werden (Gasteiger-Klicpera, 2020). Individualisierung und Binnendifferenzierung sind Konzepte, die dabei helfen können, eine passgenaue Förderung für Schülerinnen und Schüler mit unterschiedlichen Vorkenntnissen und Lernvoraussetzungen zu schaffen.

Helmke (2015, S. 252) zählt das Vorwissen neben dem Migrationshintergrund, dem Entwicklungsstand und dem Lernstil zu den zentralen Lernermerkmalen mit Blick auf Heterogenität. Das Vorwissen ist für ihn dabei das „mit Abstand wichtigste Lernermerkmal“ (Helmke, 2015, S. 252). Eine noch stärkere Berücksichtigung des Vorwissens im Leseunterricht, wie dies etwa in Bulgarien, Litauen oder Polen bereits der Fall ist, dürfte daher gewinnbringend sein. Bei der Bestimmung der Vorkenntnisse und der Einschätzung des Lesefortschritts für eine regelmäßige Kompetenzeinschätzung und einer darauf aufbauenden Individualförderung liegt im Einsatz diagnostischer Lesetests noch erhebliches Potenzial (Österbauer et al., 2023). Dafür ist zentral, dass Lehrpersonen der Nutzen von diagnostischen Tests bewusst ist und sie diese als nützlich für ihre Förderplanung wahrnehmen. Dies ist insbesondere für passgenaue individuelle Förderung zentral. Denn nur, wenn der individuelle Lernstand genau erfasst werden kann, können entsprechende Fördermaßnahmen gesetzt werden. In der Forschungsliteratur wird daher „[...] eine regelmäßige und systematische Diagnostik von Lernvoraussetzungen ebenso als fester Bestandteil des individualisierten Unterrichts angesehen [...] wie die Ausrichtung der Lernangebote an den Ergebnissen dieser Diagnostik [...]“ (Fladung, 2022, S. 87). Eine systematische Diagnose ist zentral dafür, dass individualisierte Unterrichtsformen ihre Wirkung entfalten können (Fladung, 2022, S. 88). Zwar können zur Diagnose auch nichtstandardisierte Verfahren gewinnbringend eingesetzt werden, doch „Individualisierung erfordert [auch] den regel-

mäßigen Einsatz kompetenz- und standardorientierter Tests, um die Adaptivität des Unterrichts zu erhöhen [...]“ (Helmke, 2015, S. 257). Dass eine genaue Einschätzung des Lernstands – vor allem ohne geeignete diagnostische Instrumente – alles andere als trivial ist, zeigt auch Fladung (2022). Hilfreich dabei kann eine Orientierung an der evidenzbasierten Förderung sein: Zunächst werden die vorhandenen Kompetenzen diagnostiziert (Eingangsdiagnostik), anschließend erfolgt eine darauf aufbauende Förderung und schließlich wird der Lernfortschritt wiederum mithilfe einer Wiederholung der eingangs durchgeführten Diagnoseinstrumente eingeschätzt. Die individuelle Kompetenzmessung PLUS (iKM^{PLUS}) ist dafür ein mögliches Instrument, da es die Messung des Lesefortschritts von der dritten zur vierten Schulstufe beinhaltet (Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen [IQS], 2023). Die so gewonnenen Erkenntnisse können in weiterer Folge auch als Grundlage für die Bildung von Gruppen für die Binnendifferenzierung genutzt werden (Helmke, 2015, S. 252). Diagnostische Kompetenz bei Lehrpersonen ist also eine wichtige Voraussetzung für gelingende Individualisierung und Binnendifferenzierung (Altrichter et al., 2009, S. 351; Helmke, 2015, S. 257). Auf Individualisierung, die vorwiegend von der Lehrperson ausgeht (s. Abschnitt 2) und auf der Diagnose der Lernvoraussetzungen aufbaut, wird im Lehrplan in der Fassung von 2012 nicht explizit eingegangen (BGBl. I Nr. 303/2012). Im neuen Lehrplan der Volksschule, der seit dem Schuljahr 2023/24 rollierend und beginnend mit der ersten Schulstufe eingeführt wird, wird dem Rechnung getragen. Grundsatz 1 lautet:

