Hartmut Esser
(Universität Mannheim)

Effekte der Zweigliedrigkeit

Die Wirkung von Zusammenlegung, Öffnung und Lockerung der Schulstrukturen auf Leistungsniveau und Bildungsungleichheit in der Sekundarstufe in den deutschen Bundesländern.

(13042025)

Effects of Two-Tier Systems

The impacts of merging, opening and releasing school structures on achievement and educational inequality at secondary level in the German federal states.

Zusammenfassung/Abstract

Deutschland ist international eines der wenigen Länder noch mit einem differenzierenden Bildungssystem. Allerdings gibt es innerhalb dieses Rahmens deutliche Unterschiede zwischen den Bundesländern bei den Regelungen für die vorgesehenen Schulstrukturen und die Implementation der Differenzierung. Sie beziehen sich im Wesentlichen zwei Aspekte: die Anzahl der Optionen neben dem Gymnasium, Drei- oder Zweigliedrigkeit, und die Stringenz, in der der Zugang zu Bildungswegen geregelt und die schulischen Abläufe organisiert sind. Im Zuge der Ergebnisse der internationalen Vergleichsstudien seit PISA 2000 gab es, verstärkt etwa seit 2012, einen deutlichen Trend zur Zusammenlegung der unteren Bildungswege und zur Öffnung und Lockerung des Zugangs nach oben wie bei der Strenge der Anforderungen. Damit sollte den negativen Folgen der Differenzierung und der Zunahme der gesellschaftlichen Heterogenität entgegengewirkt werden. Der Beitrag untersucht die Effekte dieser Entwicklungen auf die Leistungen in der Sekundarstufe nach dem Übergang von der Grundschule. Es gibt Befunde zu zwei Aspekten: Die Ergebnisse von Zweigliedrigkeit und Stringenz im inter- und innernationalen Vergleich nach den PISA- und IQB-Berichten von 2000 bis 2022 und multivariate Analysen der Effekte von Zweigliedrigkeit und Stringenz in Vergleich und in Kombination mit den Daten der „National Educational Panel Study“ (NEPS). Die Befunde bestätigen die Hypothesen von den positiven Wirkungen der Zusammenlegung und der Öffnung nicht. Es sieht eher danach aus, dass die stringenten und gleichzeitig dreigliedrigen Systeme, gerade auch im internationalen Vergleich, leistungsfähiger, leistungsgerechter und gegen die Krisen seit 2020 resilienter gewesen sind, auch schon in den Grundschulen. Bei den wenigen Untersuchungen, die positive Effekte berichten, gibt es Fragen an die Angemessenheit des Ansatzes. Zu weiteren Öffnungen, welcher Art auch immer, geben die Befunde bis zu deren Klärung jedenfalls keinen Anlass.

Germany is one of the few countries in the world that still has a differentiated education system. However, within this framework, there are clear differences between the federal states in the regulations for the provided school structures and the implementation of differentiation. They essentially relate to two aspects: the number of options in addition to the Gymnasium, three- or two-tier system, and the stringency with which access to educational pathways is regulated and school processes are organized. In the course of the results of the international comparative studies since PISA 2000, there has been a clear trend towards merging the lower educational pathways and opening up and loosening access to the upper levels, as well as the stringency of requirements, which has intensified since 2012. This was intended to counteract the apparently negative consequences of differentiation and the increase in social heterogeneity. This article examines the effects of these developments on secondary school achievement after the transition from elementary school. There are findings on two aspects: The results for the two-tier systems and stringency of the federal states in inter- and intra-national comparison according to the PISA and IQB reports from 2000 to 2022 and multivariate analyses of the effects of two-tier system and stringency in comparison and in combination with the data of the “National Educational Panel Study” (NEPS). The findings do not confirm the hypotheses of the positive effects of merging and opening up. It looks more likely that the stringent and at the same time tripartite systems have been more efficient, more performance-oriented and more resilient to the crises since 2020, especially in an international comparison, even in elementary school. In the few studies that report positive effects, there are serious questions about the appropriateness of the approach to identifying the effects. In any case, the findings do not give rise to any further openings of any kind.

Ability-Tracking, Mulit-tiered school systems. Achievement, Educational Inequality, German federal states

Declaration of Interests

Es bestehen keinerlei konkurrierende Interessen

1. Die Befunde für PISA 2022 bedeuteten auch für die Bildungspolitik in Deutschland eine Art von Zeitenwende: Nie waren das Niveau in den Leistungen und das Erreichen der Mindeststandards so niedrig, nicht einmal beim Schock von 2000. Die Situation ist nicht neu, denn so ähnlich hatte es in Deutschland schon einmal, bei PISA 2000, ausgesehen. Damals wurde gerade auch an der Kompensation von Nachteilen bei den Kindern in den schwierigen Verhältnissen angesetzt: Förderung des Vorschulbesuchs, Verbesserungen der Unterrichtsqualität, regelmäßiges Monitoring und eine plötzlich hohe öffentliche Aufmerksamkeit – und das zeigte tatsächlich alsbald Wirkungen (vgl. Klieme et al. 2010): von 2006 an hatte sich Deutschland nach und nach international stark verbessert und dabei sogar einige der Vorzeigeländer mit ihren integrierten Systemen in den Leistungen und der sozialen Durchlässigkeit eingeholt und sogar übertroffen (vgl. die Übersicht in Tabelle 3 unten).

Nahezu parallel gab es jedoch eine andere Entwicklung noch. Weil vermutet worden war, dass die schlechte Positionierung bei PISA 2000 mit dem gerade für Deutschland so typischen Bildungssystem der frühen Differenzierung in drei Schultypen und Bildungswegen zu tun hätte, gab es Bestrebungen für eine Annäherung an die „Integration“ ohne Differenzierung in der Pflichtschulzeit. Das erfolgte im Wesentlichen über zwei Entwicklungen: Die Zusammenlegung der für besonders problematisch angesehenen und immer mehr gemiedenen Hauptschulen mit den Realschulen zu „Gemeinschaftsschulen“ oder „Sekundarschulen“ und die Öffnung des Zugangs über die Abschaffung der Verbindlichkeit der schulischen Empfehlungen und die Lockerung der organisatorischen Kontrollen in den schulischen Abläufen. Das setzte etwa ab 2012 ein. In der Folge stagnierte das Niveau der Leistungen und sank sogar über 2015 und 2018, der letzten PISA-Erhebung unter regulären Verhältnissen vor den Krisen (vgl. Wößmann 2023). In der ZEIT wurde das Phänomen des absehbaren neuen PISA-Absturzes 2021 recht früh schon als „trauriger Smiley“ bezeichnet. Nicht nur auf den ersten Eindruck hin hätte man also vermuten können, dass der Abschwung nach 2012 etwas mit diesen Umstellungen hin zur Quasi-Integration einer zweizügigen und liberalisierten Differenzierung zu tun haben könnte. Dafür gab es, ebenfalls kaum einmal bemerkt, schon bei PISA 2000 aus den Berichten der PISA-Sonderauswertungen für die 16 deutschen Bundesländer einige deutliche Hinweise: Die Bundesländer mit den striktesten Regelungen der Differenzierung hatten ein deutlich höheres Leistungsniveau, schon nahe und sogar über dem OECD-Durchschnitt, und eine keineswegs stärkere soziale Stratifikation. Das waren damals Baden-Württemberg, Bayern und Sachsen. Diese drei Bundesländer hatten zudem alle ein

dreigliedriges System mit eigenen Abschlüssen für die Haupt- und die Realschulen. Und sie hatten auch schon seit längerer Zeit keine öffnenden Reformen erlebt.

Die schockierenden Ergebnisse fanden sich also schon bei PISA 2000 gerade *nicht* für die Bundesländer mit der traditionell strikten Differenzierung wie sie Deutschland allgemein als Archetypen dafür zugeschrieben wurden, sondern dort, wo es die Öffnungen schon gab. Das setzte sich in einem so wohl nicht beabsichtigten Feldexperiment fort: Baden-Württemberg schaffte 2012/13 die Verbindlichkeit ab, lockerte die Dreigliedrigkeit mit der Einrichtung von „Gemeinschaftsschulen" parallel zu Haupt- und Realschulen, die etwas umetikettiert wurden – und schied bald danach aus der Spitzengruppe aus. Sachsen und Bayern blieben dagegen bei ihren Strukturen, zogen ungerührt weiter davon, auch international, und lagen 2018, dem letzten regulären Jahr vor den Krisen, sogar *vor* einst so hochgelobten Ländern wie Finnland und Schweden, auch Kanada – und 2022, mit den Krisen, nur wenig hinter Estland, dem Überraschungssieger bei der bisher letzten PISA-Runde (s. dazu noch Tabelle 4 unten)

Um dieses, von Öffentlichkeit, Politik und auch der Bildungsforschung kaum bemerkte Puzzle geht es in dem folgenden Beitrag. Es ist die Frage nach den Effekten der Zweigliedrigkeit im Vergleich zur Dreigliedrigkeit und das dann gerade auch in Bezug und Kombination mit der Stringenz der Regelungen der Differenzierung nach Verbindlichkeit und Kontrolle in den deutschen Bundesländern. Dazu gibt es zunächst einen Überblick über die Forschungslage (Abschnitt 2). Daran anschließend wird die theoretische Grundlage zu den verschiedenen Positionen skizziert (Abschnitt 3). Dann kommen in Abschnitt 4 der Untersuchungsansatz und in Abschnitt 5 die Datengrundlage, die der „National Educational Panel Study" (NEPS), sowie die Zuordnung der Bundesländer zu den jeweiligen Systemregelungen von Mehrgliedrigkeit und Stringenz. Die Analysen und die Befunde beziehen sich daran anschließend auf zwei Aspekte (Abschnitt 6): die einfache deskriptive Positionierung der deutschen Bundesländer untereinander und im internationalen Vergleich nach den jeweiligen PISA- und IQB-Berichten von PISA 2000 bis 2022 und auf die Ergebnisse von nun auch multivariaten Analysen zur Zweigliedrigkeit auf die Leistungen, in den Grundschulen schon und in der Sekundarstufe, jeweils in Relation und Konditionalisierung zur Stringenz, in der die Differenzierung in den deutschen Bundesländern institutionell geregelt ist. In Abschnitt 7 kommen einige Anmerkungen zu den Limitationen und in Abschnitt 8 gibt es noch eine kurze zusammenfassende Bewertung.

