Unterrichtswiss
https://doi.org/10.1007/s42010-021-00132-9

ALLGEMEINER TEIL

Leistungsentwicklung in jahrgangsgemischten und jahrgangshomogenen dritten und vierten Klassen

Meike Munser-Kiefer · Sabine Martschinke · Alfred Lindl · Andreas Hartinger

Eingegangen: 4. Januar 2021 / Überarbeitet: 3. August 2021 / Angenommen: 23. September 2021

Zusammenfassung In einer quasi-experimentellen Längsschnittstudie wurde die Leistungsentwicklung jahrgangsgemischt und jahrgangshomogen unterrichteter Schüler*innen des dritten und vierten Schuljahres ($N = 1644$) aus 125 Klassen ($n_{jahrgangsgemischt} = 68$, $n_{jahrgangshomogen} = 57$) zu drei Messzeitpunkten miteinander verglichen. Die Ergebnisse der gematchten Gesamtstichprobe zeigen – bei vergleichbaren Ausgangswerten zu Beginn der dritten Jahrgangstufe – am Ende der vierten Jahrgangsstufe keine Unterschiede. Die Leistungsentwicklung bis zum Ende der dritten Jahrgangsstufe belegt hingegen insgesamt einen kleinen Effekt zugunsten jahrgangsgemischter Klassen. Zusätzlich werden differenzielle Effekte für verschiedene Leistungsgruppen geprüft: Hier sind am Ende der dritten Klasse signifikante kleine bis mittlere positive Effekte der Jahrgangsmischung in den oberen drei Leistungsquartilen zu erkennen. Am Ende der vierten Klasse finden sich nur für Kinder aus dem untersten Leistungsquartil tendenziell Vorteile durch die Jahrgangsmischung. Die Studie weist somit auf die Bedeutung differenzieller Effekte in Abhängigkeit von der Altersgruppe bzw. vom Leistungsstand hin, die dahinterliegende Änderungen in der Tiefenstruktur von Unterricht vermuten lassen und Anregungen für die Weiterentwicklung des jahrgangsgemischten Unterrichts geben können.

Meike Munser-Kiefer (✉)
Professur für Pädagogik (Grundschulpädagogik), Institut für Bildungswissenschaft, Universität Regensburg, Regensburg, Deutschland
E-Mail: meike.munser-kiefer@ur.de

Sabine Martschinke
Lehrstuhl für Grundschulpädagogik und -didaktik mit dem Schwerpunkt Umgang mit Heterogenität, Institut für Grundschulforschung Nürnberg, Friedrich-Alexander-Universität, Nürnberg, Deutschland

Alfred Lindl
Methoden der empirischen Bildungsforschung, Institut für Bildungswissenschaft, Universität Regensburg, Regensburg, Deutschland

Andreas Hartinger
Lehrstuhl für Grundschulpädagogik und -didaktik, Universität Augsburg, Augsburg, Deutschland

Published online: 02 November 2021

Schlüsselwörter Jahrgangsgemischter Unterricht · Jahrgangsmischung · Grundschule · Leistungsentwicklung · Differenzielle Effekte

Development of performance in multi-grade and mono-grade classes in third and fourth grades

Abstract This paper presents a quasi-experimental longitudinal study about the performance development of multi-grade and mono-grade students in third and fourth grade ($n = 1644$, 68 multi-grade classes and 57 mono-grade classes) at three measurement times. The results of the matched sample show with comparable starting values at the beginning of the third grade no differences at the end of the fourth grade. The performance development up to the end of the third grade, however, shows a small overall effect in favor of multi-grade classes. In addition, differential effects are examined for different ability groups: At the end of the third grade, positive significant small or medium effects of the multi-grade classes can be seen in the top three performance quartiles. At the end of the fourth grade, however, only children from the lowest performance quartile tend to benefit from the multi-grade composition. The study points to the importance of differential effects depending on age or level of proficiency, which suggest underlying changes in the deep structure of lessons und provides suggestions for further development of multi-grade teaching.

Keywords Multi-grade · Multi-age · Stage classes · Primary school · Learning development · Differential effects

Jahrgangsgemischtes Lernen hat historisch in nationalen und internationalen Kontexten eine lange Tradition und spielt gegenwärtig und vermutlich auch künftig besonders in der Grundschule eine Rolle. Jahrgangsgemischte Klassen werden teils aus pragmatischen Gründen gebildet, wenn zu wenige Schüler*innen einer Klassenstufe an einem Ort zusammenkommen; teils stehen pädagogisch-didaktische Gründe im Vordergrund, wenn die bewusst gespreizte Heterogenität das Lernen der Schüler*innen bereichern und verbessern soll (z. B. durch Tutorensysteme, Lernen durch Lehren; individuelle Verweildauer, differenziertes Lernangebot). Mit der „neuen Schuleingangsstufe“ hat das jahrgangsgemischte Lernen in Deutschland seit den 1990er-Jahren auch im Regelschulbetrieb eine Renaissance erfahren (Sonnleitner 2021). Die Anzahl der jahrgangsgemischten Klassen im deutschsprachigen Raum steigt seither an, mittlerweile nicht nur für die Eingangsklassen, sondern auch für die dritte und vierte Jahrgangsstufe (z. B. in Bayern: Schuljahr 2013/2014: 228, Schuljahr 2016/2017: 351, Schuljahr 2019/2020: 465). Allerdings fehlt gerade für die höheren Jahrgangsstufen die empirische Evidenz zur Wirksamkeit, insbesondere auf Basis repräsentativer Längsschnittstudien mit Kontrollgruppe.

Die vorliegende Untersuchung bearbeitet dieses Desiderat für die dritte und vierte Jahrgangsstufe: Sie prüft Unterschiede in der Leistungsentwicklung zwischen jahrgangsgemischten und jahrgangshomogenen Klassen und beleuchtet die Effekte jahrgangsgemischten Lernens – differenziert nach Leistungsgruppen – an einem

bedeutsamen Übergang im deutschen Bildungswesen, der selektiv über die Bildungschancen in den weiterführenden Schulen entscheidet (Ditton 2019).

1 Lernen in jahrgangsgemischten Klassen

Die Jahrgangsmischung ist ein Merkmal der Oberflächenstruktur. Sie kann sich jedoch über die alters- bzw. schulstufenbezogene Heterogenisierung auf die curriculare Anordnung der Inhalte und die Interaktion zwischen Lehrenden und Lernenden bzw. der Lernenden untereinander auswirken und so zu Änderungen in der Tiefenstruktur des Unterrichts führen (Decristan et al. 2020; Hahn 2019). Diese Tiefenstrukturen lassen sich durch Merkmalsdimensionen der Unterrichtsqualität beschreiben wie Klassenführung, effektive Lernzeitnutzung, Inhaltsauswahl, individuelle Unterstützung, kognitive Aktivierung und Konsolidierung (Praetorius und Charalambous 2018; Wisniewski et al. 2020). Diese sind für eher geschlossene Formen des Unterrichts in jahrgangshomogenen Klassen differenziert operationalisiert und in der Wirksamkeit empirisch gut erforscht (z. B. Baumert et al. 2010; Praetorius et al. 2018, 2020). Für das Lernen in jahrgangsgemischten Klassen ist davon auszugehen, dass die Kriterien lernförderlicher Tiefenstrukturen analog gelten – es gibt jedoch kaum Studien zur Frage, inwieweit sich diese in der Unterrichtspraxis finden lassen.

Allerdings lassen sich theoretische Annahmen zur Wirkweise von jahrgangsgemischtem Unterricht aus der Unterrichtsqualitätsforschung ableiten, zu denen es teilweise Indizien aus der empirischen Forschung zum jahrgangsgemischten Lernen gibt: So kann aus theoretischer Perspektive beispielweise in jahrgangsgemischten Klassen das Modell der älteren Schüler*innen den jüngeren das Ankommen im schulischen Lernen und somit die *Klassenführung* für die Lehrkräfte erleichtern. Indizien dafür und für eine damit verbundene inhaltsbezogenere und *effektivere Lernzeitnutzung* finden sich in einer Interviewstudie mit Lehrkräften (Sonnleitner 2020). Auch kann die Jahrgangsmischung zu einer *veränderten Inhaltsauswahl und -organisation* führen. Ronksley-Pavia et al. (2019) unterscheiden zwischen parallelen, rotierenden, spiralförmigen, entwicklungsorientierten und projektbasierten Curricula, die unterschiedliche Effekte auf das Lernen erwarten lassen. Diese Formen spiegeln sich international – im Gegensatz zum deutschen Sprachraum – auch in den Begrifflichkeiten wider (Cornish 2010): Begriffe wie composite classes oder multi-grade classes bezeichnen eher eine pragmatische Umsetzung im Rahmen eines Unterrichts, der Abteilungs- und Frontalunterricht einschließt und sich stark an den jeweils gültigen, einzelnen Jahrgangsstufenlehrplänen orientiert. Hier sind allenfalls durch die kleineren Gruppengrößen bei der getrennten Beschulung Effekte zu erwarten – wenn z. B. ein Teil der Klasse eine Einführung erhält, während der andere für sich arbeitet (vgl. Brahm 2006). Stage classes oder (in Annäherung) multi-age classes sind dagegen oft mit der Möglichkeit des kürzeren oder längeren Verweilens kombiniert und haben einen hohen Anspruch an *individuelles, entwicklungsorientiertes Lernen*. Gerade im deutschen Sprachraum finden sich dann auch Hinweise, dass jahrgangsgemischtes Lernen das Unterrichtsangebot methodisch hin zu mehr Binnendifferenzierung verändert (vgl. z. B. Berthold 2010; Pape 2016; Thoren und Brunner 2019) und zu einer erhöhten Adaptivität des Unterrichts durch individuelle

Lernangebote führt (Munser-Kiefer et al. 2017). In jahrgangsgemischten Klassen gibt es zudem intensivere Möglichkeiten der individuellen Unterstützung, da diese nicht nur durch die Lehrkraft, sondern auch durch Peers bzw. durch Lerntandems oder Lernpatenschaften in einer Kombination aus jüngeren und älteren Kindern erfolgen kann. Dadurch könnten sich für die jüngeren Schüler*innen Gelegenheiten beschleunigten Lernens ergeben, indem sie in einem Lerntandem bereits mit den Lerninhalten der nächsten Jahrgangsstufe arbeiten können und dabei individuelle Unterstützung erfahren. Ein Indiz für diesen positiven Effekt sind die Befunde der Interviewstudie von Feuchtenberger et al. 2019, in der die befragten Lehrpersonen v. a. für Kinder im ersten Besuchsjahr der jahrgangsgemischten Klasse Chancen und Vorteile angeben. Gerade für die Schüler*innen im zweiten Besuchsjahr kann dagegen eine *Konsolidierung der Inhalte durch Wiederholung* erwartet werden, wenn in spiralcurricular angelegten Klassenlehrplänen die Inhalte im zweiten Jahr auf höherem Niveau wiederkehren.

Zusammenfassend lässt sich festhalten, dass die Hoffnungen in Bezug auf die Jahrgangsmischung von der Qualität der Tiefenstruktur abhängig sind, die sich durch theoretische Annahmen und in Teilen auch durch Indizien aus der empirischen Forschung stützen lassen. Davon könnten Effekte sowohl auf die Leistung im Allgemeinen als auch differenzielle Effekte ausgehen.