Lehrerinnen und Lehrer nehmen Schülerinnen und Schüler individuell wahr und ermöglichen individuelle Lernprozesse. Schülerinnen und Schüler bringen unterschiedliches Vorwissen, verschiedene sprachliche Vorkenntnisse, vielfältige und unterschiedliche Vorerfahrungen, Interessen und Lernpräferenzen mit. An diese Unterschiede muss im Unterricht angeschlossen werden, um sie für die Stärkung der individuellen Lernmotivation und Leistungsfähigkeit nutzbar zu machen. Lehrerinnen und Lehrer verstehen es als ihre Aufgabe, Schülerinnen und Schüler individuell wahrzunehmen und zu fördern und vermeiden stereotype Zu- und Festschreibungen. Lehrerinnen und Lehrer kennen und nutzen geeignete pädagogische Diagnoseinstrumente, um die Lernvoraussetzungen der Schülerinnen und Schüler festzustellen und deren Lernprozesse entsprechend begleiten zu können. Sie fördern individuelle Lernprozesse durch unterschiedliche und abwechslungsreiche Lernsettings und verwenden dazu passende Lernmaterialien. Sie geben individuelle, lernförderliche Rückmeldungen und ermöglichen den Schülerinnen und Schülern, ihren Kompetenzzuwachs bewusst wahrzunehmen (BGBl. II Nr. 1/2023).

Wie unter anderem Helmke (2015) und Altrichter et al. (2009, S. 357–360) hervorheben, geht wirksame Individualisierung und Binnendifferenzierung mit einem erheblichen Aufwand aufseiten der Lehrpersonen, aber auch anderer Akteurinnen und Akteure einher. Diese bedürfen nicht nur einer entsprechenden Ausbildung und Fortbildung, Professionalisierung und diagnostischen Kompetenz, sondern auch entsprechender Lehrmaterialien und finanzieller Ressourcen. Denn für eine gelungene Umsetzung ist – wie sich anhand der Zusammenhangsmaße mit der Leseleistung ablesen lässt – nicht die Tatsache, dass differenziert wird, entscheidend, sondern allen voran die Art und Weise, wie Individualisierung und Binnendifferenzierung umgesetzt werden. Gerade in diesem Bereich

erscheint nicht nur das von Altrichter et al. (2009) formulierte Desiderat nach vertiefenden Studien nach wie vor aktuell. Es bedarf darüber hinaus schnell einsetzbarer und günstiger Diagnoseinstrumente sowie Materialien zur Differenzierung, auf die Lehrpersonen unkompliziert zugreifen können. Neben einer Erhöhung des bereits wachsenden Angebots ist jedoch auch eine Steigerung der Nachfrage nach solchen Instrumenten und Materialien wichtig.

5.4 Ausblick

Wie wichtig die Lesekompetenz von Kindern und Jugendlichen ist und welche Forschungsdesiderata konkret getroffen werden sollten, zeigt der Beitrag von Bachinger, Bruneforth und Schmich (2021) im Nationalen Bildungsbericht 2021 auf. Mit der Initiative „LESEN – Schwerpunkt ab dem Schuljahr 2023/2024“ hat das österreichische Bundesministerium für Bildung, Wissenschaft und Forschung eine richtungsgebende Maßnahme eingeleitet, um die Lesekompetenz der Schülerinnen und Schüler nachhaltig zu verbessern (Bundesministerium für Bildung, Wissenschaft und Forschung, 2023; Fink & Kroisleitner, 2023).