2. Der Stand der Dinge

Differenzierung und Integration bilden die Pole eines Kontinuums: die volle Integration ohne jede Differenzierung bis zum Ende der Pflichtschulzeit gegenüber der frühen und strikten, auch räumlichen Differenzierung. Daneben gibt es Kombinationen mit jeweils unterschiedlicher Dominanz: Differenzierungen in der Integration, etwa „Gesamtschulen“ mit interner Differenzierung oder integrative Öffnungen in der Differenzierung, wie bei der Zusammenlegung von Schultypen, der Verlängerung des „gemeinsamen Lernens“ oder Öffnungen und Lockerungen in Zugang, Standards und Organisation (vgl. aktuell dazu Schindler et. al. 2024.). Dazu gehören dann auch informelle Strukturierungen, etwa über soziale und kognitive Netzwerke innerhalb von Schulklassen, die sich über die formellen Einteilungen legen können und informelle Trennungen gerade da erzeugen, wo es für das Lernen am relevantesten ist: in den Schulklassen und den (peer-)Interaktionen dort (vgl. Dollmann und Rudolphi 2022, Engzell und Raabe 2024).

Differenzierung und Integration

Die gängige Auffassung für die Unterschiede einfach zwischen Differenzierung und Integration jeweils allein für sich ist, dass mit der Trennung der Bildungswege und Schulen das Leistungsniveau nicht steige, und wenn, dann nur bei den talentierteren Kindern und denen aus den oberen sozialen Schichten. Mit der Integration wäre das anders. Daher sei die Umstellung dahin angeraten, speziell unter Bedingungen der Zunahme von gesellschaftlicher Diversität und Heterogenität in den Schulen. Diese Position stützt sich im Wesentlichen auf Befunde der internationalen Vergleichsstudien, insbesondere PISA, und auf Analysen zu integrativen Reformen insgesamt.

Nach den *Vergleichsstudien* scheint es daran keinen Zweifel zu geben, speziell nicht für Deutschland, das von Beginn an als Prototyp von strikter Differenzierung und einer nahe ständestaatlichen Bildungsungleichheit gilt (vgl. u.a. Hanushek und Wößmann 2006, Gamoran 2009, Wößmann et al. 2009, van de Werfhorst und Mijs 2010, Chmielewski 2014, Wößmann 2016, Skopek et al. 2019, Roller and Steinberg 2022, Strello et al. 2022, Terrin und Triventi 2022, Dräger et al. 2023). Das Problem bei so gut wie allen diesen Vergleichen ist, dass zentrale Bedingungen für die Erklärung des Kompetenzerwerbs gefehlt haben, wie die kognitiven

Fähigkeiten, die Leistungen vorher in der Grundschule und die Effekte der kognitiven Komposition der Schulklassen, und dass daher die Schätzung der Effekte verzerrt sein dürften oder die Unterschiede schon vorher entstanden waren (Waldinger 2007, Dunne 2010, Bol et al. 2014, Dronkers und Skopek 2015, Cord and Giuliano 2016, Korthals and Dronkers 2016). Es ist daher nicht ohne Risiko, in internationalen Vergleichen zu definitiven Schlüssen über Effekte der Differenzierung zu kommen, zumal sich die Bedingungen nach Phasen des Bildungsverlaufs, für Bildungsbeteiligung und Leistungen und bei späteren Entwicklungen, etwa auf dem Arbeitsmarkt, ändern können (s. die Analysen von Jackson und Jonsson 2013 für die Bildungsbeteiligung, die Meta-Analysen bei Terrin und Triventi (2022) oder den Vergleich bei Schindler et al. 2024 für die höhere Sekundarstufe und den Arbeitsmarkt).

Das gilt ähnlich für die Untersuchungen zu den *Reformen* (Meghir und Palme 2005 für Schweden, Malamud und Pop-Eleches 2011 für Rumänien, Kerr et al. 2013 für Finnland, Guyon et al. 2012 für Nordirland, Jakubowski et al. 2016 für Polen oder van de Werfhorst 2018). Sie alle weisen – mehr oder weniger –auf *positive* Effekte der Integration hin, weniger allerdings zum Niveau der Leistungen als zu den Effekten der sozialen Herkunft. Ob das tatsächlich so ist, weiß man jedoch nicht definitiv. Die Bestimmung der Reformeffekte unterliegt generell einigen Problemen, die zur Vorsicht mahnen: In den Untersuchungen werden meist erst spätere Phasen des Lebenslaufs erfasst, es gibt zur Vorgeschichte und damit zur Selektivität beim Übergang keine Informationen, oft sind mit der Reform mehrere Änderungen gleichzeitig verbunden, deren Effekte nicht zu trennen sind, und es treten so gut wie immer Nebenwirkungen aus dem Vorgang der Umstellung selbst auf, die sog. Reformeffekte.

Das gilt für zwei Studien nicht. Die eine vermeidet mögliche Verzerrungen über die Randomisierung der Zuordnung bei der Differenzierung und findet positive Effekte der Differenzierung: Duflo et al. (2011). Sie ist die bisher einzige mit eindeutig *positiven* Effekten der *Differenzierung*. Aber es gibt auch hier Fragen: an die externe Validität, an die genauen Mechanismen und an mögliche Reformeffekte beim Einsatz des Lehrpersonals (Cummins 2016). Die andere Analyse ist die von Galindo-Rueda und Vignoles (2004) zu Reformen in England und Wales. Sie umfasst auch die Vorgeschichte und die Selektivität der Aufteilung auf die Bildungswege und damit auch Systemeffekte der Antizipation von möglichen Folgen nach dem Übergang. Danach verbessert die Differenzierung die Leistungen gerade der talentierteren Kinder, schadet aber im unteren Bereich der kognitiven Fähigkeiten nicht. Alles

in allem sprechen die Befunde für die Annahme, dass die Differenzierung die Leistungen für alle verbessern kann, wenngleich in der Tat dann nach oben relativ mehr.

Differenzierung bei Integration

Nun die Differenzierung bei Integration. Das betrifft die internen Differenzierungen bei intergierten Schulen, teils institutionalisiert über Leistungskurse, teils informell gebildet über Netzwerke innerhalb der Schulklassen, nach sozialer Herkunft und kognitiven Fähigkeiten vor allem, und teils als eher verborgene Praxis, wenn es Schwierigkeiten mit der Heterogenität in den Schulklassen gibt, wie in Estland (Esser 2024a). Für Effekte aller drei Formen gibt es Hinweise: Über die interne Differenzierung werden – wenigstens teilweise – Trennungen wie bei der externen Differenzierung innerhalb der Schulen und Schulklassen wieder eingeführt, speziell nach der sozialen Herkunft (Lucas 1999) und es bilden sich innerhalb der heterogenen Schulklassen nach Fähigkeiten und Herkunft getrennter Netzwerke (Dollmann und Rudolphi 2021, Engzell und Raabe 2023). Das schwächt die evtl. Effekte und zwar so weit, dass es kaum noch etwas ausmacht.

Integration bei Differenzierung

Die Integration bei Differenzierung nimmt ebenfalls verschiedene Formen an. Manche sind verdeckt, wie die Senkung der Anforderungen, manche über Parallelsysteme institutionalisiert wie bei den Integrierten Gesamtschulen in differenzierten Systemen, bei der Verschiebung des Alters bei der ersten Trennung, bei der Zusammenlegung von Bildungswegen bis auf den Rest eines Zweisäulenmodells oder aber auch die Schwächung von Regelungen bei der Implementation der Differenzierung, wie die der Stringenz nach Verbindlichkeit und Kontrolle.

Für die *Integrierten Gesamtschulen* wie es sie in einigen der deutschen Bundesländer innerhalb der übergreifend geltenden Differenzierung gibt finden sich auch nach etwa 50 Jahren an Erfahrung und Forschung keine wirklich überzeugenden Hinweise, dass sie die Probleme des Leistungsrückgangs oder der sozialen Stratikationen hätten können. Das Niveau bewegt sich nach Baumert et al. (2006) zwischen dem von Haupt- und Realschulen, und die soziale Durchlässigkeit ist auch nicht geringer als bei dreigliedriger Differenzierung (Lauterbach und Fend 2016, Fend 2017). Das entspricht dem, was Lucas (1999) für die USA und die interne Differenzierung dort gefunden hatte (s. oben schon dazu). Für die Niederlande zeigt van de

Werfhorst (2022), dass die „mixed schools“, die niederländische Variante der Integrierten Gesamtschulen, nicht nur nicht helfen, sondern eher eine Art von Mobilitätsfalle darstellen: Gerade die talentierten Kinder aus den unteren Schichten werden mit der ihnen leichter scheinenden Option der Gesamtschulen von dem riskanteren, aber zu höheren Leistungen führenden Weg nach oben in den differenzierenden Schulen abgelenkt. Das wiederum passt zu den Befunden bei Galindo-Rueda und Vignoles (2004) zu den Reformen in England und Wales und den Unterschieden zwischen Differenzierung und Integration, von denen oben auch schon die Rede war.

Zum *Alter bei der ersten Trennung* gibt es zwei Studien. In einer Untersuchung von Horn (2013) werden Schuleinzugsdistrikte mit Unterschieden beim Alter der ersten Trennung verglichen. Es geht um die Bildungsbeteiligung und die Leistungen in der Sekundarstufe. Die Bildungsbeteiligung über die Wahl einer *früh* differenzierenden Schule folgt, wenn alle relevanten Einflüsse kontrolliert werden, den kognitiven Fähigkeiten und darin unterscheiden sich die talentierten Kinder aus den unteren Schichten nicht von den anderen. Und die Leistungen sind bei den *früh* differenzierenden besser. Die Effekte nach der sozialen Herkunft unterscheiden sich dabei nicht. Das entspricht der gängigen Meinung von den fraglos positiven Effekten der integrativen Aufschiebung der Differenzierung nicht.

Die zweite Untersuchung zum Alter beim ersten Übergang stammt von Wößmann (2010) in einem Vergleich von Brandenburg und Berlin, die erst mit 12 Jahren trennen und nicht schon mit 10 wie die anderen Bundesländer. Der Zusammenhang ist danach negativ: Je *später* der Übergang erfolge (Berlin und Brandenburg), um so *geringer* wäre die soziale Bildungsungleichheit bei der Bildungsbeteiligung. Bei den Leistungen gebe es keine Unterschiede. Das würde die aus den OECD-Vergleichen bekannten Befunde von den positiven Effekten der Integration bestätigen. Zu beachten ist dabei jedoch, dass es sich bei der Analyse nur um *Aggregat*daten handelt, bei denen (nach dem Ko-Varianztheorem) vorausgesetzt werden müsste, dass es *keine* Interaktionseffekte zwischen den individuellen Beziehungen und den Systemeigenschaften gibt. Aber gerade das wird ja angenommen: Es gebe *System*effekte der *Moderation* der individuellen Zusammenhänge, etwa der sozialen Herkunft auf Bildungsbeteiligung und Leistungen. Die Befunde von Wößmann (2010) sind, wie viele der frühen ökonometrischen Untersuchungen mit PISA-Daten, offenbar ein Artefakt: fehlende Daten für die Individualebene, unerlaubte Fixierung der Systemeffekte und ein darüber entstehender „ökologischer Fehlschluss“.