1.1 Effekte jahrgangsgemischten Lernens auf Leistung

Die Befundlage zur Leistungsentwicklung in jahrgangsgemischten Klassen ist inkonsistent: Auf der einen Seite finden sich seit den 1990er-Jahren in Deutschland durch die bildungspolitisch begünstigte Zunahme von Jahrgangsmischungen zahlreiche Modellversuche, die mit eher positiven Evaluationsergebnissen einhergehen (z. B. Bayern: Klöver 2014; Brandenburg: Krüsken 2008; Baden-Württemberg: Ministerium für Kultus, Jugend und Sport Baden-Württemberg 2006). Erklärt werden kann dieser Effekt zumindest teilweise durch eine Positivauswahl an Lehrkräften, deren Innovationsfreude und meist günstige Einstellungen für Effekte mitverantwortlich sein könnten. Hinweise hierfür finden sich auch in der Begleituntersuchung der flächendeckend eingeführten jahrgangsübergreifenden Schulanfangsphase in Berlin (Thoren und Brunner 2019): Es ließen sich Typen – „‚Pioniere' und ‚allg. Nichtüberzeugte'" (S. 291) – identifizieren, die sich in Unterrichtsmerkmalen unterschieden, die mit effektivem Unterricht in heterogenen Lerngruppen assoziiert sind.

Auf der anderen Seite finden sich Studien, die eher keine Effekte auf die Leistung nachweisen. Für Deutschland dient die repräsentative Ländervergleichsstudie des Instituts für Qualitätsentwicklung im Bildungswesen IQB (Kuhl et al. 2013) als gewichtiger Beleg. Auch die Begleituntersuchung zur Schulanfangsphase in Berlin konnte keine generellen Unterschiede aufdecken (Thoren 2017; Thoren und Brunner 2019). International kommen Ronksley-Pavia et al. (2019) in ihrem systematischen Review zu empirischen Artikeln für die Jahre 1997 bis 2017 – allerdings nur für kleine Schulen – zu dem unbefriedigenden Ergebnis, dass in manchen Studien keine Unterschiede im Leistungsbereich nachgewiesen werden können, andere Studien mit (kleinen) positiven, wiederum andere mit (kleinen) negativen Entwicklungen aufwarten. Dies deckt sich auch mit älteren Metaanalysen: Veenman (1996) fand

keine signifikanten Unterschiede, Gutiérrez und Slavin (1992) entdeckten positive Effekte auf die Leistung, bei Sundell (1994) und Russel et al. (1998) wurden Nachteile des jahrgangsgemischten Lernens für die Leistung festgestellt.

Diese widersprüchlichen Ergebnisse sind jedoch nicht völlig unerwartet: Zum einen fehlen bei diesen Überblicksstudien Informationen über die Umsetzung des jahrgangsgemischten Lernens und zum andern wird nicht darauf eingegangen, inwieweit diese Form der Klassenzusammensetzung für bestimmte Schüler*innen besonders geeignet oder ungeeignet ist.

1.2 Differenzielle Effekte jahrgangsgemischten Lernens auf Leistung

Differenzielle Effekte werden hier aus drei empirischen Suchrichtungen berichtet: erstens bezogen auf verschiedene Klassenstufen (Eingangsstufe vs. 3./4. Klasse vs. höher), zweitens auf Schüler*innen innerhalb einer jahrgangsgemischten Klasse (erstes vs. zweites/letztes Besuchsjahr) sowie drittens auf unterschiedliche Leistungsgruppen.

Zur Frage, ob Effekte der Jahrgangsmischung nach der Schuleingangsstufe zu finden sind, gibt es bislang nur wenige Studien. So ist auch heute noch auf die ältere Metaanalyse von Veenman (1996) zurückgreifen, die zwar keine unterschiedlichen Effekte für Leistungsentwicklung in jahrgangsgemischten oder -homogenen Klassen der Jahrgangsstufen 1 bis 6 zeigt, aber eine signifikante Varianzaufklärung für die Klassenstufe: Die neun Studien aus Jahrgangsstufe 1 und 2 deckten einen kleinen Leistungsvorsprung für die Jahrgangsmischung auf, für die Jahrgangstufe 3 und 4 zeichneten sich keinerlei Effekte (20 Studien) ab, für die fünfte und sechste Jahrgangsstufe war sogar ein kleiner negativer Effekt (5 Studien) nachweisbar (vgl. auch Lindström und Lindahl 2011). Im Widerspruch zu der Annahme, dass bei steigender Klassenstufe ungünstigere Effekte für Jahrgangmischungen nachweisbar sind, steht dagegen die groß angelegte norwegische Studie von Leuven und Rønning (2011): Für kombinierte Klassen zeigen sich sogar noch in siebten bis neunten Jahrgangsstufen günstigere Leistungsergebnisse. Die Erklärung liegt dabei aber nicht auf der Klassenstufe oder den Klassenstufen generell, sondern auf den unterschiedlichen Effekten durch die Altersgruppe der Peers innerhalb der Jahrgangsmischung.

Erklären lässt sich dies durch Unterschiede *innerhalb* einer jahrgangsgemischten Gruppe und damit durch differenzielle Effekte für die Jüngeren und die Älteren bzw. bei den unterschiedlichen Alters- oder Jahrgangsgruppen. Bei Leuven und Rønning (2011) deutet sich an, dass der Lerngewinn älterer Schüler*innen abnahm, während die anderen von den älteren Peers signifikant profitieren. Auch Hartinger et al. (2011) fanden für die jahrgangsgemischte Eingangsstufe einen steileren Leistungszuwachs für die Schüler*innen im ersten Schulbesuchsjahr, der im zweiten Schulbesuchsjahr abflachte, sodass sich der Vorsprung der jahrgangsgemischt unterrichteten Schüler*innen wieder nivellierte. Nach Laging (2010) lernen die Kinder dabei in asymmetrischer Interaktion und einem durch den Altersunterschied geprägten Rollenbewusstsein. Hier ließen sich in altersgemischten Gruppen mehr Empathie und Unterstützung finden als bei der konkurrenzhaltigeren Kooperation Gleichaltriger. Indizien für eine Unterstützung der jüngeren Schüler*innen durch die älteren fand Campana Schleusener (2014) in einer Beobachtungsstudie in Basisstu-

fenklassen (4 bis 8-Jährige), wobei die Hilfestellungen der Älteren sich vor allem auf das Begleiten des Lösungswegs (71 % für Vorzeigen, Zurechtweisen, Anleiten) beziehen. Matz und Knauf (2010) beobachteten darüber hinaus in einer Jahrgangsmischung 1–4 den Trend, dass Hilfsangebote mit der Jahrgangsstufe zuzunehmen schienen (7 % der Erstklässler, 20 % der Zweitklässler, 13 % der Drittklässler, 60 % der Viertklässler). Perren und Malti (2016) konnten ferner zeigen, dass sich die Fähigkeit der Schüler*innen, adaptiv Hilfestellung zu leisten, während der Grundschulzeit zunehmend auszudifferenzieren scheint. Das lässt ein Potenzial verstärkter individueller Unterstützung für das einzelne Kind – zumindest durch die Peers – vermuten. Diese Befunde lassen Veränderungen der Unterrichtsprozesse in jahrgangsgemischten Klassen erwarten und zeigen das Potenzial, das Lernen und damit die Leistungen gerade der jüngeren Kinder oder auch der unteren Leistungsgruppen zu verbessern.

Von daher ist es sinnvoll, auch die Lernvoraussetzungen der Kinder zu betrachten: Gölitz (2008) untersuchte im Rahmen der Studie „Schulanfang auf neuen Wegen“ den Einfluss der Jahrgangsmischung auf eine Risikogruppe mit defizitären Ausgangslagen (unter anderem) in der phonologischen Bewusstheit sowie im Mengenvorwissen. Er fand in der ersten Klasse einen kleinen negativen Effekt für die Jahrgangsmischung für den Bereich Lesen, der sich zum Ende der zweiten Klasse nivellierte; für Mathematik zeigte sich zum Ende der ersten Klasse dagegen ein kleiner Vorteil, tendenziell leicht steigend zum Ende der zweiten Klasse. Von Waaden (2017) begleitete Risikokinder in jahrgangsgemischten Klassen in Mathematik und konnte hier feststellen, dass niedrige Ausgangswerte sich auch in der Jahrgangsmischung manifestierten. Grittner et al. (2013) zeigten dagegen, dass Unterschiede vor allem auf Schüler*innen mit günstigen Leistungsvoraussetzungen zurückzuführen sind. Erklären lässt sich dies unter anderem durch die zusätzliche Anregung und die ergänzenden Angebote, die für diese Schüler*innen passend und förderlich sein können.

2 Forschungsfragen und Hypothesen

Die Befundlage zum Vergleich von Effekten von jahrgangsgemischt und jahrgangshomogen unterrichteten Klassen ergibt kein eindeutiges Bild. Positive Effekte lassen sich vorrangig in Modellversuchen in der Eingangsstufe nachweisen; speziell für die dritte und vierte Klasse weist der Forschungsstand eher auf eine Pattsituation hin, allerdings mit nur wenigen Studien aus dem deutschsprachigen Raum. Aufgrund der dürftigen Befundlage verbleiben die Forschungsfragen hier zunächst auf der Oberflächenstruktur und sind eher explorativ angelegt. Die erste Forschungsfrage richtet sich auf einen grundsätzlichen Vergleich jahrgangsgemischten und -homogenen Unterrichts in der dritten und vierten Jahrgangsstufe:

*1 Unterscheidet sich die schulische Leistungsentwicklung von Schüler*innen, die in der dritten und vierten Jahrgangsstufe in jahrgangsgemischten Klassen unterrichtet wurden, von denen aus jahrgangshomogenen Klassen?*

Die Forschungslage legt die Vermutung nahe, dass keine großen Unterschiede zu finden sind, wenn man alle Schüler*innen über die beiden Organisationsformen hinweg vergleicht.

Untersucht werden zudem folgende weiterführende differenzielle Fragen:

*2 Wirkt sich die Unterrichtsorganisation (jahrgangsgemischt vs. jahrgangshomogen) unterschiedlich auf Kinder der verschiedenen Schulbesuchsjahre (Dritt- bzw. Viertklässler*innen) aus?*

3 Gibt es Effekte für unterschiedliche Leistungsgruppen?

Aufgrund des skizzierten Forschungsstands sind bei einer repräsentativen Stichprobe zwischen jahrgangsgemischten und jahrgangshomogenen Klassen keine signifikanten Leistungsunterschiede zum Ende der vierten Jahrgangsstufe (H1), aber zum Ende der dritten Jahrgangsstufe zu erwarten (H2). Außerdem ist bei einem Vergleich von jahrgangsgemischten und -homogenen Klassen anzunehmen, dass leistungsstarke Schüler*innen gerade während der dritten Jahrgangsstufe profitieren (H3a), u. U. weil sie durch die vorgezogenen Inhalte der vierten Jahrgangsstufe herausgefordert und ihre Leistung katalysiert wird. Leistungsschwache Schüler*innen sollten dagegen vor allem während der vierten Jahrgangsstufe vergleichsweise positive Leistungsentwicklungen zeigen (H3b), weil hier vielleicht die Wiederholung der Inhalte und ihre Rolle als fortgeschrittene Lerner die Leistungsentwicklung begünstigen könnten.