Literatur

- Altrichter, H., Trautmann, M., Wischer, B., Sommerauer, S. & Doppler, B. (2009). Unterrichten in heterogenen Gruppen: Das Qualitätspotenzial von Individualisierung, Differenzierung und Klassenschülerzahl. In W. Specht (Hrsg.), *Nationaler Bildungsbericht Österreich 2009, Band 2: Fokussierte Analysen bildungspolitischer Schwerpunktthemen* (S. 341–360). Graz: Leykam.
- Bachinger, A., Bruneforth, M. & Schmich, J. (2021). Lesekompetenz und Leseunterricht in Österreich – Ergebnisse, Entwicklungen und Forschungsinteressen aus fachdidaktischer und aus bildungspolitischer Perspektive. In BMBWF (Hrsg.), *Nationaler Bildungsbericht Österreich 2021. Teil 3 – Ausgewählte Entwicklungsfelder* (S. 471–529). Wien: BMBWF. doi:10.17888/nbb2021-03
- BIFIE/IQS (2021). B410I_MASTER_RAW_2_0_Almanach.rtf, *Standardüberprüfung Ausgangsmessung, 4. Schulstufe (2010), Schülerebene (B410I) (V2)*. doi:10.14472/IQSULSAN2/WYIQRA
- Bohl, T., Batzel, A. & Richey, P. (2012). Öffnung – Differenzierung – Individualisierung – Adaptivität. Charakteristika, didaktische Implikationen und Forschungsbefunde verwandter Unterrichtskonzepte zum Umgang mit Heterogenität. In T. Bohl, M. Bönsch, M. Trautmann & B. Wischer (Hrsg.), *Binnendifferenzierung. Teil 1: Didaktische Grundlagen und Forschungsergebnisse zur Binnendifferenzierung im Unterricht* (Bd. 17, S. 40–69). Immenhausen bei Kassel: Prolog-Verlag.

- Bönsch, M. (1995). *Differenzierung in Schule und Unterricht. Ansprüche, Formen, Strategien*. München: Ehrenwirth.
- Bönsch, M. (2009). Methodik der Differenzierung. Ordnung und Umsetzungsmöglichkeit von Differenzierungsformen. *Pädagogik*, 61 (9), 36–40.
- Bönsch, M. (2012). Strategien zur Lernprozessoptimierung – Innere Differenzierung. In T. Bohl, M. Bönsch, M. Trautmann & B. Wischer (Hrsg.), *Binnendifferenzierung. Teil 1: Didaktische Grundlagen und Forschungsergebnisse zur Binnendifferenzierung im Unterricht* (S. 9–23). Immenhausen bei Kassel: Prolog-Verlag.
- Breit, S., Bruneforth, M. & Schreiner, C. (Hrsg.). (2016). *Standardüberprüfung 2015. Deutsch, 4. Schulstufe. Bundesergebnisbericht*. Salzburg: Bundesinstitut für Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens. Verfügbar unter: <https://www.iqs.gv.at/downloads/archiv-des-bifie/bildungsstandardueberpruefungen/ergebnisberichte>
- Bundesministerium für Bildung, Wissenschaft und Forschung (BMBWF). (2023). *LESEN – Schwerpunkt ab dem Schuljahr 2023/24*. Verfügbar unter: <https://www.bmbwf.gv.at/Themen/schule/zrp/lesen.html>
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd edition). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Fink, A. G. & Kroisleitner, O. (2023, 4. September). Polaschek will Lesekompetenz stärken. *Der Standard*. Verfügbar unter: <https://www.derstandard.at/story/3000000185464/bildungsminister-polaschek-kuendigt-massnahmen-fuer-bessere-lesekompetenz-an>
- Fishbein, B., Liquin, Y. & Foy, P. (Hrsg.). (2024). *PIRLS 2021 User Guide for the International Database* (2nd edition). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Verfügbar unter: https://pirls2021.org/data/downloads/P21_UG_International-Database.pdf
- Fladung, I. T. (2022). *Deutschunterricht im Vorbereitungsdienst adaptiv planen. Eine empirische Studie zum Stellenwert von Diagnostik und Differenzierung in schriftlichen Unterrichtsplanungen*. Münster: Waxmann.
- Gasteiger-Klicpera, B. (2020). Diversität in der Entwicklung des Lesens. In L. Paleczek & S. Seifert (Hrsg.), *Inklusiver Leseunterricht: Leseentwicklung, Diagnostik und Konzepte* (S. 3–21). Wiesbaden: Springer Fachmedien Wiesbaden. doi:10.1007/978-3-658-24221-3_1
- Hattie, J. A. C. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. London, New York: Routledge.
- Helmke, A. (2013). Individualisierung: Hintergrund, Missverständnisse, Perspektiven. *Pädagogik* (2), 34–37.
- Helmke, A. (2015). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts* (6. Auflage). Seelze-Velber: Kallmeyer.
- Hess, M. & Lipowsky, F. (2017). Lernen individualisieren und Unterrichtsqualität verbessern. In F. Heinzel & K. Koch (Hrsg.), *Individualisierung im Grundschulunterricht. Anspruch, Realisierung und Risiken* (S. 23–31). Wiesbaden: Springer.
- Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen (IQS). (2023). *Individuelle Kompetenzmessung PLUS (iKM^{PLUS})*. Verfügbar unter: <https://iqs.gv.at/themen/nationale-kompetenzerhebung/ikm-plus>