Es sieht danach also keineswegs so aus als würden integrative Elemente innerhalb der Differenzierung eine Besserung bewirken. Teilweise schon ältere Analysen zu den Effekten von Differenzierung und Integration auf der Ebene von *Einzelschulen* (ohne besondere Bindungen an übergreifende Regelungen) bestärken das: Wenn auch die internen Differenzierungen einbezogen werden, schneiden die differenzierenden Schulen *besser* ab als die integrierten (Betts und Shkolnick 2000 und Figlio und Page 2002; Esser 2023, Abschnitte 3.2 und 3.2 zu der längeren Debatte darüber). Dabei kommt es darauf an, dass bei der Einrichtung der Differenzierung auf eine besondere „Selektivität" geachtet wird, also Fehlplatzierungen bei der Aufteilung möglichst vermieden werden (Korthals und Dronkers 2016; vgl. dazu noch Abschnitt 2 ausführlich).

(Quasi-)Integration: Differenzierung, Zweigliedrigkeit, Öffnung und Lockerung

Deutschland mit seinen 16 Bundesländern ist ein Fall unterschiedlicher Grade der Annäherung an eine (Quasi-)Integration innerhalb der Differenzierung, etwa die Verlängerung der ersten Trennung oder das Angebot an Integrierten Gesamtschulen. Hinzu kommen zwei zusätzliche Regelungen, die *unmittelbar* bei den Strukturen für die Differenzierung selbst ansetzen: Die Anzahl der Optionen, die Mehrgliedrigkeit also, und die Stringenz, in der die Differenzierung implementiert und vollzogen wird. Hier setzen die bildungspolitischen Debatten am ehesten an, soweit sie Systemregelungen betreffen: Die Umstellung von der Drei- auf die Zweigliedrigkeit und die Aufhebung der Verbindlichkeit und die Lockerung der Kontrollen, die Auflösung der Stringenz also, in der das „Ability"-Tracking auch wirklich als *Ability*-Tracking implementiert wird – mit Verbindlichkeit *und* Kontrolle also (s. dazu gleich unten Abschnitt 2).

Begründungen für die *Zweigliedrigkeit* finden sich u.a. bei Bohl et al. (2017) und in den Beiträgen von van Ackeren und Kühn (2017) und Wacker (2017) darin, bei Maaz (2017) oder Autor*inengruppe Neue Sekundarschule (2024), früh auch schon in einer Ausarbeitung des wissenschaftlichen Dienstes des Deutschen Bundestages (Deutscher Bundestag 2006). Als empirische Belege gäbe es einige Beiträge zu entsprechenden Reformen in deutschen Bundesländern: Piopiunik (2014) für Bayern, Neumann et al. (2017) für Berlin, Maaz et al. (2019) für Bremen. Die Ergebnisse zu den Strukturreformen in Berlin und Bremen lassen sich leicht zusammenfassen: Es gab zwar einige Effekte, etwa auf die Akzeptanz der Änderungen und die Kooperation, aber so gut keine für das, worauf es hauptsächlich angekommen wäre: Leistungen und Mindeststandards (für Berlin: Baumert et al. 2013, Becker et al 2017. 181ff.,

Baumert et al. 2017: 217ff.; für Bremen: Maaz et al. 2019: 219ff.; für Deutschland insgesamt: Marx und Maaz 2023: 194f., s. auch die eher ernüchternden Feststellungen zur (Un-)Wirksamkeit der verschiedenen Maßnahmen bisher bei Maaz und Lörz 2024).

In der Untersuchung von Piopiunik (2014) für Bayern war es anders. Dort war im Jahr 2000, also noch vor Veröffentlichung der ersten PISA-Ergebnisse, im nicht-gymnasialen Bereich die *Drei*gliedrigkeit mit der Trennung von Haupt- und Realschule *ein*geführt worden, eine Reform also gerade *gegen* den Trend der Zusammenlegung. Das Ergebnis: Die Leistungen *sanken* und die Streuung *vergrößerte* sich mit der Einführung der Differenzierung in den unteren Bereichen, also gerade *gegen* die Hypothesen der Differenzierungsposition und in Übereinstimmung mit den Forderungen später nach Zusammenlegung. Die Effekte wurden mit gewissen Nebenfolgen der Umstellung und ihrer verzögerten Implementation begründet, ein Problem, das es bei Reformen nicht selten gibt (Piopiunik 2014: 31f.). Auch mit dieser Studie gibt es jedoch Probleme. Das wohl wichtigste: Es gibt ein sog. Collider-Problem (vgl. Elwert und Winship 2014: 36 ff.): Die Leistungen werden über die Schuleffekte *und* die kognitiven Fähigkeiten bestimmt, man weiß aber nicht wie die untereinander zusammenhängen. Die Analysen beruhten auf PISA-Daten, die keine Informationen über die kognitiven Fähigkeiten oder die Leistungen vorher enthalten. Es lässt sich daher auch hier nichts über den Effekt der Reform wirklich sagen. Der Spitzenstellung von Bayern im Leistungsniveau generell auch in den unteren Leistungsbereichen, damals schon und später, hat das alles nicht geschadet.

Direkte *Vergleiche* für die deutschen Bundesländer zur Zweigliedrigkeit gibt es ebenfalls nicht oft. Die Befunde beziehen sich oft nur auf Nebenaspekte. Maaz (2017: 200) verweist etwa auf die Studie von Baumert et al. (2010) über leistungsstarke Kinder in den Berliner Grundschulen, die sich in den Leistungen nach der 5. und 6. Klasse nicht von Kindern in den sog. „Grundständiges Gymnasium“ unterscheiden. Das wird als Beleg gewertet, dass die Leistungsheterogenität den begabten Kindern nicht schade. Das aber könnte auch daran liegen, dass die Berliner Gymnasien ohnehin nicht besonders stark sind. Auch auf Hattie (2009) wird hingewiesen, der, allgemein für die Schulen, belegt hätte, dass „insbesondere leistungsschwächere Schülerinnen und Schüler von einer größeren leistungsbezogenen Heterogenität profitieren“ (Maaz 2017: 200). Davon kann jedoch keine Rede sein: Hattie (2009) findet für den (nicht-konditionalen) Effekt beim Ability-Grouping in Schulklassen einen nur schwachen d-Wert von 0.12 (Hattie 2009: 89) und Hattie (2023: 151ff.) in der Nachfolgearbeit einen von 0.09. In dieser Größenordnung sind danach auch äquivalente Vorgänge und Strukturen wie für die Auflösung des Tracking oder das jahrgangsübergreifende Lernen, also

jeweils für eine höhere Heterogenität. Das aber hieße zusammengefasst, dass es auch nach den großen Meta-Analysen keine Effekte der Heterogenität in den Schulklassen gibt, und das weder oben, noch unten im Leistungsniveau.

Es gibt eine Studie, die *positive* Effekte der Zweigliedrigkeit im *direkten* Vergleich der deutschen Bundesländer insgesamt findet (Matthewes 2021). Der Ansatz lehnt sich im Design an Piopiunik (2014) an: die Beschränkung auf die unteren Bildungswege und Kontrollen der unbeobachteten Heterogenität bei den Bundesländern bzw. für Parallelentwicklungen bei den Gymnasien über ein zwei- bzw. dreifaches *difference-in-difference*-Verfahren. Unter Kontrolle insbesondere von Einflüssen der sozialen Herkunft und verschiedener Merkmale der Schulen, wie Ausstattung, Klassengröße und Lehrpersonal, finden sich nennenswerte nicht-konditionale Effekte zugunsten der Zweigliedrigkeit, beim Lesen stärker als bei der Mathematik, und dazu noch einmal, als bivariate Beziehung, positive Effekte der Kinder mit den *geringsten* Leistungen in der Grundschule vorher, also gerade das, was die Integrationsposition immer betont hatte. Allerdings stellen sich auch hier wieder die schon geläufigen Fragen. Wie bei Piopiunik werden die kognitiven Fähigkeiten nicht kontrolliert, so dass es auch hier ein Collider-Problem aus der nicht-gemessenen Beziehung etwa der sozialen Herkunft über die nicht gemessenen kognitiven Fähigkeiten auf die Leistungen in der Sekundarstufe könnte. Wichtiger aber noch: Es werden auch die Leistungen vorher in der Grundschule nicht kontrolliert, so dass Effekte der Antezipation von bestimmten Folgen in der Grundschule, die es mit der Verbindlichkeit geben könnte, nicht erfasst werden. Das aber wäre eine Voraussetzung für das *difference-in-difference*-Verfahren gewesen, das zur Kontrolle der unbeobachteten Heterogenität zwischen den Bundesländern und des Verlaufstrends mit den Gymnasien eingesetzt wird (Matthewes 2022: F. 43: 1303): Eine einfache VA-Schätzung kann es mit den Fixierungen nicht geben, weil die Verfahren nicht ineinander vernestet sind. Der gefundene (positive) Effekt könnte damit aber auch ein Artefakt des Verzichts auf die VA-Schätzung sein: Keine Kontrolle der individuellen Selektion und übergangene Effekte der Antezipation im Vergleich von Systemen mit unterschiedlichen Anreizen für Anstrengungen schon vor dem Übergang (oder auch noch später und auf allen Ebenen der Schulstruktur, auch jener des gymnasialen Zweigs, die in den VA-Schätzungen einbezogen sind; vgl. dazu auch noch die Abschnitte 5.4 und 8 unten).

Soweit der Stand der Dinge für die Zweigliedrigkeit. Für die Effekte der von Öffnung und Lockerung auf die Leistungen, die Aufgabe der Stringenz also, gibt es einige Befunde (Esser

2023, Kapitel 7 und 9 insbesondere, s. die Zusammenfassung bei Münch 2024), auch Repliken und Gegenrepliken (Heisig und Matthewes 2022, Lorenz et. al 2023, Esser und Seuring 2023, Esser 2024b). Das Kernergebnis ist, dass mit der Stringenz die Leistungen in der Grundschule *steigen*, in der Sekundarstufe auch, dass dabei die Kinder in den Klassen mit den *niedrigeren* Leistungen profitieren, wenn die Klassen homogen sind. Verstärkungen der Effekte der sozialen Herkunft oder der sozialen Segregation in den Schulklassen gibt es dabei *nicht*. In den Klassen mit den höheren Leistungen bringt dagegen gerade die kognitive Heterogenität etwas. Die Integration in der Differenzierung ist offenbar etwas nur für Konstellationen, in denen das Lernen kein Problem ist.