3 Methode

3.1 Untersuchungsdesign

Zur Untersuchung dieser Forschungsfragen wurde eine quasi-experimentelle Längsschnittstudie durchgeführt, bei der die Leistungsentwicklung jahrgangshomogen und -gemischt unterrichteter Schüler*innen des dritten und vierten Schuljahres miteinander verglichen wurde. Die Studie startete im Schuljahr 2014/2015; die Daten wurden an drei verschiedenen Messzeitpunkten erhoben: 1) zu Beginn, 2) am Ende der dritten und 3) am Schluss der vierten Jahrgangsstufe. Zur Messung der Lernentwicklung der Schüler*innen wurden lehrplanvalide Tests zu den Fächern Deutsch (im Bereich Lesen) bzw. Mathematik (in den Themenbereichen Algebra, Geometrie und Sachrechnen) verwendet. Diese beinhalteten ausschließlich zentrale Lerninhalte der Grundschule, sodass davon ausgegangen werden kann, dass diese in allen untersuchten Klassen thematisiert und unterrichtet wurden. Die Befunde werden für die beiden Fächer getrennt betrachtet, da beide gute, aber differenzielle Indikatoren für den Lern- und Leistungsfortschritt der Kinder darstellen. Die Untersuchung möglicher fachbezogener Effekte steht jedoch nicht vergleichend im Fokus dieser Studie. Zusätzlich wurden auf der Ebene der Schüler*innen verschiedene Kovariaten wie Geschlecht, Bildungsnähe des Elternhauses, fachbezogenes Selbstkonzept, Motivation und Schulfreude erfasst und in den Analysen berücksichtigt.

3.2 Stichprobe

Die Stichprobe umfasst Schulen aus den Städten Augsburg und Nürnberg sowie deren Umland. Ausgegangen wurde bei der Rekrutierung der Stichprobe von jahrgangsgemischten Klassen, in denen ausschließlich die Jahrgänge 3 und 4 kombiniert wurden. Weitere Formen der Jahrgangsmischung gingen nicht in die Untersuchung ein. Es gelang, ca. 90 % der jahrgangsgemischten Klassen der untersuchten Gebiete für eine Teilnahme zu gewinnen. Die wenigen Gründe für eine Absage streuten und ließen keine Systematik erkennen, sodass die Stichprobe als repräsentativ für die beiden Regionen gelten kann. Für die Kontrollgruppe der jahrgangshomogenen Klassen wurden die zuständigen Schulrät*innen gebeten, Schulen mit vergleichbarem Sprengel und Lehrkräfte mit vergleichbarem Engagement zu nennen. Da vorab unklar war, ob deren Einschätzungen zuträfen, wurde bewusst eine umfangreichere Kontrollgruppe anvisiert, um etwaigen Stichprobenverzerrungen mit geeigneten Matchingverfahren begegnen zu können.

Insgesamt nahmen 1644 Schüler*innen aus 125 Klassen (davon 68 jahrgangsstufengemischt) an 58 Grundschulen teil, die von 125 Lehrkräften (91,7 % weiblich; mittleres Dienstalter 15,8 Jahre, $SD = 11{,}3$) unterrichtet wurden. Eine detailliertere Beschreibung der Zusammensetzung der Stichprobe der Schüler*innen getrennt nach jahrgangsgemischten und -homogenen Klassen unter Berücksichtigung zentraler Kovariaten findet sich in Tab. 1.

3.3 Untersuchungsinstrumente

Lesen Zur Erfassung der Leseleistung wurde zu Beginn der dritten und am Ende der vierten Klasse derselbe Test aus VERA 2006 verwendet (kontinuierlicher Sachtext, geschlossenes und offenes Antwortformat, Subskalen: hierarchieniedrige und hierarchiehöhere Verstehensprozesse, 13 Items, Cronbachs $\alpha_{MZP1} = 0{,}73$; $\alpha_{MZP3} = 0{,}72$). Am Ende der dritten Klasse wurde die Leseleistung mithilfe der bayernweit durchgeführten Vergleichsarbeiten gemessen.

Mathematik Unter Rückgriff auf Aufgaben aus ILEA (LISUM 2008) sowie des Probeunterrichts für weiterführende Schulen (vgl. z. B. ISB 2014) wurden lehrplankonforme Tests zur Erhebung der Mathematikleistung (Zahl- und Mengenerfassung, Rechnen, Sachrechnen, Geometrie) zu Beginn der dritten (47 Items, $\alpha_{MZP1} = 0{,}89$) und am Ende der vierten Klasse (20 Items, $\alpha_{MZP3} = 0{,}81$) entwickelt. Am Ende der dritten Klasse wurde die Mathematikleistung ebenfalls mithilfe der bayernweit durchgeführten Vergleichsarbeiten erhoben.

Kovariaten Alle Kovariaten wurden zum ersten Messzeitpunkt über Fragebögen mittels Einzelitems (z. B. Geschlecht, Anzahl der Bücher im Haushalt für Bildungsnähe, Eltern- bzw. Familiensprache als Hauptkommunikationssprache zwischen den jeweiligen Familienmitgliedern) oder entsprechende Skalen erfasst. Diese wiesen jeweils gute Reliabilitäten auf: Einstellung zu Mitschüler*innen und Schule (8 Items, $\alpha = 0{,}82$), Selbstkonzept Lesen (5 Items, $\alpha = 0{,}82$), Selbstkonzept Mathematik (10 Items, $\alpha = 0{,}87$) (zu den einzelnen Skalen aus der KILIA-Studie, vgl.

Tab. 1 Überblick über Mittelwerte (*M*), Standardabweichungen (*SD*) und *z*-Differenzen zu zentralen Stichprobenmerkmalen vor und nach dem Matching in Bezug auf die Lese- und Mathematikleistungen von Schüler*innen in jahrgangshomogenen (*JH*) und jahrgangsgemischten (*JM*) Klassen

	Gesamtstichprobe ($N = 1644$)					Matchingstichprobe Lesen ($N = 1326$)					Matchingstichprobe Mathematik ($N = 1330$)				
	JH		JM		*z*-Diff.	JH		JM		*z*-Diff.	JH		JM		*z*-Diff.
N	979		665			663		663			665		665		
	M	*SD*	*M*	*SD*		*M*	*SD*	*M*	*SD*		*M*	*SD*	*M*	*SD*	
Geschlecht (% weiblich)	48	–	50	–	0,04	52	–	50	–	0,04	51	–	50	–	0,02
Bildungsnähe	2,90	1,04	2,82	1,04	0,07	2,84	1,07	2,83	1,04	0,02	2,85	1,05	2,82	1,04	0,03
Elternsprache	0,20	0,36	0,18	0,36	0,06	0,18	0,34	0,17	0,36	0,01	0,19	0,35	0,18	0,36	0,03
Familiensprache	0,13	0,30	0,13	0,31	0,02	0,13	0,30	0,12	0,31	0,01	0,12	0,29	0,13	0,31	0,01
Intrinsische Motivation	3,43	1,38	3,53	1,36	0,07	3,51	1,40	3,53	1,36	0,02	3,52	1,42	3,53	1,36	0,01
Identifizierte Motivation	3,85	1,00	3,71	1,02	*0,13*	3,72	1,01	3,71	1,02	0,00	3,69	1,01	3,71	1,02	0,02
Introjizierte Motivation	1,58	1,25	1,70	1,21	*0,10*	1,71	1,27	1,71	1,22	0,00	1,72	1,30	1,70	1,21	0,01
Externale Motivation	3,14	1,14	3,05	1,22	0,08	3,07	1,15	3,05	1,22	0,01	3,07	1,13	3,05	1,22	0,02
Einstellung zu Schüler*innen und Schule	2,32	0,75	2,36	0,71	0,05	2,36	0,71	2,36	0,71	0,01	2,35	0,73	2,36	0,71	0,00
Selbstkonzept Lesen	2,38	0,54	2,38	0,50	0,01	2,40	0,52	2,38	0,50	0,04	–	–	–	–	–
Leseleistung	–0,03	1,01	0,04	0,98	0,07	0,00	1,01	0,00	0,99	0,00	–	–	–	–	–
Selbstkonzept Mathematik	2,07	0,48	2,02	0,49	*0,10*	–	–	–	–	–	2,03	0,51	2,02	0,49	0,02
Mathematikleistung	–0,01	1,01	0,01	0,99	0,01	–	–	–	–	–	0,00	1,01	0,00	0,99	0,00
Mittlere absolute standardisierte Diff.	–	–	–	–	0,06	–	–	–	–	0,01	–	–	–	–	0,02

Bücher: 0 = keine oder wenige, 1 = zusammen etwa ein Regalbrett, 2 = zusammen etwa ein Regal, 3 = zusammen etwa drei Regale, 4 = über 200 Bücher; Eltern- bzw. Familiensprache: 0 = deutsch, 1 = nichtdeutsch; Motivation: Summenscore aus je zweimal drei Items pro Kategorie (intr., ident., intro., ext.): 0 = Ablehnung, 1 = Zustimmung; Einstellung: 0 = trifft nicht zu, ..., 3 = trifft voll zu; Selbstkonzept: 0 = trifft nicht zu, ..., 3 = trifft voll zu; Lese- und Mathematikleistung: *z*-standardisiert; *kursiv*: $p \leq 0{,}05$, signifikant

Kammermeyer und Martschinke 2006). Die Motivation, mit den – auch im Self-Regulations-Questionnaire (Ryan und Connell o. J.) erhobenen – Motivationsstilen intrinsisch, identifiziert, introjiziert und external wurde mithilfe eines Dominanz-Paarvergleichs erhoben (vgl. Hartinger et al. 2004). Dazu wurden zu jedem dieser vier Motivationsstile zwei Items formuliert, sodass sich insgesamt zwölf Paarvergleiche ergeben (z. B. *Im Unterricht arbeite ich mit, a) weil ich mich schämen würde, wenn ich schlecht bin* oder *b) weil die Schule sehr wichtig ist*). Die Kinder mussten sich dann für eine der beiden Optionen entscheiden. Als Maß der Konsistenz solcher Paarvergleiche schlagen Bortz et al. (2008, S. 489 ff.) die Berechnung eines Kennwerts auf der Grundlage der (zu vermeidenden) inkonsistenten Triaden vor. Hier zeigt sich, dass keiner der Dominanzpaarvergleiche inkonsistente Triaden aufweist, sodass die Dominanzpaarvergleiche als reliabel angesehen werden können.