- Kaźmierczak, J. (2022). Poland. In K. A. Reynolds, E. Wry, I. V. S. Mullis & M. von Davier (Hrsg.), *PIRLS 2021 Encyclopedia: Education Policy and Curriculum in Reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Verfügbar unter: <https://pirls2021.org/encyclopedia>
- Kellermann, F. (2017, 04. September). *Polnisches Schulsystem. Rückkehr zum autoritären Unterricht?* Deutschlandfunk.
- Klafki, W. (1994). *Neue Studien zur Bildungstheorie und Didaktik. Zeitgemäße Allgemeinbildung und kritisch-konstruktive Didaktik*. Weinheim, Basel: Beltz.
- Kopas, G. (2022). Slovak Republic. PIRLS 2021 Encyclopedia. In K. A. Reynolds, E. Wry, I. V. S. Mullis & M. von Davier (Hrsg.), *PIRLS 2021 Encyclopedia: Education Policy and Curriculum in Reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Verfügbar unter: <https://pirls2021.org/slovak-republic>
- Landerl, K. & Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *Journal of Educational Psychology*, 100 (1), 150–161.
- Mavrodieva, M. & Damyanova, S. (2022). Bulgaria. PIRLS 2021 Encyclopedia. In K. A. Reynolds, E. Wry, I. V. S. Mullis & M. von Davier (Hrsg.), *PIRLS 2021 Encyclopedia: Education Policy and Curriculum in Reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Verfügbar unter: <https://pirls2021.org/bulgaria/>
- Mielach, E. (2023). Teilnahmequoten im Haupttest von PIRLS 2021. In J. Schmich (Hrsg.), *PIRLS 2021. Technischer Bericht*. Salzburg: Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen (IQS). doi:10.17888/pirls2021-tb
- Mullis, I. V. S., von Davier, M., Foy, P., Fishbein, B., Reynolds, K. A. & Wry, E. (Hrsg.). (2023). *PIRLS 2021 International Results in Reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Verfügbar unter: <https://pirls2021.org/results>
- Neubacher, M., Ober, M., Wimmer, C. & Hartl, M. (2019). Die Kompetenzen der Schüler/innen in der Zusammenschau aller Standardüberprüfungen des ersten Zyklus. In A. C. George, C. Schreiner, C. Wiesner, M. Pointinger & K. Pacher (Hrsg.), *Fünf Jahre flächendeckende Bildungsstandardüberprüfungen in Österreich. Vertiefende Analysen zum Zyklus 2012 bis 2016* (S. 55–85). Münster, New York: Waxmann. Verfügbar unter: <https://waxmann.com/buch3925>
- OECD. (2021). *The state of school education: One year into the COVID pandemic*. Paris: OECD Publishing. doi:10.1787/201dde84-en
- Österbauer, V., Illetschko, M. & Mödlhamer, C. (2023). Die Eignung von diagnostischen Instrumenten für den Leseunterricht aus Sicht der Lehrpersonen. In J. Schmich, C. Wallner-Paschon & M. Illetschko (Hrsg.), *PIRLS 2021. Die Lesekompetenz am Ende der Volksschule. Erste Ergebnisse* (S. 58–61). Salzburg: Institut für Qualitätssicherung im österreichischen Schulwesen. doi:10.17888/pirls2021-eb.2
- Schrader, F.-W. & Helmke, A. (2008). Unter der Lupe: Wie in deutschen Klassenzimmern differenziert wird. *Klexer. Magazin für die Grundschule* (20), 3–5.
- Schreiner, C. & Wiesner, C. (2019). Die Überprüfung der Bildungsstandards in Österreich: der erste Zyklus als Meilenstein für die Schul- und Unterrichtsentwicklung – eine gelungene Innovation im österreichischen Schulsystem. In A. C. George, C. Schreiner, C. Wiesner, M. Pointinger & K. Pacher (Hrsg.), *Fünf Jahre flächendeckende Bildungs-*