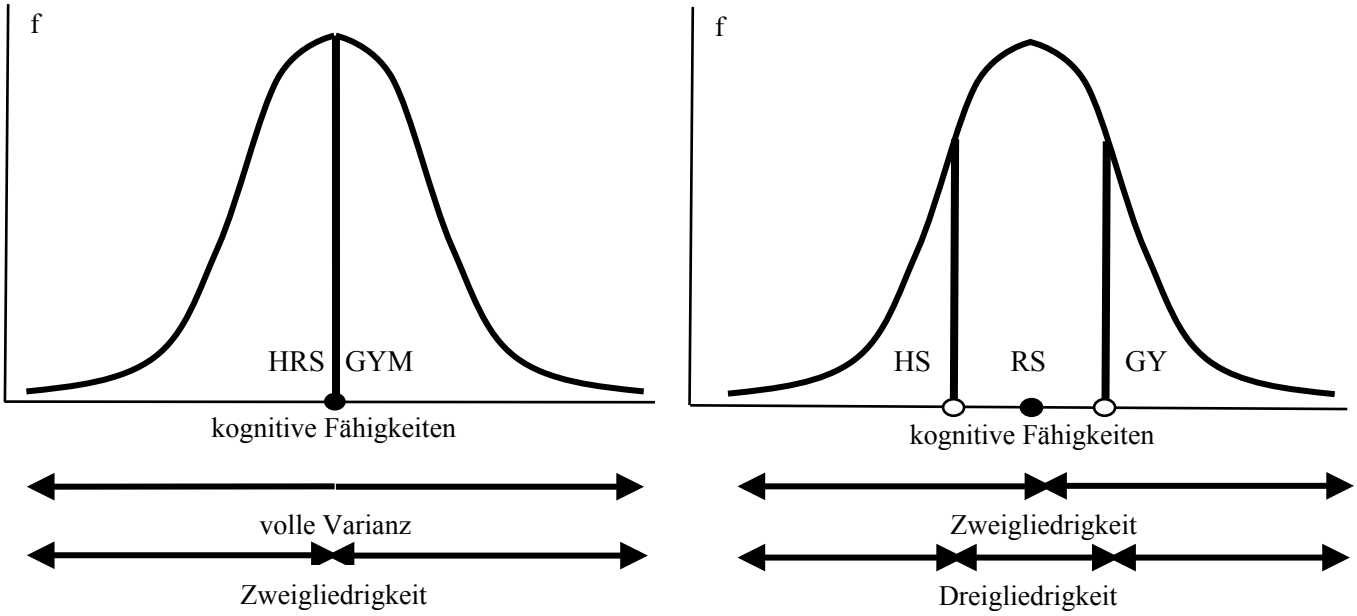
3. Der theoretische Rahmen

In der Debatte um die Differenzierung der Bildungswege stehen sich zwei Positionen gegenüber. Nach der *Differenzierungsposition* erlaubt die Trennung in unterschiedliche Bildungswege nach den kognitiven Fähigkeiten wegen der stärkeren kognitiven *Homogenität* in den Schulklassen einen effizienteren Unterricht und darüber höhere Leistungen, ohne dass die soziale Bildungsungleichheit zunehmen müsse. (Sörensen 1970, Sörensen und Hallinan 1977). Nach der *Integrationsposition* wäre es geradezu umgekehrt: dem Lernen und den Leistungen förderliche peer-Interaktionen seien nur bei kognitiver und sozialer *Heterogenität* möglich, und nur ohne die Stratifikationen nach unten ließen sich die Marginalisierungen beheben, die die räumliche, soziale und kognitive Segregation der Bildungswege mit sich brächten (Oakes 1985, Zimmer und Toma 2000, Domina et al. 2017, Domina et al. 2019).

Die Effekte von kognitiver Homogenität und Heterogenität auf Unterricht und peer-Interaktionen lassen sich vor diesem Hintergrund über einfache Skizzen von Verteilungen verdeutlichen, in Abbildung 1 zunächst für Integration und Mehrgliedrigkeit). [1]

[1] Die Darstellung orientiert sich an den weiter gefassten Konzeptionen des Modells der Leistungsdifferenzierung (abgekürzt als „MoAbiT“ nach „Model of Ability-Tracking“)bei Esser 2016a, Esser 2016b, Esser 2021, Kapitel 8, Esser 2023: Vorkapitel „Hintergrund“).

Abbildung 1: Verteilung der kognitiven Fähigkeiten in den Schulklassen für Integration und Differenzierung bei Zwei- und Dreigliedrigkeit.



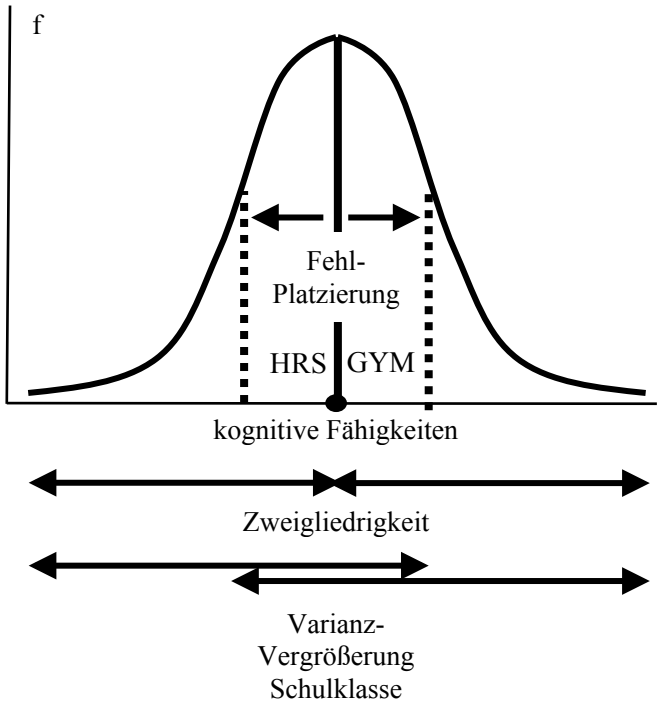
Dargestellt sind die (Normal-)Verteilungen der kognitiven Fähigkeiten über eine Population von Kindern in den Schulklassen. Im linken Diagramm finden sich die Verteilungen für die (vollauf) integrierten Systeme im Vergleich zur Zweigliedrigkeit, rechts die der Zweigliedrigkeit gegenüber der Dreigliedrigkeit.

Nach der Differenzierungsposition verringern sich mit der Aufteilung die Varianzen und das wirkt „wie“ eine Verringerung der Klassengröße. Das erlaubt gerade auch einen auf Einzelprobleme eingehenden „individualisierten“ Unterricht. Der Extremfall wäre der Privatunterricht – was weder zu finanzieren, noch wünschenswert wäre. Die Integrationsposition betont andere Folgen: Das „längere gemeinsame Lernen“ mit peer-Interaktionen über die gesamte Spannweite der sozialen und kognitiven Zusammensetzung der Schulklassen hinweg verbessert die Leistungen, und ohne die Stratifikation nach unten gäbe es keine Entmutigungen, Abwertungen und Marginalisierungen, etwa in den sog. „Restschulen“, den mehr und mehr gemiedenen, stigmatisierten und vernachlässigten Hauptschulen in sozialen „Brennpunkten“.

Die beiden Positionen nehmen in ihren Begründungen offenbar die *gleichen* Konstrukte an, schreiben ihnen aber *unterschiedliche* Effekte und Gewichtungen zu. Das aber ist nicht alles. Die Effekte der Differenzierung hängen davon ab, ob es sich jeweils auch *wirklich* um eine Aufteilung nach den *kognitiven* Fähigkeiten gehandelt hat, um ein „Ability“-Tracking also, das

diese Bezeichnung verdient, und auch, dass es die nötigen Umstellungen in den Curricula und im Unterricht auch *tatsächlich* gegeben hat (vgl. dazu schon früh: Sörensen und Hallinan 1977, Hallinan 1994). In Abbildung 2 ist das für den Fall von Fehlplatzierungen bei der Aufteilung auf die Bildungswege in einem zweigliedrigen System im Anschluss an Abbildung 1 oben skizziert.

Abbildung 2: Folgen von Fehlplatzierungen für die Zusammensetzung der Schulklassen nach den kognitiven Fähigkeiten (bei Zweigliedrigkeit)



Ohne Fehlplatzierungen gibt es keine Überscheidungen der objektiven kognitiven Fähigkeiten mit der Zuordnung zu den beiden Schultypen (HRS und GYM). Oft schon ist gezeigt worden, dass das empirisch nicht zutrifft, sei es über abweichende Bildungsentscheidungen der Eltern oder über ungenaue und gerechte Bewertungen des Lehrpersonals. Die Fehlplatzierungen kann es nach oben und nach unten geben: Es kommen mehr auf das Gymnasium (Pfeil nach links) oder weniger (Pfeil nach rechts) als es nach den Fähigkeiten sein sollte. Die Folge sind Überlappungen in den kognitiven Fähigkeiten bei den Schultypen mit einer dann für den faktischen Unterricht vergrößerten Varianz auch in den Schulklassen (s. die Pfeile unten). Dadurch würde der Unterricht mit der Differenzierung ineffizienter und es gäbe eine Minderung im Leistungsniveau, und zwar unabhängig von der Gliedrigkeit der Differenzierung.

Es käme also nach der Differenzierungsposition ganz besonders darauf an, derartige Fehlplatzierungen bei der Aufteilung zu vermeiden. Das wäre möglich, und es wird auch versucht: über *Zusatz*-Regelungen als Vorgabe für die Implementation der Differenzierung. Hier wären es insbesondere zwei: Die Verbindlichkeit der schulischen Empfehlungen und die Kontrolle der schulischen Organisation.

Für die *Verbindlichkeit* sprechen es zwei Gründe. Höhere *Anreize* für eine frühe kognitive Entwicklung in der Familie und die Leistungen in der Grundschule und die *Blockade* von unangemessenen Aspirationen bei den Bildungsentscheidungen, besonders bei den oberen Schichten. Die *Kontrolle* bezieht sich auf die faktische *Implementation* der Anpassung und die Fokussierung des Unterrichts, sowie auf die *Dämpfung* der tertiären Effekte, die auch nach der sozialen Herkunft unterschiedliche Bewertung der Leistungen der Kinder, insbesondere bei den Empfehlungen.