3.4 Datenaufbereitung und -analyse

3.4.1 Umgang mit fehlenden Werten

Fehlende Werte bei einer oder mehreren Variablen wurden für jeden Messzeitpunkt mithilfe des Expectation-Maximization-Algorithmus geschätzt (vgl. z. B. Enders 2010). Dies war möglich, da die fehlenden Werte eine unsystematische Verteilung aufwiesen und ihr Anteil bei allen Variablen maximal 14,7 % betrug (z. B. Madley-Dowd et al. 2019). Anschließend wurden zuerst für jeden Messzeitpunkt einzeln, dann über die verschiedenen Messzeitpunkte hinweg die fehlenden Werte der Schüler*innen geschätzt, bei denen hierfür eine Mindestdatenmenge von 70 % aller Variablen verfügbar war.

3.4.2 Balancierung von Stichprobenunterschieden (Propensity-Score-Matching)

Um im Rahmen des quasi-experimentellen Studiendesigns für möglichst viele Einflussgrößen bei der Analyse zu kontrollieren, wurde mithilfe eines Propensity-Score-Matchingverfahrens adjustiert (Guo und Fraser 2015; Kuss et al. 2016). Zur Schätzung des Propensity-Scores wurde mit Blick auf die Mathematik- bzw. Leseleistung je ein separates logistisches Regressionsmodell mit dem dichotomen Kriterium jahrgangshomogener bzw. -gemischter Unterricht berechnet. Alle darin als unabhängige Variablen eingehenden Merkmale (zu Beginn der dritten Klasse) wurden a priori auf Basis theoretischer Überlegungen ausgewählt. Aus diesem Grund wurden sie trotz ihrer geringen Beiträge zur Verbesserung der jeweiligen Modellgüte (McFaddens Pseudo R^2 für Lesen bzw. Mathematik 0,01) beibehalten (vgl. Tab. 1). Dabei bleiben mathematikbezogene Variablen im Modell für Lesen, lesespezifische im Modell für Mathematik unberücksichtigt. Um den Datenpool der Experimentalgruppe bei den Analysen möglichst vollständig auszuschöpfen, wurde ein 1:1-Matching unter Verwendung eines „nearest neighbour"-Algorithmus durchgeführt (Guo und Fraser 2015), wobei aufgrund der deutlich umfangreicheren Kontrollgruppe die Festlegung einer maximalen Äquivalenzunschärfe bei der Fallzuordnung (Caliper-Weite) nicht notwendig war (und – wie zusätzliche Analysen zeigten – eine zur Schätzung von

Mittelwertdifferenzen angemessene Caliper-Weite von 0,30 vergleichbare Ergebnisse erzielte; cf. Austin 2010; Wang et al. 2013).

Wie aus Tab. 1 ersichtlich wird, konnten dadurch die signifikanten Unterschiede zwischen Kontroll- und Experimentalgruppe hinsichtlich dreier Kovariaten (identifizierte und introjizierte Motivation, Selbstkonzept Mathematik) ausgeglichen und jedem Kind in einer jahrgangsgemischten Klasse genau ein*e Matchingpartner*in mit ähnlichen Merkmalen in der anderen Gruppe zugewiesen werden.[1] Eine hinreichende Balancierung der Daten indiziert schließlich auch ein Vergleich der einzelnen *z*-Differenzen ebenso wie der mittleren absoluten standardisierten Differenzen, die in der gematchten Stichprobe für Analysen sowohl zur Lese- als auch zur Mathematikleistung geringer ausfallen.

3.4.3 Aufbereitung der Leistungsdaten und Bildung von Leistungsquartilen

Um die differierenden Rohpunktskalen der inhaltlich vergleichbaren Testinstrumente, die für Lesen bzw. Mathematik zu den drei verschiedenen Zeitpunkten eingesetzt wurden, jeweils zu vereinheitlichen und gemeinsam analysieren zu können, wurden die entsprechenden Leistungsdaten im Anschluss an das Matchingverfahren *z*-standardisiert. Im Folgenden wird über diese *z*-standardisierten Werte (Tab. 2, 3, 4, 5 und 6) berichtet.

Nach Auswertungen mit der gesamten gematchten Stichprobe zur Untersuchung der ersten beiden Forschungsfragen wird diese für die dritte Fragestellung in Quartile aufgeteilt, um unterschiedliche Leistungsentwicklungen in den jeweiligen Teilgruppen wie auch differenzielle Effekte der Jahrgangsmischung in Abhängigkeit von den Leistungsgruppen zu betrachten (Balancierung von Stichprobenunterschieden). Die Einteilung in die vier Leistungsgruppen erfolgte dabei nicht auf Basis der Matchinggewichte, da in diese neben Leistungs- auch andere Kovariaten aus dem Persönlichkeitsbereich (vgl. Tab. 1) eingehen, sondern in einem anschließenden separaten Schritt auf Basis der *z*-Werte im Lesen bzw. in Mathematik zu Anfang der dritten Jahrgangsstufe. Dadurch weisen die Startwerte der einzelnen Quartilsgruppen hier im Vergleich zu späteren Zeitpunkten relativ wenig Streuung auf (vgl. Abb. 1 und 2).

3.4.4 Zur Auswertung eingesetzte Verfahren

Um Entwicklungsverläufe und Unterschiede zwischen jahrgangshomogenen und -gemischten Gruppen insgesamt (vgl. Tab. 2 und 3) wie auch für die vier Leistungsquartile (vgl. Tab. 4, 5 und 6) zu den einzelnen Messzeitpunkten zu bestimmen, wurden für Lesen und Mathematik jeweils gemischte lineare Modelle geschätzt. Diese berücksichtigen nicht nur die personenspezifischen Abhängigkeiten in den Längsschnittdaten mit drei Messzeitpunkten, die aus den wiederholten individuellen

[1] Da die Leseleistung zweier Schüler*innen aus der Jahrgangsmischung am Ende der vierten Jahrgangsstufe bei einer Ausreißeranalyse unplausible Werte aufwies (u. a. mehr als 3,5 Standardabweichungen geringer als der Gruppenmittelwert), wurden diese vor der Matching-Prozedur aus der Stichprobe entfernt. Hinsichtlich der Mathematikleistung bestanden keine Auffälligkeiten.

Leistungsmessungen resultieren (vgl. die Werte der Intraklassenkorrelation [ICC] in Tab. 3, 5 und 6), sondern besitzen darüber hinaus auch weitere methodische Vorteile in Hinblick auf günstigere Analysevoraussetzungen, Teststärke oder den Umgang mit fehlenden Werten (vgl. für Details: Hilbert et al. 2019). Die Prädiktoren Gruppenzugehörigkeit (jahrgangshomogen vs. -gemischt) und die Zeitvariable (Anfang bzw. Ende der dritten, Ende der vierten Klassenstufe) werden dummy-kodiert (0/1), wobei jahrgangshomogener Unterricht und der Zeitpunkt Ende der dritten Klasse als Referenzkategorien angelegt werden. Die Wahl dieses mittleren Zeitpunkts als Referenzkategorie ist deshalb von Vorteil, weil in einem einzigen Modell (d. h. ohne Alphafehlerkumulation) Effekte zwischen Kontroll- und Experimentalgruppe ebenso wie zwischen den drei Messzeitpunkten direkt paarweise geschätzt werden können. Von besonderem Interesse sind hierbei die jeweiligen Interaktionseffekte (Gruppe × Anfang 3. Jahrgangsstufe bzw. Gruppe × Ende 4. Jahrgangsstufe), da diese die *zusätzliche* Veränderung in der jahrgangsgemischten Gruppe ausdrücken (unter Berücksichtigung der Veränderung der jahrgangshomogenen Gruppe).

Die Analysevoraussetzungen (z. B. Normalverteilung) wurden graphisch und inferenzstatistisch überprüft und schränken die Interpretierbarkeit der Ergebnisse nicht ein. Alle weiterführenden Auswertungen wurden mit der Statistiksoftware *R* (*R* Core Team 2020) durchgeführt; genutzt wurden vor allem die folgenden Zusatzpakete: MatchIt (Ho et al. 2011), ggplot2 (Wickham 2016), multilevel (Bliese 2016), lme4 (Bates et al. 2014), lmerTest (Kuznetsova et al. 2017) und MuMIn (Barton 2020).

4 Ergebnisse

Nachstehend werden zuerst die Ergebnisse für Lesen und Mathematik in Bezug auf die gematchte Gesamtstichprobe, daraufhin nach den vier gebildeten Leistungsquartilen getrennt präsentiert. Im Vordergrund steht damit anfangs die Überprüfung der Forschungshypothesen 1 und 2 zum Einfluss von jahrgemischtem Unterricht auf die Leistungsentwicklung in Lesen und Mathematik – generell bzw. bezogen auf die Schulbesuchsjahre. Anschließend wird die Analyse verfeinert und berichtet, ob jahrgangsgemischter Unterricht in einzelnen Leistungsquartilen zu differenziellen Effekten führt (H3a/b).

4.1 Einfluss jahrgangsgemischten Unterrichts auf die Leistungsentwicklung in Bezug auf die Gesamtstichprobe

Einen ersten deskriptiven Überblick über die Leistungsentwicklungen gibt Tab. 2. Aufgrund des Matchings mit anschließender *z*-Standardisierung sind die arithmetischen Mittelwerte in beiden Gruppen zu Anfang der dritten Jahrgangsstufe im Lesen wie auch in Mathematik identisch; die Leistungsstreuung erstreckt sich über mehr als vier Standardabweichungen (und ist in Mathematik ausgeprägter als im Lesen).

Am Ende der dritten Jahrgangsstufe unterscheiden sich jahrgangsgemischte und -homogene Klassen sowohl im Lesen als auch in Mathematik durchschnittlich um ein Viertel der Standardabweichung ($d_{JH\text{-}JM} = 0{,}26$, $d_{JH\text{-}JM} = 0{,}24$, Tab. 2). Dies entspricht nach Cohen (1992) einem kleinen Effekt, der zudem signifikant ist ($p \leq 0{,}01$), wie

Tab. 2 Deskriptive Übersicht über die Leistungen in Lesen ($N = 1326$) und Mathematik ($N = 1330$) zu drei Messzeitpunkten (z-Werte)

	Anfang 3. Jahrgangsstufe				Ende 3. Jahrgangsstufe				Ende 4. Jahrgangsstufe			
	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*
Lesen	–2,47	1,91	0,00	1,00	–3,37	1,64	0,00	1,00	–3,57	2,11	0,00	1,00
Jahrgangshomogen	–2,47	1,91	0,00	1,01	–3,37	1,64	–0,13	1,10	–3,40	2,11	0,02	1,01
Jahrgangsgemischt	–2,47	1,91	0,00	0,99	–2,91	1,41	0,13	0,87	–3,57	2,00	–0,02	0,99
$d_{JH\text{-}JM}$ | 95 % KI für $d_{JH\text{-}JM}$	–	–	0,00 [–0,10; 0,11]		–	–	0,26 [0,15; 0,37]		–	–	–0,03 [–0,14; 0,08]	
Mathematik	–4,05	1,45	0,00	1,00	–2,71	2,11	0,00	1,00	–2,26	2,54	0,00	1,00
Jahrgangshomogen	–3,88	1,45	0,00	1,01	–2,54	2,11	–0,12	1,03	–2,26	2,54	0,00	0,99
Jahrgangsgemischt	–4,05	1,45	0,00	0,99	–2,71	2,11	0,12	0,96	–2,26	2,54	0,00	1,01
$d_{JH\text{-}JM}$ | 95 % KI für $d_{JH\text{-}JM}$	–	–	0,00 [–0,11; 0,10]		–	–	0,24 [0,13; 0,35]		–	–	0,00 [–0,11; 0,11]	