standardüberprüfungen in Österreich. Vertiefende Analysen zum Zyklus 2012 bis 2016 (S. 13–53). Münster, New York: Waxmann. Verfügbar unter: <https://waxmann.com/buch3925>

Schulz-Heidorf, K. (2017). Die Erfassung individualisierten Unterrichts in Large-Scale-Assessments. Herausforderungen und Perspektiven am Beispiel von IGLU. *Tertium comparationis*, 23 (1), 9–27.

Von der Groeben, A. & Kaiser, I. (2011). Entwicklungsziel Individualisierung. *Pädagogik*, 63 (12), 42–47.

Autorinnen und Autoren

em. o. Univ.-Prof. Dr. Herbert Altrichter

Tanja Auer, MA

Univ.-Prof. i. R. Dr. Johann Bacher

Michael Bruneforth, MA

Carina Gargitter, BA MSc

Priv.-Doz.ⁱⁿ Dr.ⁱⁿ Dipl.-Stat. Ann Cathrice George

Univ.-Prof. Dr. Burkhard Gniewosz

Prof.ⁱⁿ Dr.ⁱⁿ habil. Jana Groß Ophoff

Univ.-Prof. Dr. Christoph Helm

Univ.-Prof.ⁱⁿ Dr.ⁱⁿ Nele Kampa

HS.-Prof. Mag. Dr. David Kemethofer

Mag.^a Katharina Meusburger

MMag.^a Dr.ⁱⁿ Veronika Österbauer

Martina Ott, PhD MA

Assoz. Prof.ⁱⁿ Dr.ⁱⁿ Marion Reindl

Mag.^a Dr.ⁱⁿ Silvia Salchegger

Mag.^a Dr.ⁱⁿ Juliane Schmich

Mag.^a Christina Wallner-Paschon

Dr. Christoph Weber

Mag.^a Lisa Wiesinger, BA

Ramona Zintl, BSc MSc MSc

Johannes Kepler Universität Linz

Paris Lodron Universität Salzburg

Johannes Kepler Universität Linz

IQS

IQS

IQS

Paris Lodron Universität Salzburg

PH Vorarlberg

Johannes Kepler Universität Linz

Universität Wien

PH Oberösterreich

PH Vorarlberg

IQS

PH Vorarlberg

Paris Lodron Universität Salzburg

IQS

IQS

IQS

PH Oberösterreich

IQS

IQS