Verbindlichkeit und Kontrolle würden, so kann man vermuten, gerade in ihrer *Kombination* wirken, weil sie alle Akteure und Vorgänge gleichzeitig umfasst. Diese Kombination sei als die *Stringenz* eines Bildungssystems bezeichnet. Womöglich entsteht oder stabilisiert sich darüber auch so etwas wie ein übergreifendes „Bildungsklima", das die Vorgänge auch dann wie selbstverständlich trägt, wenn es einmal Schwierigkeiten gibt. Die Stringenz sorgt, so könnte man sagen, für eine ganz besondere Resilienz der Strukturen und Abläufe gegen Irritationen von außen. Ihre Effekte könnten zu möglichen anderen Systemeigenschaften hinzutreten, wie zur Mehrgliedrigkeit, sich ggf. aber auch *konditional* damit verstärken oder abschwächen. Darum geht es nun.

4. Der Untersuchungsansatz

Das theoretische Modell unterscheidet zwei Arten von Effekten: *allgemeine* Einflüsse auf die Leistungen und *System*effekte der *Änderung* der allgemeinen Einflüsse in bestimmten Kontexten übergreifend. Die Änderungen können sich zwei Aspekte beziehen: Der Effekt einer bestimmten Systemeigenschaft als zusätzlichen *nicht*-konditionalen Einfluss zu dem der allgemeinen Bedingungen. Das kann empirisch über die Kontrolle der *Mediation* des Systemeffekts mit den allgemeinen Bedingungen festgestellt werden. Hinzu kommt die Analyse einer möglichen *Moderation* der allgemeinen Bedingungen, der nach bestimmten

Systemmerkmalen *konditionalen* Änderung ihrer Wirkungen, identifizierbar über entsprechende Interaktionseffekte von allgemeinen und Systembedingungen.

In Tabelle 1 sind die allgemeinen Bedingungen und die Systemeffekte nach dem theoretischen Modell des MoAbiT zusammenfassend und jeweils mit den Hypothesen für die beiden Ansätze und Systemeigenschaften der Mehrgliedrigkeit und der Stringenz dargestellt. Die empirischen Analysen folgen dieser Aufstellung.

Tabelle 1: Allgeneine Bedingungen und Systemeffekte für die Leistungen in der Grundschule und in der Sekundarstufe nach Mehrgliedrigkeit (D für Dreigliedrigkeit) und Stringenz (T für Tracking), ... nicht vorgesehen oder berücksichtigt.

Bedingungen	Kürzel	Positionen	
		Integration/ Standardansatz	Differen-zierung
allgemeine Effekte (nicht konditional)			
soziale Herkunft	SES	+	+
Niveau soziale Herkunft	NSES	+	+
Homogenität soziale Herkunft	HSES	-	0
Interaktion NSES*HSES	NSES*HSES	-	0
kognitive Fähigkeiten	ABL	…	+
Leistungen Grundschule	ACE	+	+
Niveau kogn. Fähigkeiten	NABL	…	+
Homogenität kogn. Fähigkeiten	HABL	…	0
Interaktion NABL*HABL	NABL*HABL	…	0
Schultyp	GYM	+	+
Dreigliedrigkeit (nicht konditional)	D	≤0	≥0
Stringenz (nicht konditional)	T	≤0	≥0
Systemeffekte (konditional)			
D*soziale Herkunft	D*SES	+	0
D*Niveau soziale Herkunft	D*NSES	+	0
D*Homogenität soziale Herkunft	D*HSES	+	0
D*Niveau*Homogenität s. Herkunft	D*NSES*HSES	+	0
D*kognitive Fähigkeit	D*ABL	…	0
D*Niveau kogn.Fähigkeit	D*NABL	…	+
D*Homogenität kogn.Fähigkeit	D*HABL	…	+
D*Niveau*Homogenität kogn.Fähigk.	D*NABL*HABL	…	0
D*Schultyp	D*GYM	+	+
T*soziale Herkunft	T*SES	+	0
T*Niveau soziale Herkunft	T*SES	+	0
T*Homogenität soziale Herkunft	T*HSES	-	0
T*Niveau*Homogenität s. Herkunft	T*NSES*HSES	-	0
T*kognitive Fähigkeit	T*ABL	…	0
T*Niveau Niveau kogn.Fähigkeit	T*NABL	…	+
T*Homogenität kogn.Fähigkeit	T*HABL	…	+
T*Niveau*Homogenität kogn.Fähigk.	T*NABL*HABL	…	0
T*Schultyp	T*GYM	+	+
Dreigliedrigkeit*Stringenz	D*T	≤0	≥0
Dreigliedrigkeit (konditional)	(D)	≤0	≥0
Stringenz (konditional)	(T)	≤0	≥0

Das Explanandum sind die Leistungen in der Sekundarstufe (ACS in den Analysen). In der linken Spalte stehen alle in den jeweiligen Positionen benannten oder nach dem MoAbiT vorgesehenen Konstrukte und Einflüsse. Beim Vergleich der Hypothesen der beiden Ansätze, Integrations- und Differenzierungsposition, fällt auf, dass es für einige Bereiche keine expliziten Hypothesen gibt oder, wie beim Standardansatz, bestimmte Konstrukte in den Analysen nicht berücksichtigt wurden.

Sieht man davon ab, werden deutliche Unterschiede zwischen den Ansätzen erkennbar. Sie beziehen sich auf die *System*effekte, nicht-konditionale und konditionale (grau unterlegte Zeilen oben bzw. unten). Es sind im Wesentlichen vier Aspekte. *Erstens*: Bei den nicht-konditionalen Systemeffekten nimmt die Integrationsposition sowohl für die Dreigliedrigkeit wie für die Stringenz keine Verbesserung, sondern eher eine *Verringerung* der Leistungen (in Grundschule und Sekundarstufe) an, ebenso wie bei den entsprechenden konditionalen (Haupt-)Effekten für die Interaktion mit den allgemeinen Bedingungen (soziale Herkunft und kognitive Fähigkeiten und die Effekte der sozialen und kognitiven Zusammensetzung der Schulklassen). Das ist nach der Differenzierungsposition bzw. dem MoAbiT diametral anders: Keine Verringerung, sondern eher eine *Verbesserung* des Leistungsniveaus, ebenfalls in Grundschule und Sekundarstufe. *Zweitens*: Die Differenzierungsposition erwartet für die (System-)Effekte der sozialen Herkunft weder für die Dreigliedrigkeit, noch für die Stringenz besondere Veränderungen der Effekte der sozialen Herkunft, eher eine Verringerung. Das ist anders für die Integrationsposition. Hier werden teils Verstärkungen erwartet, wie die Schereneffekte in Schulklassen mit einem höheren sozialen Niveau, teils Abschwächungen wie in Schulklassen mit hoher sozialer Homogenität. *Drittens*: Der Schultyp, Gymnasium, Realschule, Hauptschule, hat nach *beiden* Ansätzen einen zu allem anderen dann noch einmal größeren positiven Effekt, der sich, wieder nach beiden Ansätzen, mit der Dreigliedrigkeit wie mit der Stringenz noch verstärkt. Nach der Integrationsposition ist das ein Teil der (beklagten) Schereneffekte, nach der Differenzierungsposition eine weitere Verbesserung der (pareto-optimalen) Zusatzgewinne. *Viertens:* Die positiven Effekte der Stringenz sollten sich nach der Differenzierungsposition mit der Dreigliedrigkeit noch weiter verstärken, und das sowohl in der Grundschule schon wie in der Sekundarstufe. Der Grund: Mit der Dreigliedrigkeit steigen die Anreize zu Anstrengungen vorher, weil sich dann im Vergleich zur Zweigliedrigkeit sozusagen die Fallhöhe verstärkt, dem bei Stringenz nicht ausgewichen werden kann, damit bereits die nicht-konditionalen Systemeffekte nach oben gehen und in der Sekundarstufe die Effizienz über die größere kognitive Homogenität steigt, was besonders den Kindern im unteren

Leistungsbereich zugute käme, weil es für sie dann mehr Möglichkeiten der individuellen Betreuung bei Schwierigkeiten gibt.

5. Daten, Konstrukte und Analysen

Die internationalen Vergleichsstudien erlauben eine empirische Prüfung des theoretischen Modells nicht, weil einige, nach *beiden* Ansätzen, der Integrations- wie der Differenzierungsposition, relevanten Daten fehlen (s. oben dazu schon). Die Daten der "National Educational Panel Study" (NEPS) enthalten alles, was für eine empirische Prüfung der Annahmen des Modells, des MoAbiT, für die deutschen Bundesländer zum Zusammenspiel von Mehrgliedrigkeit und Stringenz notwendig wäre (s. Tabelle 1 unten). Hier werden die Datengrundlage, die Verteilung der Variablen und die Zuordnung der Bundesländer zur Mehrgliedrigkeit und zu den Regelungen der Stringenz und die Analyseverfahren beschrieben. [2]

5.1 Die Grundlage

Das NEPS wird vom Leibniz-Institut für Bildungsverläufe (LIfBi, Bamberg) in Kooperation mit einem deutschlandweiten Netzwerk durchgeführt (vgl. Blossfeld und Roßbach 2011). Die hier betrachtete Startkohorte 3 (NEPS-Netzwerk 2021) verfolgt Schülerinnen und Schüler, die im Schuljahr 2010/2011 in Deutschland die fünfte Klasse besuchten, in jährlichen Befragungen und Kompetenztests. Die verwendeten Variablen entstammen der ersten bis dritten Welle (Klasse 5-7) einer schriftlichen Befragung und Kompetenztestung der Schülerinnen und Schüler sowie einer telefonischen Befragung jeweils eines Erziehungsberechtigten. In die Analysen wurden die drei spät sortierenden Länder Berlin, Brandenburg und Mecklenburg-Vorpommern *nicht* aufgenommen, weil in der dritten Welle die erforderlichen Daten für den Übergang und die Leistungen noch nicht vorlagen. Aus den verbleibenden 13 Bundesländern haben 5.248 Schülerinnen und Schüler an der ersten Welle und 4.719 an der dritten Befragungswelle teilgenommen. Nach dem Ausschluss von Fällen ohne verfügbare Ergebnisse aus den Leistungstests in Klasse 5 (N=503) und dann in Klasse 7 (N=769) und derjenigen mit unvollständigen Angaben für die Modellvariablen (N=1.703 bzw. N=1.314), umfasst die Analysestichprobe für die Untersuchung von Sortierung, Übergang und Bildungsbeteiligung

[2] Wie zum theoretischen Modell gibt es auch für die Einzelheiten der Operationalisierung der verwendeten Konstrukte ausführlichere Informationen aus anderen Publikationen; vgl. Fußnote 1 oben.

ein N von 3.042 Schülerinnen und Schüler in 207 Schulen und 382 Schulklassen in der 5. Klasse (für die Analysen zur Bildungsbeteiligung und Strukturierung der Schulklassen), bzw. für die Analysen zur Strukturierung auf Klassenebene ein N von 339 Schulklassen, und für die Analysen der Leistungen in der Sekundarstufe in der 7. Klasse ein N von 2.636 Schülerinnen und Schüler in 171 Schulen und 313 Schulklassen.