Vgl. zur Signifikanzüberprüfung Tab. 3

N Stichprobengröße, *Min* Minimum, *Max* Maximum, *M* arithmetischer Mittelwert, *SD* Standardabweichung, $d_{JH\text{-}JM}$ Effektstärke Cohens *d* (nach Cohen 1992: 0,2 klein, 0,5 mittel, 0,8 groß), *95 % KI* 95 % Konfidenzintervall

Tab. 3 Gemischte lineare Modelle für Lesen und Mathematik auf Basis der gematchten Gesamtstichprobe (unter Berücksichtigung der nach Schüler*innen sowie Klassen geordneten Datenstruktur)

	Lesen					Mathematik				
N | *Beob.* | *ICC*	1326		3978		45,9 %	1330		3990		51,6 %
Feste Effekte	*b*	*SE*	*df*	*t*	*p*	*b*	*SE*	*df*	*t*	*p*
Konstante	–0,13	0,06	172,32	–2,39	0,018	–0,13	0,06	154,30	–2,13	0,034
Jahrgangsmischung	0,26	0,08	184,96	3,35	<0,001	0,24	0,08	164,45	2,93	0,004
Anfang 3. Jahrgangsstufe	0,13	0,04	2647,98	3,19	0,001	0,12	0,04	2656,00	3,22	0,001
Ende 4. Jahrgangsstufe	0,15	0,04	2647,98	3,65	<0,001	0,12	0,04	2656,00	3,20	0,001
Anfang 3. Jahrgangsstufe × Jahrgangsmischung	–0,26	0,06	2647,98	–4,52	<0,001	–0,25	0,05	2655,99	–4,56	<0,001
Ende 4. Jahrgangsstufe × Jahrgangsmischung	–0,29	0,06	2647,98	–5,16	<0,001	–0,24	0,05	2655,99	–4,52	<0,001
Marg. R^2 | *kond.* R^2	0,01			0,47		0,01			0,52	

Alle im Modell enthaltenen Prädiktoren sind dummy-kodiert (0/1). Da der Messzeitpunkt am Ende der dritten Jahrgangsstufe zeitliche Referenzkategorie ist, sind die Regressionsgewichte bezüglich des Beginns der dritten Jahrgangsstufe in der Interpretationslogik zu invertieren

N Stichprobengröße, *Beob.* Beobachtungen, *ICC* Intraklassenkorrelation, *b* standardisierter Regressionskoeffizient, *SE* Standardfehler, *df* Freiheitsgrade; *t* *t*-Wert, *p* Wahrscheinlichkeit für den Fehler 1. Art, R^2 Determinationskoeffizient

Tab. 4 Deskriptive Übersicht über die Leistungen im Lesen und in Mathematik nach Messzeitpunkten, homogenen und gemischten Jahrgangsstufen (JH vs. JM) sowie Leistungsquartilen getrennt (*z*-Werte)

	Anfang 3. Jahrgangsstufe					Ende 3. Jahrgangsstufe					Ende 4. Jahrgangsstufe				
	JH		JM			JH		JM			JH		JM		
	M	*SD*	*M*	*SD*	$d_{\text{JH-JM}}$ | 95 % KI	*M*	*SD*	*M*	*SD*	$d_{\text{JH-JM}}$ | 95 % KI	*M*	*SD*	*M*	*SD*	$d_{\text{JH-JM}}$ | 95 % KI
Lesen															
1. Quartil	–1,32	0,31	–1,30	0,31	0,04 [–0,17; 0,25]	–0,61	1,06	–0,46	0,92	0,15 [–0,07; 0,36]	–0,79	0,86	–0,67	1,06	0,12 [–0,10; 0,33]
2. Quartil	–0,37	0,26	–0,36	0,27	0,05 [–0,17; 0,27]	–0,27	1,13	–0,07	0,77	0,21 [0,01; 0,43]	–0,16	0,93	–0,26	0,87	–0,11 [–0,32; 0,11]
3. Quartil	0,44	0,26	0,46	0,25	0,06 [–0,15; 0,28]	–0,03	1,09	0,34	0,72	0,41 [0,19; 0,63]	0,25	0,78	0,23	0,77	–0,03 [–0,25; 0,18]
4. Quartil	1,26	0,25	1,29	0,27	0,11 [–0,11; 0,32]	0,39	0,86	0,75	0,51	0,50 [0,27; 0,72]	0,77	0,73	0,68	0,67	–0,14 [–0,36; 0,08]
Mathematik															
1. Quartil	–1,41	0,77	–1,39	0,78	0,02 [–0,19; 0,24]	–0,69	0,91	–0,60	0,93	0,10 [–0,12; 0,31]	–0,74	0,85	–0,68	0,90	0,06 [–0,15; 0,28]
2. Quartil	–0,11	0,22	–0,14	0,21	–0,11 [–0,33; 0,10]	–0,41	0,93	–0,10	0,77	0,37 [0,15; 0,59]	–0,15	0,90	–0,28	0,83	–0,16 [–0,37; 0,06]
3. Quartil	0,50	0,17	0,51	0,18	0,03 [–0,17; 0,24]	0,10	0,88	0,42	0,74	0,39 [0,18; 0,60]	0,22	0,79	0,32	0,89	0,12 [–0,09; 0,33]
4. Quartil	1,05	0,18	1,06	0,18	0,05 [–0,17; 0,27]	0,52	0,95	0,79	0,79	0,31 [0,09; 0,54]	0,69	0,84	0,68	0,86	–0,01 [–0,23; 0,22]

Vgl. zur Signifikanzüberprüfung Tab. 5 und 6

N Stichprobengröße, *M* arithmetischer Mittelwert, *SD* Standardabweichung, $d_{JH\text{-}JM}$ Effektstärke Cohens *d* (nach Cohen 1992: 0,2 klein, 0,5 mittel, 0,8 groß), *95 % KI* 95 % Konfidenzintervall für $d_{\text{JH-JM}}$

Tab. 5 Gemischte lineare Modelle für Lesen nach Leistungsquartilen differenziert (unter Berücksichtigung der nach Schüler*innen sowie Klassen geordneten Datenstruktur)

	1. Quartil					2. Quartil					3. Quartil					4. Quartil				
N | *Beob.* | *ICC*	342		1026		8,8 %	322		966		6,4 %	337		1011		14,4 %	325		975		6,6 %
Feste Effekte	*b*	*SE*	*df*	*t*	*p*	*b*	*SE*	*df*	*t*	*p*	*b*	*SE*	*df*	*t*	*p*	*b*	*SE*	*df*	*t*	*p*
Konstante	–0,61	0,07	290,27	–8,68	<0,001	–0,26	0,07	341,16	–3,94	<0,001	–0,02	0,07	291,60	–2,62	0,722	0,38	0,05	221,99	7,43	<0,001
Jahrgangs-mischung (JM)	0,16	0,10	335,12	1,62	0,106	0,19	0,09	381,56	2,09	0,038	0,38	0,09	298,36	4,23	<0,001	0,36	0,07	275,30	4,87	<0,001
Anfang 3. Jahr-gangsstufe	–0,70	0,08	680,00	–8,83	<0,001	–0,10	0,08	640,00	–1,22	0,221	0,48	0,07	670,48	6,37	<0,001	0,87	0,06	646,00	14,94	<0,001
Ende 4. Jahrgangs-stufe	–0,17	0,08	680,00	–2,18	0,030	0,11	0,08	640,00	1,39	0,166	0,28	0,07	670,48	3,79	<0,001	0,38	0,06	646,00	6,53	<0,001
Anfang 3. Jgst. × JM	–0,14	0,11	680,00	–1,19	0,236	–0,19	0,11	640,00	–1,69	0,091	–0,35	0,1	670,48	–3,52	<0,001	–0,33	0,09	646,00	–3,85	<0,001
Ende 4. Jgst. × JM	–0,04	0,11	680,00	–0,32	0,751	–0,30	0,11	640,00	–2,62	0,009	–0,39	0,1	670,48	–3,91	<0,001	–0,45	0,09	646,00	–5,30	<0,001
Marg. R^2 | *kond.* R^2	0,14			0,29		0,02			0,14		0,05			0,19		0,23			0,36	

Alle im Modell enthaltenen Prädiktoren sind dummy-kodiert (0/1). Da der Messzeitpunkt am Ende der dritten Jahrgangsstufe zeitliche Referenzkategorie ist, sind die Regressionsgewichte bezüglich des Beginns der dritten Jahrgangsstufe in der Interpretationslogik zu invertieren

N Stichprobengröße, *Beob.* Beobachtungen, *ICC* Intraklassenkorrelation, *b* standardisierter Regressionskoeffizient, *SE* Standardfehler, *df* Freiheitsgrade, *t* *t*-Wert, *p* Wahrscheinlichkeit für den Fehler 1. Art, R^2 Determinationskoeffizient

Tab. 6 Gemischte lineare Modelle für Mathematik nach Leistungsquartilen differenziert (unter Berücksichtigung der nach Schüler*innen sowie Klassen geordneten Datenstruktur)

	1. Quartil					2. Quartil					3. Quartil					4. Quartil				
N | *Beob.* | *ICC*	330		999		23,3 %	332		996		12,4 %	352		1056		20,9 %	313		939		16,7 %
Feste Effekte	*b*	*SE*	*df*	*t*	*p*	*b*	*SE*	*df*	*t*	*p*	*b*	*SE*	*df*	*t*	*p*	*b*	*SE*	*df*	*t*	*p*
Konstante	–0,67	0,08	197,88	–8,88	<0,001	–0,41	0,06	293,15	–6,39	<0,001	0,10	0,06	219,12	1,67	0,096	0,53	0,07	212,50	8,05	<0,001
Jahrgangs-mischung (JM)	0,09	0,11	217,43	0,87	0,388	0,30	0,09	305,45	3,44	<0,001	0,31	0,08	242,59	3,80	<0,001	0,26	0,09	235,82	2,84	0,005
Anfang 3. Jahr-gangsstufe	–0,72	0,08	662,00	–9,56	<0,001	0,30	0,07	659,99	4,07	<0,001	0,40	0,06	700,00	6,14	<0,001	0,53	0,07	622,00	7,55	<0,001
Ende 4. Jahrgangs-stufe	–0,05	0,08	662,00	–0,67	0,502	0,26	0,07	659,99	3,58	<0,001	0,12	0,06	700,00	1,78	0,076	0,17	0,07	622,00	2,36	0,019
Anfang 3. Jgst. × JM	–0,07	0,11	662,00	–0,67	0,501	–0,34	0,10	659,99	–3,29	0,001	–0,31	0,09	700,00	–3,40	<0,001	–0,26	0,10	622,00	–2,61	0,009
Ende 4. Jgst. × JM	–0,04	0,11	662,00	–0,35	0,729	–0,45	0,10	659,99	–4,38	<0,001	–0,21	0,09	700,00	–2,35	0,019	–0,28	0,10	622,00	–2,76	0,006
Marg. R^2 | *kond.* R^2	0,14			0,44		0,02			0,16		0,04			0,26		0,07			0,27	

Alle im Modell enthaltenen Prädiktoren sind dummy-kodiert (0/1). Da der Messzeitpunkt am Ende der dritten Jahrgangsstufe zeitliche Referenzkategorie ist, sind die Regressionsgewichte bezüglich des Beginns der dritten Jahrgangsstufe in der Interpretationslogik zu invertieren

N Stichprobengröße, *Beob.* Beobachtungen, *ICC* Intraklassenkorrelation, *b* standardisierter Regressionskoeffizient, *SE* Standardfehler, *df* Freiheitsgrade, *t* *t*-Wert; *p* Wahrscheinlichkeit für den Fehler 1. Art, R^2 Determinationskoeffizient

die Haupteffekte „Jahrgangsmischung“ der zugehörigen gemischten linearen Modelle unterstreichen (Tab. 3; temporale Referenzkategorie hier Ende 3. Jahrgangsstufe). Auch die relativen Leistungszunahmen in der jahrgangsgemischten (unter Berücksichtigung der Abnahmen der -homogenen) Gruppe bis zum Ende der dritten Jahrgangsstufe sind mit Blick auf die entsprechenden Interaktionseffekte („Anfang 3. Jahrgangsstufe × Jahrgangsmischung“, Tab. 3) signifikant. Diese im Lesen wie auch Mathematik jeweils höheren Werte in den jahrgangsgemischten Klassen verschwinden bis zum Ende der vierten Klasse jedoch vollständig, wobei diese gegenläufigen Entwicklungen gerade vor dem Hintergrund der relativen Leistungssteigerung der jahrgangshomogenen Klassen in diesem Zeitabschnitt überzufällig sind (vgl. die Interaktionseffekte „Ende 4. Jahrgangsstufe × Jahrgangsmischung“, Tab. 3).

Aus diesen Analysen ist bezüglich der Forschungsfragen 1 und 2 festzuhalten, dass sich die Leistungsentwicklungen (in Lesen und Mathematik) von Schüler*innen in jahrgangsgemischten und -homogenen Klassen unterscheiden. Zudem geht aus den vorliegenden Daten einerseits gemäß der ersten Hypothese hervor, dass gegen Ende der vierten Jahrgangsstufe keine Leistungsunterschiede zwischen beiden untersuchten Organisationsformen existieren. Anderseits unterstreicht diese aber auch die zweite Annahme, dass Schüler*innen gerade in der dritten Jahrgangsstufe von einem jahrgangsgemischten Unterricht profitieren können. Inwiefern dies insbesondere auf bestimmte Leistungsgruppen zutrifft, wird im Folgenden betrachtet.

4.2 Differenzielle Effekte jahrgangsgemischten Unterrichts auf die Leistungsentwicklung

In Tab. 4 sind die arithmetischen Mittelwerte und Standardabweichungen im Lesen und in Mathematik dargestellt – getrennt nach den Leistungsquartilen, die zu Zeitpunkt 1 gebildet wurden. Außerdem finden sich hier die Effektstärken zwischen den jeweiligen jahrgangshomogenen und -gemischten Gruppen zu den drei Erhebungszeitpunkten. Aufgrund des Matchings liegen am Anfang der dritten Jahrgangsstufe erwartungsgemäß weder in Bezug auf Lesen noch Mathematik in irgendeinem Quartil signifikante Unterschiede zwischen den beiden Gruppen vor. Zum Ende der dritten Jahrgangsstufe schneiden jedoch die jahrgangsgemischten Klassen in allen Leistungsquartilen sowohl im Lesen als auch in Mathematik im Mittel besser ab als die jahrgangshomogenen Klassen. Abgesehen vom Quartil mit den jeweils geringsten Ausgangswerten zu Beginn der dritten Jahrgangsstufe (hier ist diese Tendenz nur deskriptiv zu erkennen) zeigen sich in den übrigen Quartilen kleine ($d_{JH\text{-}JM} = 0{,}21$) bis mittlere ($d_{JH\text{-}JM} = 0{,}50$) Effekte, die mit Blick auf den Haupteffekt „Jahrgangsmischung“ der zugehörigen gemischten linearen Modelle (mit temporaler Referenzkategorie Ende 3. Jahrgangsstufe) signifikant sind ($p \leq 0{,}05$, Tab. 5 und 6). Ferner ist in Tab. 4 zu erkennen, dass diese Effekte im Lesen umso größer ausfallen, je höher die Ausgangswerte sind. In Mathematik zeigen sich in allen drei oberen Quartilen nahezu identische Effekte.

Die gegen Ende der dritten Jahrgangsstufe erreichten Effekte verschwinden am Schluss der vierten Jahrgangsstufe nicht nur, sondern sie kehren sich tendenziell sogar zugunsten der jahrgangshomogenen Klassen um (Tab. 4). Hiervon ausgenom-

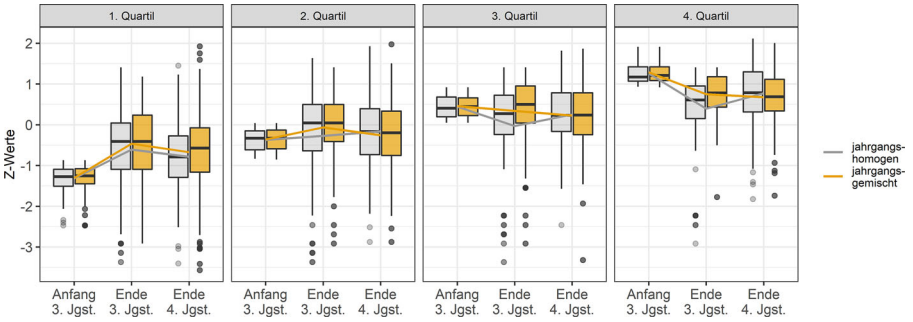

Abb. 1 Entwicklung der Leseleistung vom Anfang der dritten bis zum Ende der vierten Jahrgangsstufe (nach Leistungsquartilen getrennt)

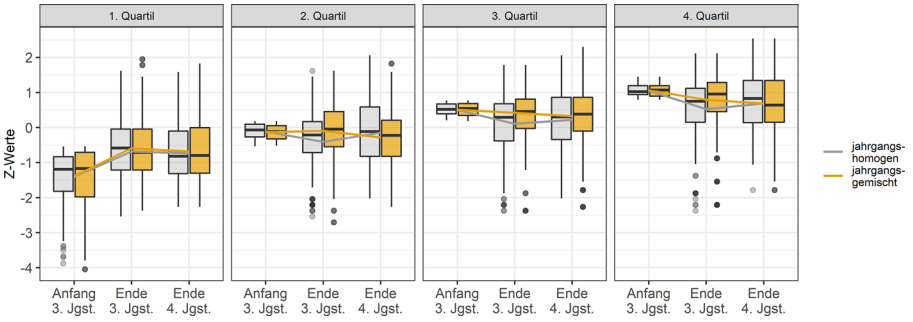

Abb. 2 Entwicklung der Mathematikleistung vom Anfang der dritten bis zum Ende der vierten Jahrgangsstufe (nach Leistungsquartilen getrennt)

men sind jedoch die niedrigsten Leistungsquartile in beiden Bereichen sowie das dritte Leistungsquartil in Mathematik.

Eine Überprüfung der beschriebenen Entwicklungen mittels der entsprechenden gemischten linearen Modelle für alle Leistungsquartile zu Lesen und Mathematik verdeutlicht zudem (Tab. 5 und 6), dass die Unterschiede in der Leistungsveränderung zwischen den jahrgangshomogenen und -gemischten Klassen in allen Teilgruppen bis auf die in den ersten Quartilen signifikant sind. Dies gilt sowohl für die Leistungszunahmen der Quartile in den jahrgangsgemischten Klassen bis zum Ende der dritten wie auch der -abnahmen bis zum Schluss der vierten Jahrgangsstufe, die sie *zusätzlich* zur Leistungsveränderung in den jahrgangshomogenen Klassen erfahren (vgl. die jeweiligen Interaktionseffekte „Anfang 3./Ende 4. Jahrgangsstufe × Jahrgangsmischung“).

Betrachtet man die Entwicklungsverläufe pro Quartil (Leseleistung: Abb. 1, Mathematikleistung: Abb. 2) so fallen weiterhin die breiten Leistungsspektren auf, die trotz ähnlicher Ausgangsleistungswerte (und des vorausgehenden Matchingalgorithmus) schon zum Ende der dritten wie auch der vierten Jahrgangsstufe wieder ersichtlich werden. Dies zeigt in beiden Unterrichtsformen, dass die Eingangsquartile keineswegs leistungsstabil sind, sondern jeweils im Verhältnis zur Gesamtgruppe individuelle Leistungssteigerungen als auch -abnahmen auftreten. Über den betrach-

teten Zeitraum hinweg verbleiben im Bereich Lesen durchschnittlich 40 %, im Bereich Mathematik im Mittel 45 % der Schülerinnen und Schüler in demselben Leistungsquartil; je 40 % verbessern oder verschlechtern sich um ein Quartil, 20 % bzw. 15 % sogar um mehrere. Ohne hier differenzierter auf individuelle Entwicklungsverläufe einzugehen, ist schließlich anzumerken, dass gerade die äußerst schwachen Mathematikleistungen, die zu Erhebungsbeginn noch sichtbar sind (mehr als drei Standardabweichungen unter dem Durchschnitt), bei beiden Unterrichtsformen in nachfolgenden Erhebungszeitpunkten nicht mehr vorkommen.

Eine nach Leistungsquartilen getrennte Betrachtung des Einflusses der Unterrichtsorganisationsform auf die Leistungsentwicklung von Schüler*innen stützt folglich Hypothese 3a, dass Schüler*innen der drei oberen Leistungsquartile gerade in der dritten Jahrgangsstufe von einem jahrgangsgemischten Unterricht profitieren können, und zwar im Lesen sogar umso mehr, je höher ihr vorheriges Ausgangsniveau ist. Auf die leistungsschwachen Schüler*innen (des ersten Quartils) trifft dies nur tendenziell zu. Entgegen der Hypothese 3b können sie aber auch in der vierten Jahrgangsstufe keinen beträchtlichen Vorteil aus jahrgangsgemischtem Unterricht gewinnen.

5 Diskussion und Zusammenfassung

Die Studie liefert einen Beitrag zum Vergleich der schulischen Leistungen zwischen jahrgangsgemischt und jahrgangshomogen unterrichteten Klassen, speziell für die dritte und vierte Jahrgangsstufe. Am letzten Messzeitpunkt (Ende der vierten Jahrgangsstufe und damit Ende der Grundschulzeit) lassen sich keine Unterschiede in der Lernleistung (mit den Ergebnissen aus Tests in Deutsch/Lesen und Mathematik als Indikatoren) zwischen den beiden Gruppen feststellen. Damit stärkt das Ergebnis den einschlägigen nationalen Befund aus der repräsentativen Ländervergleichsstudie (Kuhl et al. 2013). Aufgrund der internationalen inkonsistenten Befundlage ist dies aus unserer Sicht von Bedeutung. Anders sieht das Bild jedoch aus, wenn man die Lernentwicklung genauer betrachtet, da am Ende der dritten Jahrgangsstufe Kinder aus jahrgangsgemischten Klassen signifikant besser abschneiden. Wie erwartet, profitieren die Drittklässler*innen als die jüngeren Lernpartner*innen vom Altersgefälle in der Jahrgangsmischung (Leuven und Rønning 2011; Grittner et al. 2013; Campana Schleusener 2014). Erklärbar könnte dies u. U. durch zusätzliche, herausfordernde Angebote durch den spiralcurricularen Aufbau in vielen jahrganggemischt unterrichteten Klassen sein. Eine weitere mögliche Erklärung ergibt sich durch die Unterstützung im Austausch und die Hilfe, welche sie durch die Viertklässler*innen erfahren (Matz und Knauf 2010).