5.2 Konstrukte und Verteilungen

Tabelle 2 beschreibt die Konstrukte des MoAbiT für die Analysen und deren Verteilung für die beiden Wellen 1 und 3 der Startkohorte 3 des NEPS bzw. für die Klassen 5 und 7. Alle kontinuierlichen Variablen aus Klasse 5, sowohl individuelle als auch Kontext-Merkmale, die vorwiegend als unabhängige Variablen verwendet werden (SES, ABL, ACE, NSES, HSES, NABL und HABL), wurden für die Analysen so transformiert, dass 0 jeweils dem geringsten empirischen Wert und 1 dem höchsten empirischen Wert entspricht.

Tabelle 2: Wertebereich, Mittelwerte und Streuungen bzw. prozentuale Verteilung der in den Analysen verwendeten Konstrukte

			Klasse 5		Klasse 7	
	Kürzel	min/max	av/%	sd	av/%	sd
Individuen						
soziale Herkunft	SES	0/1	0.54	0.21	0.54	0.22
kognitive Fähigkeiten	ABL	0/1	0.60	0.21	0.60	0.21
Leistungen Grundschule	ACE	0/1	0.47	0.15	0.48	0.15
Leistungen Sekundarstufe	ACS	3.5/3.9	0.56	0.16	0.75	0.16
Schulklassen						
Niveau SES	NSES	0/1	0.61	0.16	0.55	0.19
Homogenität SES	HSES	0/1	0.68	0.12	0.67	0.13
Niveau ABL	NABL	0/1	0.62	0.20	0.60	0.20
Homogenität ABL	HABL	0/1	0.66	0.12	0.57	0.16
Schultyp	GYM	0/1	0.55	0.50	0.48	0.50
Sozio-Demographie						
Geschlecht	FEM	0/1	0.48	-	0.49	-
Migrationshintergrund	MHG	0/1	0.30	-	0.29	-
Vorschulbesuch (Monate)	VSB	0/83	39.30	10.91	39.35	10.85
n (SuSen			3042		2636	
N (Schulklassen)			382		313	

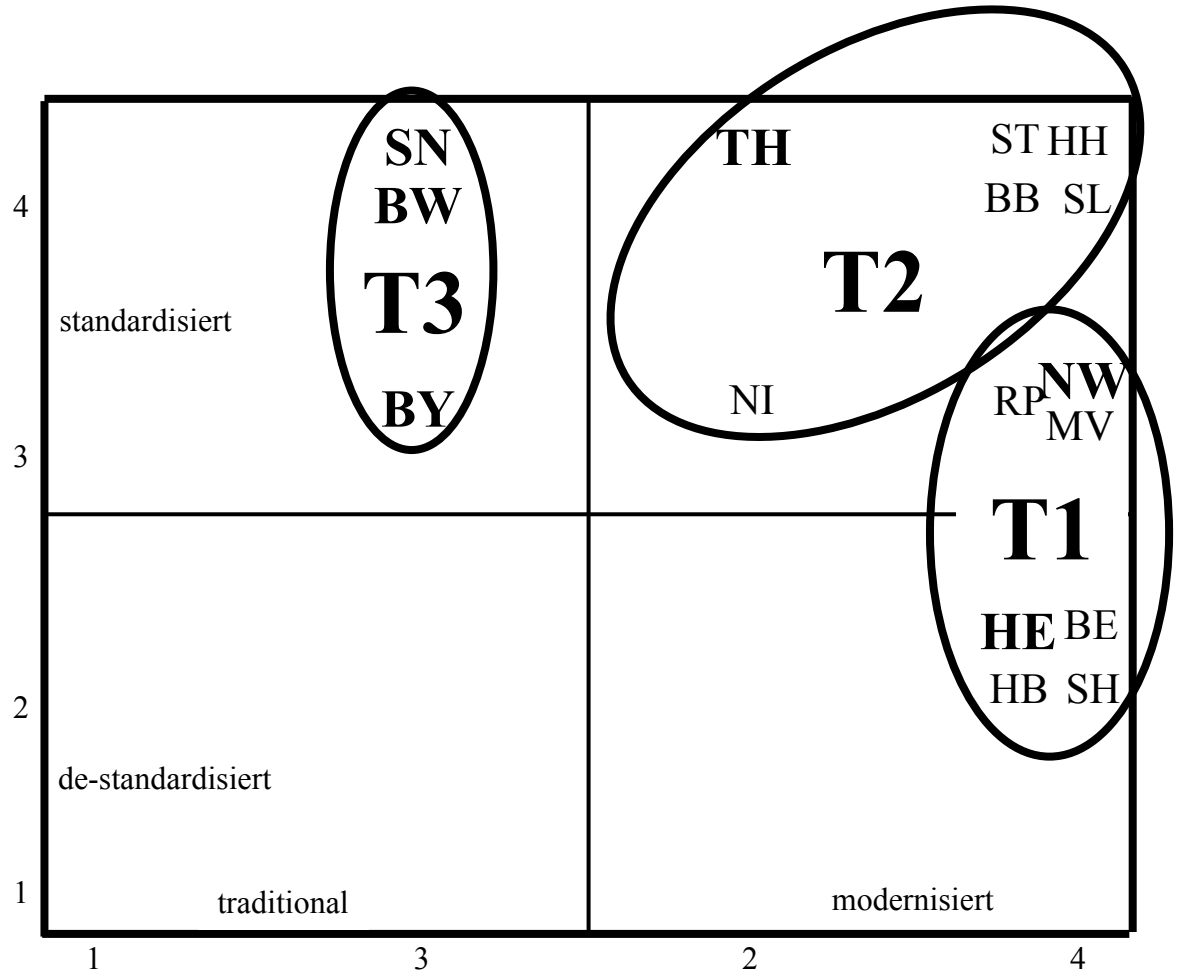
Besondere Auffälligkeiten in den Verteilungen lassen sich nicht erkennen, auch nicht bei der Bildungsbeteiligung (GYM) mit 55% bzw. 48% Gymnasialbesuch nach dem Übergang.

5.3 Die Zuordnung der Bundesländer

Den Kern der Analysen zu den Systemeffekten bildet die Einteilung der Bundesländer nach der der Mehrgliedrigkeit und der Stringenz der Leistungsdifferenzierung. Die Zuordnung der Bundesländer zur Stringenz folgt dem Vorgehen bei den Typologien bei von Below (2011, Abschnitt 4.3) und der noch weiter differenzierenden Klassifikation bei Helbig und Nikolai (2015, Abbildung 28, S. 286) in zwei grundlegende Dimensionen, die mit Modernisierung vs. Traditionalität und De-Standardisierung vs. Standardisierung bezeichnet werden. Sie werden für die weiteren Analysen zu drei Typen aus der im theoretischen Teil begründeten

Kombination von *Verbindlichkeit* der Empfehlungen und der *Kontrolle* der schulischen Abläufe zusammengefasst: In T1 gibt es weder Verbindlichkeit noch Kontrolle, in T3 sowohl Verbindlichkeit wie Kontrolle und in T2 nur jeweils eine dieser Bedingungen. Hinzu kommt die Einteilung nach der Mehrgliedrigkeit mit der Dreigliedrigkeit, gekennzeichnet über die fett hervorgehobenen Kürzel für die Bundesländer.

Abbildung 3: Die deutschen Bundesländer in einem Merkmalsraum von Stringenz und Mehrgliedrigkeit (für 2010/11) nach einer Illustration bei Helbig and Nikolai 2015: Abbildung 28: 286.



BB=Brandenburg, BE=Berlin, BW=Baden-Württemberg, BY=Bayern, HB=Bremen, HE=Hessen, HH=Hamburg, MV=Mecklenburg-Vorpommern, NI=Niedersachsen, NW=Nordrhein-Westfalien, RP=Rheinland-Pfalz SH=Schleswig-Holstein, SL=Saarland, SN=Sachsen, ST=Sachsen-Anhalt, TH=Thüringen.

Baden-Württemberg, Bayern und Sachsen sind die drei Bundesländer mit der höchsten Stringenz (Verbindlichkeit und Kontrolle; T3), Berlin, Bremen, Hessen, Mecklenburg-Vorpommern, Nordrhein-Wetsfalen und Rheinland-Pfalz die mit der geringsten (weder Verbindlichkeit, noch Stringenz; T1), alle anderen liegen dazwischen (Verbindlichkeit oder Kontrolle; T2). Die dreigliedrigen Bundesländer sind Baden-Württemberg, Bayern, Hessen,

Nordrhein-Westfalen, Sachsen und Thüringen, zweigliedrig Berlin, Brandenburg, Bremen, Hamburg, Mecklenburg-Vorpommern, Niedersachsen, Rheinland-Pfalz, das Saarland, Sachsen-Anhalt und Schleswig-Holstein. Es zeigt sich, dass die Verteilung der beiden Dimensionen über die Bundesländer eine gewisse Unabhängigkeit aufweist. Ein Feld in dem Diagramm bleibt leer: Es gibt in der Startkohorte 3 im Jahr 2010/11 *kein* Bundesland mit Stringenz *und* Zweigliedrigkeit.

5.4 Die Analysen

Die verschiedenen Bedingungen zur Erklärung der Leistungen in der Sekundarstufe beziehen sich auf die Konstrukte des theoretischen Modells nach den Tabellen 1 und 2 oben. Die Analysen zur multivariaten Bestimmung der jeweiligen Systemeffekte folgen dem „Value Added"-Ansatz (VA): Die Identifikation des *marginalen* Zuwachses bei den Leistungen der *Individuen* über eine bestimmte Bedingung, die Eigenschaften der Schultypen und Schulklassen insbesondere. Das setzt die Kontrolle der *Selektion* in die jeweiligen Kontexte voraus, nach dem MoAbiT für die Differenzierung im Wesentlichen über die kognitiven Fähigkeiten und die Leistungen zuvor in der Grundschule. Damit können auch Vorgänge der *Antezipation* erfasst werden, wie die Zunahme von Anstrengungen *vor* dem Übergang, wenn dafür bestimmte Leistungen verlangt werden und dem nicht ausgewichen werden kann, wie etwa bei der Verbindlichkeit der Empfehlungen.