Aus didaktischer Sicht ist festzuhalten, dass sich die asymmetrische Peerstruktur jedoch nur für die jüngeren Schüler*innen und nicht in gleicher Form für die Viertklässler*innen zu „lohnen“ scheint: Die Hoffnung, dass diese Kinder z. B. durch Wiederholung oder durch ihr Erklären in der jahrgangsgemischten Variante mehr lernen, bestätigt sich zumindest in der generellen Analyse des Oberflächenstrukturmerkmals Jahrgangsmischung in seinen Effekten auf Leistung nicht. Eine mögliche Erklärung könnte in fehlenden, über die Jahrgangsstufe hinausführenden Inhalten

und Lernangeboten liegen, sodass die curricularen Vorgaben die Leistungsentwicklung deckeln und einen inhaltlichen Vorsprung gegenüber jahrgangshomogen unterrichteten Schüler*innen verhindern. Eine Rolle kann auch die spezielle Übertrittssituation spielen: Im hier untersuchten Bundesland Bayern ist ein bestimmter Notenschnitt erforderlich, um für den Besuch eines Gymnasiums oder einer Realschule zugelassen zu werden. Es ist denkbar, dass durch die damit zugeschriebene hohe Bedeutung schulischer Leistungen in den Fächern Deutsch und Mathematik (diese sind in beiden Organisationsformen identisch) Unterschiede zwischen jahrgangsgemischten und jahrgangshomogenen Klassen nivelliert werden. Feststellen lässt sich damit, dass das Potenzial der Jahrgangsmischung, das sich in der dritten Jahrgangsstufe gut zeigt, anschließend anscheinend (noch) nicht optimal genutzt wird.

Mit Blick auf die bislang offene Frage, ob bestimmte Leistungsgruppen besonders profitieren, zeigen unsere Daten Chancen für die jeweils jüngeren *und* leistungsstärkeren Schüler*innen. In der vorliegenden Studie ist der Effekt zugunsten der jahrgangsgemischten Klassen in der dritten Jahrgangsstufe auf alle Leistungsgruppen zurückzuführen, in besonderem Maße jedoch auf die leistungsstärkeren Quartile. Hier gibt es signifikante kleine bis mittlere Effekte, die im untersten Quartil ausbleiben. Gerade die Schüler*innen mit höheren Vorkenntnissen scheinen die Zusammenarbeit mit den älteren Schülern*innen nutzen zu können. Diese Effekte passen zu Befunden aus der ersten und zweiten Jahrgangsstufe (Gölitz 2008; Grittner et al. 2013) und bestärken die oben benannte Vermutung, dass die zusätzlichen Anregungen der höheren Jahrgangsstufe eine Ursache für die besseren Lernergebnisse der jahrgangsgemischten Gruppen sind – diese Anregungen sind vor allem für die Gruppe der leistungsstärkeren Kinder von Relevanz, die die Lernziele ihrer eigenen Jahrgangsstufe gut bzw. sehr gut erreichen und von daher auf weiterführende Angebote zugreifen können.

Umgekehrt dazu „verliert“ in der vierten Klasse lediglich das unterste Leistungsquartil in den jahrgangsgemischten Klassen *nicht* im Vergleich zur jahrgangshomogenen Gruppe. Gerade für diese Kinder könnte sich günstig auswirken, dass sie durch den Unterricht für die jüngeren Schüler*innen noch nicht verstandene Lerninhalte des letzten Jahres wiederholen können. Denkbar ist auch, dass die beratende und unterstützende Rolle, die diese Kinder für die jüngeren Kinder übernehmen, günstige Entwicklungen – evtl. unterstützt durch eine positive Entwicklung des Selbstkonzepts – anstößt.

Die Studie bestätigt, dass der differenzielle Blick auf die Effekte des jahrgangsgemischten Lernens von Bedeutung ist: Stärkere (differenzielle) Wirkungen zeigen sich in Übereinstimmung mit dem Forschungsstand für die jüngere Altersgruppe und für die leistungsstärkeren Schüler*innen (auch und gerade in dieser Kombination). Überraschend ist der Hinweis auf kleinere Gewinne der älteren und leistungsschwächeren Schüler*innen, die zumindest tendenziell am Ende der vierten Jahrgangsstufe besser in jahrgangsgemischten Klassen abschneiden.

6 Limitationen und Ausblick

Bei der Interpretation und Verallgemeinerung der voranstehenden Befunde gilt es abschließend einige Einschränkungen zu bedenken: Um Boden- bzw. Deckeneffekte zu vermeiden, war es erforderlich, zu den drei Messzeitpunkten unterschiedliche, jeweils curricular valide Verfahren einzusetzen. Die damit verbundene Unterschiedlichkeit der Roh- und Gesamtpunktwerte zwischen den drei Messzeitpunkten, erschwert jedoch zum Teil eine direkte Ergebnisinterpretation und den Nachvollzug individueller Entwicklungsverläufe der Schüler*innen, sodass nur durch eine *z*-Transformation der entsprechenden Werte deren relative Einordnung zu jedem Messzeitpunkt und eine vergleichende Betrachtung der jeweiligen Gruppenmittelwerte möglich war.

Aufgrund der organisatorischen und administrativen Rahmenvorgaben des Schulkontexts waren randomisierte Stichprobenziehungen von Klassen sowie Schüler*innen wie auch ein experimentelles Design nicht umsetzbar. Um dennoch eine gewisse Anzahl an Einflussgrößen zu kontrollieren, wurde ein Propensity-Score-Matchingverfahren eingesetzt, bei dem äquivalente Schüler*innenpaare aus jahrgangsgemischten und -homogenen Klassen gebildet wurden. Auch wenn dies aus methodischer Sicht eine gute Alternative zur Auswertung nicht randomisierbarer Studien darstellt (Kuss et al. 2016), ist bei Schlussfolgerungen und Kausalannahmen dennoch zu bedenken, dass in die Adjustierung nur tatsächlich gemessene Merkmale eingehen; darüber hinausgehende Ursachen für Unterschiede bleiben unberücksichtigt oder möglicherweise sogar unerkannt.

Des Weiteren gründete sich die Entscheidung zur Bildung von Leistungs*quartilen* zum ersten Messzeitpunkt auf forschungspragmatische (z. B. ausreichende Fallzahl pro Quartil), didaktische (z. B. übersichtliche Darstellbarkeit) und inhaltliche Gründe (Analyse des Entwicklungsverlaufs von schwächsten, unterdurchschnittlichen, überdurchschnittlichen und höchsten Ausgangsleistungen). Daneben wären auch andere Aufteilungen mit entsprechend leicht abweichenden Resultaten denkbar (z. B. Terzile, bei denen der fehlende Effekt in der schwächsten Gruppe durch die Vergröberung der Kategorien verdeckt wird, oder Quintile, bei denen aufgrund der Verfeinerung der Kategorien die zwei unteren keinen, die drei oberen Quintile – ähnlich zu Quartilen – Effekte aufweisen).

Da es sich bei der Maßnahme Jahrgangsmischung primär um ein unterrichtliches Merkmal der Oberflächenstruktur handelt, dessen Effekt auf der Qualität tiefenstruktureller Kriterien basiert (Hahn 2019), ist schließlich die in den einzelnen gemischten linearen Modellen durch diesen Prädiktor (als festen Effekt) erzielte Varianzaufklärung zwar relativ gering, aber durchaus erwartungskonform. Angesichts der großen Varianzanteile, die hingegen auf (variabel modellierte) Ausgangsniveaus von Schüler*innen wie auch Klassen zurückgeführt werden können, bleibt bislang allerdings ungeklärt, inwiefern weitere Merkmale von Schüler*innen (z. B. Rolle der jüngeren/älteren Lerner*innen, motivationale Facetten wie Selbstkonzept und Motivationsform), verschiedene unterrichtliche Prozesse (z. B. tutorielles Lernen, Lernen durch Lehren) und deren Prozessqualität (z. B. Gestaltung jahrgangsgemischten Unterrichts, curriculare Inhaltsanordnung) oder Merkmale der Lehrkräfte (z. B. professionelles Wissen, Einstellungen, Motivation für jahrgangsgemischten Unter-

richt, Erfahrung) von besonderer Relevanz sind. Lohnen könnten auch Analysen aus fachdidaktischer Perspektive, die die Unterschiede in der Gestaltung im Fach Mathematik und Deutsch im jahrgangsgemischten Unterricht in den Blick nehmen. Diese Aspekte standen nicht im Fokus des vorliegenden Beitrags, zentrale Aspekte sollen aber in weiterführenden Analysen und Studien mehr Beachtung erhalten und in ihrer differenziellen Bedeutung systematisch untersucht werden. Gleiches gilt für Daten zum wahrgenommenen Leistungsdruck, zur Motivation sowie zu Hoffnungen und Befürchtungen bezüglich des Übertritts. Diese sollen – auch im Konnex zur Leistungsentwicklung – weiter ausgewertet werden, um mögliche Zusammenhänge an dieser im Bildungssystem höchst bedeutsamen Gelenkstelle aufzuspüren.

Funding Open Access funding enabled and organized by Projekt DEAL.

Literatur

Austin, P.C. (2010). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical statistics*, *10*(2), 150–161.

Barton, K. (2020). MuMIn: Multi-Model Inference. R package version 1.43.17. https://CRAN.R-451 project.org/package=MuMIn. Zugegriffen: 15. Dez. 2020.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*(1), 133–180.

Berthold, B. (2010). *Sekundäranalytische Rekonstruktion entwicklungskritischer Kernaufgaben und Verlaufsmuster der Unterrichtsentwicklung bei der Einrichtung der integrativen, jahrgangsgemischten und flexiblen Schuleingangsphase*. Bremen: Universität Bremen.

Bliese, P. (2016). multilevel: multilevel functions. R package version 2.6. https://CRAN.R-project.org/package=multilevel. Zugegriffen: 15. Dez. 2020.

Bortz, J., Lienert, G.A., & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik: mit 247 Tabellen* (3. Aufl.). Heidelberg: Springer.

Brahm, G. (2006). Klassengröße: eine wichtige Variable von Schule und Unterricht? *Bildungsforschung*. https://doi.org/10.25656/01:4654.

Campana Schleusener, S. (2014). Wenn Lernen und Lehren zusammentreffen: gegenseitiges Helfen in heterogenen Klassen. In B. Kopp, S. Martschinke, M. Munser-Kiefer, M. Haider, E.-M. Kirschhock, G. Ranger & G. Renner (Hrsg.), *Individuelle Förderung und Lernen in der Gemeinschaft* (S. 166–169). Wiesbaden: Springer.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Cornish, L. (2010). Multiage classes: what's in a name? *Journal of Multiage Education*, *4*(2), 7–11.
Decristan, J., Hess, M., Holzberger, D., & Praetorius, A.-K. (2020). Oberflächen- und Tiefenmerkmale – Eine Reflexion zweier prominenter Begriffe der Unterrichtsforschung. *Zeitschrift für Pädagogik 66. Beiheft*, *1*, 102–116.
Ditton, H. (2019). Mechanismen der Selektion und Exklusion im Schulsystem. In G. Quenzel & K. Hurrelmann (Hrsg.), *Handbuch Bildungsarmut* (S. 157–181). Wiesbaden: Springer VS.
Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford.
Feuchtenberger, S., Martschinke, S., Munser-Kiefer, M., & Hartinger, A. (2019). „Mehr Zeit für einzelne Kinder" oder „mehr Stress" – Eine Interviewstudie zu Chancen und Risiken jahrgangsgemischten Lernens in der dritten und vierten Jahrgangsstufe aus der Perspektive von Lehrkräften. In C. Donie, F. Foerster, M. Obermayr, A. Deckwerth, G. Kammermeyer, G. Lenske, M. Leuchter & A. Wildemann (Hrsg.), *Grundschulpädagogik zwischen Wissenschaft und Transfer* (S. 263–269). Wiesbaden: Springer.
Gölitz, D. (2008). *Profitieren Kinder mit kognitiven Entwicklungsrisiken von jahrgangsgemischtem Schulanfangsunterricht?* Göttingen: Georg-August-Universität Göttingen.
Grittner, F., Hartinger, A., & Rehle, C. (2013). Wer profitiert beim jahrgangsgemischten Lernen? *Zeitschrift für Grundschulforschung*, *6*(1), 102–113.
Guo, S., & Fraser, M. W. (2015). *Propensity score analysis: statistical methods and applications* (2. Aufl.). Thousand Oaks: SAGE.
Gutiérrez, R., & Slavin, R. E. (1992). Achievement effects of the nongraded elementary school: a best evidence synthesis. *Review of Educational Research*, *62*(4), 333–376.
Hahn, E. (2019). *Umgang mit Heterogenität an Gemeinschaftsschulen: eine multimethodische Untersuchung zu Oberflächen- und Tiefenstrukturen des Unterrichts* (6. Aufl.). Münster, New York: Waxmann.
Hartinger, A., Graumann, O., & Grittner, F. (2004). „Grundschul-Numerus Clausus" oder Orientierungsstufe? Auswirkungen verschiedener Übertrittsbedingungen auf Motivationsstile und Leistungsängstlichkeit von Grundschulkindern. *Empirische Pädagogik*, *18*(2), 173–193.
Hartinger, A., Grittner, F., & Rehle, C. (2011). *Gibt es Mathäus-Effekte vom Jahrgangsgemischten Lernen? (Vortrag auf der 20. Jahrestagung der Kommission für Grundschulforschung und Pädagogik der Primarstufe, 22.09.2011)*. Paderborn: DGfE, Universität Paderborn.
Hilbert, S., Stadler, M., Lindl, A., Naumann, F., & Bühner, M. (2019). Analyzing longitudinal intervention studies with linear mixed models. *TPM: Testing, Psychometrics, Methodology in Applied Psychology*, *26*(1), 101–119.
Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). Matchit: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28.
ISB (Staatsinstitut für Schulqualität und Bildungsforschung München) (2014an). Probeunterricht 2014 an den Gymnasien in Bayern Mathematik – Jahrgangsstufe 4. *Staatsinstitut für Schulqualität und Bildungsforschung*. https://www.isb.bayern.de/download/15525/probeunterricht_2014_mathematik_jahrgangsstufe_4_tag_1.pdf. Zugegriffen: 14. Dez. 2020.
Kammermeyer, G., & Martschinke, S. (2006). Selbstkonzept- und Leistungsentwicklung in der Grundschule–Ergebnisse aus der KILIA-Studie. *Empirische Pädagogik*, *20*(3), 245–259.
Klöver, B. (2014). *Evaluationsbericht Flexible Grundschule*. München: Staatsinstitut für Schulqualität und Bildungsforschung.
Krüsken, J. (2008). Schülerleistungen in FLEX-Klassen bei den Vergleichsarbeiten Jahrgangsstufe 2 in Brandenburg in den Jahren 2004 bis 2006. In K. Liebers, A. Prengel & G. Bieber (Hrsg.), *Die flexible Schuleingangsphase: Evaluationen zur Neugestaltung des Anfangsunterrichts* (S. 30–56). Weinheim: Beltz.
Kuhl, P., Felbrich, A., Richter, D., Stanat, P., & Pant, H. A. (2013). Die Jahrgangsmischung auf dem Prüfstand: Effekte jahrgangsübergreifenden Lernens auf Kompetenzen und sozio-emotionales Wohlbefinden von Grundschülerinnen und Grundschülern. In R. Becker & A. Schulze (Hrsg.), *Bildungskontexte* (S. 299–324). Wiesbaden: Springer.
Kuss, O., Blettner, M., & Börgermann, J. (2016). Propensity score: an alternative method of analyzing treatment effects—part 23 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt*, *113*(35/36), 597–603.
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmertest package: tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.
Laging, R. (2010). Altersheterogenität und Helfen – eine Untersuchung in der Schuleingangsstufe der Reformschule Kassel. In R. Laging (Hrsg.), *Altersgemischtes Lernen in der Schule* (Bd. 4, S. 54–71). Baltmannsweiler: Schneider-Verl. Hohengehren.

Leuven, E., & Rønning, M. (2011). *Classroom grade composition and pupil achievement*. IZA DP No. 5922.

Lindström, E.-A., & Lindahl, E. (2011). The effect of mixed-age classes in Sweden. *Scandinavian Journal of Educational Research*, *55*(2), 121–144.

LISUM (Landesinstitut für Schule und Medien, Berlin-Brandenburg) (2008). Individuelle Lernstandsanalysen. Schülerheft Mathematik 3. *Bildungsserver Berlin-Brandenburg*. https://bildungsserver.berlin-brandenburg.de/fileadmin/bbb/unterricht/lernbegleitende_Diagnostik/ilea/2010/Mathe3Schueler.pdf. Zugegriffen: 14. Dez. 2020.

Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, *110*, 63–73.

Matz, S., & Knauf, T. (2010). Altersmischung in der Praxis einer Montessori-Schule – eine Beobachtungsstudie zur Auftretenshäufigkeit ausgewählter Aspekte altersgemischter Lerngruppen. In R. Laging (Hrsg.), *Altersgemischtes Lernen in der Schule* (Bd. 4, S. 72–79). Baltmannsweiler: Schneider.

Ministerium für Kultus, Jugend und Sport Baden-Württemberg (2006). *Schulanfang auf neuen Wegen: Abschlussbericht zum Modellprojekt in Baden-Württemberg*. Stuttgart: Kultusministerium Baden-Württemberg.

Munser-Kiefer, M., Martschinke, S., & Hartinger, A. (2017). Adaptive Unterrichtsgestaltung und Überzeugungen von Lehrpersonen in jahrgangsgemischten und jahrgangshomogenen Klassen. *Zeitschrift für Grundschulforschung*, *10*(1), 147–161.

Pape, M. (2016). *Didaktisches Handeln in jahrgangsheterogenen Grundschulklassen: eine qualitative Studie zur Inneren Differenzierung und zur Anleitung des Lernens*. Bad Heilbrunn: Verlag Julius Klinkhardt.

Perren, S., & Malti, T. (2016). Soziale Kompetenz entwickeln: Synthese und Ausblick. In T. Malti & S. Perren (Hrsg.), *Soziale Kompetenz bei Kindern und Jugendlichen: Entwicklungsprozesse und Förderungsmöglichkeiten* (S. 284–294). Stuttgart: Kohlhammer.

Praetorius, A.-K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: looking back and looking forward. *ZDM*, *50*(3), 535–553.

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality. The German framework of Three Basic Dimensions. *ZDM Mathematics Education*, *50*(3), 407–426.

Praetorius, A.-K., Grünkorn, J., & Klieme, E. (2020). Empirische Forschung zu Unterrichtsqualität. Theoretische Grundfragen und quantitative Modellierungen. *Zeitschrift für Pädagogik*, *66*, 9–14.

R Core Team (2020). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Ronksley-Pavia, M., Barton, G., & Pendergast, D. (2019). Multiage education: an exploration of advantages and disadvantages through a systematic review of the literature. *Australian Journal of Teacher Education*, *44*(5), 24–41.

Russel, V. J., Rowe, K. J., & Hill, P. W. (1998). Effects of multigrade classes on student progress in literacy and numeracy: quantitative evidence and perceptions of teachers and school leaders. Victoria (Australia). http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/16/61/a2.pdf;. Zugegriffen: 23. Dez. 2020.

Ryan, R. M., & Connell, J. P. (o. J.). *Self-Regulation Questionnaires*. Abrufbar unter https://selfdeterminationtheory.org/self-regulation-questionnaires/.

Sonnleitner, M. (2020). Jahrgangsmischung aus Sicht von Schulleitungen und Lehrkräften. *Zeitschrift für Grundschulforschung*, *13*(2), 357–371.

Sonnleitner, M. (2021). *Schule entwickeln: Jahrmischung aus der Perspektive professionell Handelnder*. Bad Heilbrunn: Klinkhardt Forschung.

Sundell, K. (1994). Mixed-age groups in Swedish nursery and compulsory schools. *School Effectiveness and School Improvement*, *5*(4), 376–393.

Thoren, K. (2017). *Implementationserfolg von Schulreformen in der Berliner Schulanfangsphase*. Berlin: Freie Universität Berlin.

Thoren, K., & Brunner, M. (2019). Flächendeckende Implementation des Jahrgangsübergreifenden Lernens: Welche Typen gibt es und zeigen diese Unterschiede in der Schul- und Unterrichtsqualität? *Zeitschrift für Erziehungswissenschaft*, *22*(2), 279–300.

Veenman, S. (1996). Effects of multigrade and multi-age classes reconsidered. *Review of Educational Research*, *66*(3), 323–340.

von Waaden, S. (2017). *Mathematiklernen von Risikokindern in der Jahrgangsmischung*. Wiesbaden: Springer.

Wang, Y., Cai, H., Li, C., Jiang, Z., Wang, L., Song, L., & Xia, J. (2013). Optimal caliper width for propensity score matching of three treatment groups: a Monte Carlo study. *PLoS ONE*, *8*(12), e81045. https://doi.org/10.1371/journal.pone.0081045.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2. Aufl.). New York: Springer.

Wisniewski, B., Zierer, K., Dresel, M., & Daumiller, M. (2020). Obtaining secondary students' perceptions of instructional quality: Two level structure and measurement invariance. *Learning and Instruction*, *66*, 101303. https://doi.org/10.1016/j.learninstruc.2020.101303.