Mit dem Analyseziel der Identifikation der marginalen (System-)Effekte verbieten sich bestimmte Fixierungen, wie sie oft für die Kontrolle unbeobachteter Heterogenität bei den Kontexten eingesetzt werden, wie bei PISA gerade oft auch mit dem Fehlen wichtiger Informationen zu relevanten Einflüssen: *fixed effects* bei den Schulklassen oder *difference-in-difference*-Kontrollen bei den Bundesländern. Das wäre schon vom Ansatz des MoAbiT her nicht sinnvoll: Es werden theoretisch ja gerade bestimmte *Moderationen* der allgemeinen Zusammenhänge erwartet, die mit den Fixierungen unterdrückt würden. Außerdem wären sie überflüssig: Der für die Bundesländer gleiche nationale Kontext sorgt für eine weitgehende Einebnung der Verhältnisse und mit dem MoAbiT werden, vor allem über die Erfassung der Schulstrukturen, die wichtigsten individuellen und kontextuallen Merkmale erfasst. Was im internationalen Vergleich vielleicht nötig wäre, wird nun unnötig. Es kann eher sogar zu Verzerrungen bei der Schätzung der Systemeffekte kommen (vgl. Clarke et al. 2010). Bei den Fixierungen darf es zudem keine Antizipationseffekte geben, also gerade das, was womöglich

einen Teil der Systemeffekte ausmachen könnte und empirisch auch gut nachgewiesen ist (Grewenig 2020) oder für die Abschaffung der Verbindlichkeit in Baden-Württemberg (vgl. Bach und Fischer 2020, Osikominu 2020).

In den Analysen hier (nach dem VA-Ansatz) werden lineare (Mehrebenen-)Regressionsmodelle geschätzt. In allen Analysen wird die genestete Datenstruktur (Schülerinnen und Schüler in Schulklassen) über die Schätzung cluster-robuster Standardfehler berücksichtigt. Zur quantitativen Bestimmung der Effektstärken wurden die abhängige Variable der Leistungen in Klasse 7 z-standardisiert (av=1, sd=0). Die Regressionskoeffizienten lassen sich demnach als Unterschiede in Standardabweichungen interpretieren. In allen Modellen wird für das Geschlecht des Kindes (FEM, mit 0 für männlich und 1 für weiblich), den Migrationshintergrund (MHG, mit 0 für einheimisch und 1 für Migrationshintergrund) und die Dauer des Vorschulbesuchs vor der Einschulung (VSB, in Monaten) kontrolliert.

6. Die Befunde

Der Bericht über die Befunde beginnt mit der deskriptiven Übersicht über die Ergebnisse der PISA-Studien und der IQB-Berichte von 2000 bis 2022 für das Leistungsniveau und die Herkunftseffekte im internationalen Vergleich und für die deutschen Bundesländer (Abschnitt 6.1). Anschließend kommen in Abschnitt 6.2 die multivariaten Analysen mit den Daten des NEPS zu den Effekten von Zweigliedrigkeit und Stringenz nach dem theoretischen Modell des MoAbiT.

6.1 Die deutschen Bundesländer im internationalen und regionalen Vergleich von PISA 2000 bis 2022

Dafür, dass die Integration, Öffnung und Lockerung keineswegs in Effizienz und sozialer Durchlässigkeit der strikten Differenzierung überlegen wären, gibt es seit den ersten PISA-Erhebungen gut dokumentierte Belege. In Tabelle 3 stehen die Befunde zu den PISA-Ergebnissen bzw. denen aus den IQB-Berichten von 2000 bis 2022 im internationalen Vergleich und dem der deutschen Bundesländer und in Tabelle 4 finden sich die für die deutschen Bundesländer die Werte für die leistungsschwächeren Schüler:innen und für die

soziale Durchlässigkeit, also alles das, worum es in den Debatten über die Bildungssysteme geht: Effizienz, Gleichheit und Gerechtigkeit.

Tabelle 3: Kompetenzen (Lesen) im internationalen Vergleich und im Vergleich der deutschen Bundesländer (fett hervorgehoben) zwischen 2000 und 2022 (ohne China und Singapur) nach Mehrgliedrigkeit (Anzahl der Optionen) und Stringenz (Länder nach Integration (I) und Differenzierung (T), Bundesländer nach Differenzierung mit unterschiedlichen Graden der Stringenz (T1, T2 und T3)

	Bildungs-system		Kompetenzen Lesen Mittelwerte					
	Anzahl Optionen	Stringenz I/ T	2000	2009	2012	2015	2018	2022
Sachsen	**3**	**T3**	491	508	536	528	**530**	503
Bayern	**3**	**T3**	510	509	517	513	**524**	496
Estland		**I**	501	501	516	519	523	510
Finnland		**I**	546	536	524	527	520	484
Kanada		**I**	525	524	523	517	520	497
Korea		**I**	482	520	536	511	514	527
Thüringen	**3**	**T2**	482	497	494	500	**507**	482
Schweden		**I**	516	497	497	497	506	482
USA		**I**	504	500	498	498	505	465
GB		**I**	523	494	490	516	504	489
Japan		**I**	522	539	538	496	504	536
Baden-Württemberg	**3**	**T3/T2**	500	504	500	496	**503**	486
Norwegen		**I**	505	503	504	509	499	468
Deutschland		**T**	484	497	508	511	498	475
Brandenburg	**3**	**T2**	459	485	485	504	**493**	472
Sachsen-Anhalt	**2**	**T2**	455	496	483	492	**493**	483
Hessen	**2**	**T1**	478	492	495	498	**491**	460
Niedersachsen	**2**	**T2**	474	490	490	492	**491**	474
Nordrhein-Westfalen	**3**	**T1**	482	490	486	494	**490**	458
Rheinland-Pfalz	**2**	**T2**	485	497	497	496	**490**	466
Hamburg	**2**	**T2**	478	484	489	496	**488**	472
OECD			496	493	496	476	487	472
Schleswig-Holstein	**2**	**T2**	478	488	497	514	**486**	476
Niederlande		**T**	532	508	499	503	485	493
Österreich		**T**	507	470	504	485	484	487
Mecklenburg-Vorp.	**2**	**T2**	467	493	488	506	**482**	477
Schweiz		**T**	494	501	494	492	482	508
Saarland	**2**	**T2**	484	492	493	496	**481**	478
Berlin	**2**	**T1**	461	480	479	483	**479**	451
Island		**I**	507	500	483	482	474	459
Israel		**I**	439	474	486	479	470	458
Bremen	**2**	**T1**	448	469	471	458	**460**	436

In der Tabelle sind China und Singapur ausgelassen. Sie liegen z.T. weit über allen anderen. Die genauen Gründe dafür sind nicht geklärt, darunter auch Fragen an die Datenerhebung dort. Es wären die gleichen Bedingungen wie für die asiatischen Länder insgesamt: formal integriert und mit einigen Eigenschaften der stärkeren Kontrolle, des Personals, vor allem aber, wie auch in Korea und Japan, ein extrem rigides Selektionssystem nach der Pflichtschulzeit und nicht zuletzt die besondere Leistungskultur in Asien, also als formal integrierte Systeme darin den (stringent) differenzierenden Systemen weit voraus in Leistungsdruck und Prüfungsstress. In Baden-Württemberg wurde 2012/13 die Verbindlichkeit abgeschafft, daher die Rückstufung von T3 zu Beginn auf T2 danach (gekennzeichnet über T3/T2). Die politische Absicht dabei war die weitere Erhöhung des Leistungsniveaus, vor allem aber die Verringerung der sozialen Ungleichheit über die Stärkung des Elternwillens gegenüber den Empfehlungen der Schule

Betrachtet wird die Rangordnung der OECD-Länder und deutschen Bundesländer für das Jahr 2018, dem letzten „regulären“ Jahr vor den verschiedenen Krisen (Spalte für 2018, grau hervorgehoben). Es ist der Stand fast s 20 Jahre nach der ersten PISA-Erhebung 2000 mit dem bekannten Ergebnis. Damals hatten die Länder mit Integration die mit Differenzierung in nahezu jeder Hinsicht ausgestochen, Finnland ganz vorne, aber auch Kanada und Schweden und weit dahinter Deutschland, Österreich und die Schweiz. Aber schon da (Spalte 2000) war, wie man sehen kann, die Sache differenzierter, jedenfalls für Deutschland und die deutschen Bundesländer. Es gab zwei, die bei PISA 2000 über oder auf dem OECD-Durchschnitt lagen: Bayern (510) und Baden-Württemberg (mit 500), Bayern sogar schon an Schweden heran reichend. Das drittbeste Bundesland damals war Sachsen (mit 491), deutlich vor dem nächsten, Rheinland-Pfalz (mit 485), und den anderen dann darunter, z.T. ganz erheblich. Die drei besten Bundesländer damals hatten zwei Gemeinsamkeiten: Die *Drei*gliedrigkeit und die *Stringenz* (T3), der Kombination von Verbindlichkeit und Kontrolle bei Differenzierung (vgl. die entsprechenden Einträge in der Tabelle links). Es waren auch die einzigen Bundesländer, die nach den Umstellungen in den 80er Jahren keine öffnende Reform erlebt hatten (vgl. Helbig und Nikolai 2015). Und so sieht es dann aus: Fehlte auch nur eine dieser Bedingungen, ging es abwärts, damals schon.

Für 2018, der letzten der PISA-IQB-Erhebungen *vor* den Krisen ab etwa 2020, aber auch schon nach den Flüchtlingsbewegungen seit 2015, sieht es noch einmal anders aus. Sachsen und Bayern nun ganz vorn, auch international – *vor* Estland noch, wenngleich Bayern nur ganz knapp, und dann *vor* Finnland, Kanada, und Korea, ganz zu schweigen von Schweden, Japan oder Norwegen, allesamt mit ihren integrierten Systemen. Dritter der deutschen Bundesländer

ist Thüringen, auch mit Dreigliedrigkeit, aber ohne die volle Stringenz (T2), vierter ist Baden-Württemberg, das, wie erwähnt, 2012/13 die Verbindlichkeit abgeschafft hatte und so auch die volle Stringenz aufgegeben hatte. Dann geht es wieder deutlich bergab, besonders bei den Bundesländern mit voller Öffnung und Lockerung (T1), Hessen, Nordrhein-Westfalen, Berlin und ganz unten Bremen.

Soweit die Sachlage zur Effizienz. Nun die Befunde zur Stratifikation, zur sozialen Durchlässigkeit also, und zu den Unterschieden für die Kinder in den unteren Leistungsbereichen (Tabelle 4).

Tabelle 4: Leistungen im unteren Bereich und die Effekte der sozialen Herkunft (sozialer Gradient) für die deutschen Bundesländer von 2009 bis 2022 (IQB-Berichte); fett: höchste Werte, fett kursiv: geringste Werte, jeweils bei Leistungsniveau und sozialer Durchlässigkeit; für 2022 wurden die Leistungen im unteren Bereich über den Anteil beim Erreichen des unteren Mindeststandards gemessen),

		Leistungen unterer Bereich (10. Perzentil/ Mindeststandard erreicht)				Effekt soziale Herkunft (sozialer Gradient)			
		2009	2015	2018	2022	2009	2015	2018	2022
Sachsen	**3/T3**	**382**	**415**	**412**	**91,6**	28	32	***42***	**37**
Bayern	**3/T3**	**390**	**387**	**394**	88,5	34	**30**	**34**	***46***
Baden-Württemberg	**3/T3, 3/T2**	**391**	363	381	86,6	**32**	32	40	42
Sachsen-Anhalt	**2/T2**	380	**385**	384	**90,9**	28	37	39	39
Thüringen	**3/T2**	387	**385**	**390**	**89,9**	29	34	**29**	**36**
Saarland	**2/T2**	373	368	**359**	82,8	**36**	27	37	41
Mecklenburg-V.	**2/T2**	379	374	360	88,3	30	35	36	**37**
Schleswig-Holstein	**2/T2**	***361***	393	361	87,5	27	32	42	39
Niedersachsen	**2/T2**	362	380	380	85,5	**32**	**27**	**33**	**37**
Brandenburg	**2/T2**	368	386	379	87,7	27	39	**34**	**32**
Hamburg	**2/T2**	353	353	354	83,5	**32**	***43***	...	...
Rheinland-Pfalz	**2/T2**	373	364	364	81,2	31	**29**	***45***	39
Hessen	**3/T3**	373	***352***	389	83,2	**30**	36	41	39
Nordrhein-Westfalen	**3/T3**	376	360	364	***80,3***	**32**	37	41	***44***
Berlin	**2/T3**	***342***	***333***	**345**	***77,7***	***42***	***42***	***46***	***47***
Bremen	**2/T3**	***330***	***309***	**338**	***75,6***	***36***	***44***	***42***	43

Die Befunde zu den Kindern mit Lernschwächen, denen im 10. Quintil der Verteilung der Leistungen von unten nach oben, sind besonders überraschend. *Anders* als in Integrations- und Standardposition angenommen sind es *nicht* die offenen Länder, die hier helfen, ganz im Gegenteil: Sachsen und Bayern haben jeweils die höchsten Werten für die schwächeren Kinder

und sind sozial nicht undurchlässiger als Nordrhein-Westfalen, Berlin und Bremen. Bemerkenswert ist wieder Baden-Württemberg: Nach der Öffnung von 2012/13 fallen die Leistungswerte gerade für die Kinder mit Lernschwächen, und zwar durchaus dramatisch: von 391 in 2009 auf 363 und 381, weit hinter Sachsen mit 415 und 412, und Bayern immer noch bei 387 und 394 jeweils für 2015 bzw. 2018. Für 2022 sieht das nicht anders aus, nun für das Erreichen des Mindeststandards.

Zu erwähnen wäre in diesem Zusammenhang noch, dass es für die *Grundschulen* nicht anders ausgesehen hat (PISA-E 2000, IQB 2015 und 2021): Seit Beginn in 2000 waren die Leistungen dort in den stringenten und mehrgliedrigen Ländern ebenfalls (deutlich) besser und lagen gerade hier sogar teilweise auch wieder über dem internationalen Standard. Und es waren auch hier stets die offenen Bundeländer, die schon gleich zu Beginn der Bildungsbiographie zurückgeblieben sind.

Das Ergebnis ist in so gut wie jeder Hinsicht eindeutig – und unerwartet: Die Bundesländer mit Dreigliedrigkeit, Stringenz *und* bildungspolitischer Stabilität sind deutlich besser, von Anfang und bis zu den Krisen, und das auch gegen den (internationalen) Trend der Abnahme von Niveau und Durchlässigkeit seit etwa 2012 als die Systeme mit Zusammenlegung, Öffnung und Lockerung. Es war allen Berichten zum Vergleich der Bundesländer über die Jahre zu entnehmen und hätte schon früh Anlass geben können, gerade nach der Erfolgsschichte bis 2012 den einmal eingeschlagenen Pfad der Stringenz nicht zu verlassen: Never Change a Winning System!

6.2 Zweigliedrigkeit und Stringenz

Die deskriptiven Befunde müssen nicht auch schon jene sein, die man nach der Kontrolle der relevanten Hintergrundbedingungen finden würde, geschweige denn, dass es *kausale* Effekte der Eigenschaften der Bildungssysteme wären. Das kommt nun: Die Bestimmung der Systemeffekte unter Kontrolle der in Tabelle aufgeführten Hintergrundbedingungen. Die Analysen beziehen sich auf die beiden Stadien des Bildungsverlaufs beim Übergang: Die Leistungen in der Grundschule schon und die danach in der Sekundarstufe.

Grundschule

Das theoretische Modell des MoAbiT geht davon aus, dass sich bestimmte Regelungen der Bildungssysteme im Prinzip auf alle Phasen und Bereiche des Bildungsverlaufs auswirken können, auch schon früh in der Familie, in der Vorschule und insbesondere in der Grundschule kurz vor dem Übergang. Einer der Mechanismen dabei kann die *Antizipation* der jeweiligen Folgen sein, etwa mit der Verbindlichkeit der schulischen Empfehlungen: Wenn es keine freie Wahl gibt, wie bei der Stringenz mit Verbindlichkeit und Kontrolle, wären die (Opportunitäts-)Kosten eines Verzichts auf vorherige Anstrengungen, den Übergang zu schaffen, höher. Die (Opportunitäts-)Kosten würden noch einmal zunehmen, wenn die Stratifikation unterhalb des gymnasialen Zweigs größer ist, wie bei der Dreigliedrigkeit gegenüber der Zweigliedrigkeit. Es wird nach dem MoAbiT also erwartet, dass sowohl die Stringenz wie die Dreigliedrigkeit zu einer Erhöhung der Leistungen in der Grundschule führt, womöglich verstärkt in der Kombination beider Regelungen.[3]

Tabelle 5 beschreibt die Befunde dazu, einmal im Vergleich der beiden Regelungen jeweils für sich, Zweigliedrigkeit und Stringenz (Spalten 1 und 2) und dann die Effekte der Stringenz, getrennt nach Zwei- und Dreigliedrigkeit für den konditionalen Effekt (Spalten 3 und 4). Kontrolliert sind dabei die beiden zentralen Einflüsse auf die Leistungen, auch schon in der Grundschule: die soziale Herkunft und die kognitiven Fähigkeiten.

[3] Zu den Effekten der *Stringenz* nach dem MoAbiT gibt es seit einiger eine Reihe von empirischen Belegen: Ein früher Vergleich von Bayern und Hessen für die Leistungen in der Sekundarstufe (Esser und Relikowski 2015, später nach der Verfügung über die Daten der NEPS für die Bundesländer insgesamt zu den Leistungen in der Grundschule und die Bildungsbeteiligung: Esser und Hoenig 2018, und zu den Leistungen in der Sekundarstufe. Esser und Seuring 2020, Esser 2023, Kapitel 5 bis 8. Für eine Zusammenfassung und Bewertung der wichtigsten theoretischen Überlegungen und Befunde vgl. Münch (2024), zu einigen Repliken und Gegenrepliken siehe noch Abschnitt 7 unten.

Tabelle 5: Systemeffekte der Mehrgliedrigkeit (M) und der Stringenz (T) auf die Leistungen am Ende der Grundschule: lineare Regression; Kontrolle Geschlecht, Migrationshintergrund, Vorschulbesuch; fett: p<0.05; Abkürzungen nach Tabelle 1.

	Effekte Zweigliedrigkeit vs. Stringenz Grundschule			
	Vergleich Zweigliedrigkeit vs. Stringenz		Vergleich Stringenz getrennt nach Zwei- und Dreigliedrigkeit	
	1	2	3	4
	Zweigliedrigkeit	Stringenz	Zweigliedrigkeit	Dreigliedrigkeit
Systemeffekt (n. kond.)	0.01	**0.03**	-0.01	**0.04**
Kontrolle SES Kontrolle ABL	x x	x x	x x	x x
c R^2 n (Kinder)	 0.33 2636	 0.34 2636	 0.37 518	 0.34 2118

Das Ergebnis ist eindeutig: Für die Zweigliedrigkeit gibt es keinerlei Effekt, jeweils für sich, aber auch nicht in Kombination mit der Stringenz. Das ist für die Stringenz anders. Sie hat für sich schon den angenommenen Antizipationseffekt (mit 0.03), und der verstärkt sich noch mit der der Dreigliedrigkeit (auf 0.04).

Die Effekte sind zwar von der Stärke her nicht besonders groß. Aber sie entsprechen in der Tendenz dem, was man inzwischen dazu weiß: Es *gibt* früh schon, in Familie und Schule, Antizipationen bestimmter Folgen mit der Verbindlichkeit bzw. der Stringenz, und die wirken als Anreize für besondere Anstrengungen, bei den Kindern wie bei den Familien, schon *vor* dem Übergang in die Sekundarstufe (vgl. Bach und Fischer 2020, Osikominu et al. 2020, Grewenig 2022). Und das erzeugt – unabhängig, was sonst noch kommen mag – als nicht-konditionaler Systemeffekt schon vorher einen Leistungsvorsprung für die Sekundarstufe, noch, ohne dass dort noch weitere Effekte geben müsste. Mit den Anstrengungen können gewiss auch hemmende und entmutigende Spannungen verbunden sein. Aber die Folgen davon sind angesichts des *positiven* Bruttoeffekts offenbar geringer als die der Anreize zur erhöhten Anstrengung. Ergänzt sei noch, dass der Effekt der sozialen Herkunft etwa 0.35 beträgt, jener der kognitiven Fähigkeiten aber etwa das dreifache.

Sekundarstufe

Die Effekte der Bildungssysteme auf die Leistungen in der Sekundarstufe bestehen, wie in der Grundschule, zunächst aus den *nicht-konditionalen* Effekten der Systemeigenschaften, ermittelt über Analysen der Mediation mit der Kontrolle nach den *allgemeinen Bedingungen*. Daneben gibt es, speziell für die Sekundarstufe, die *konditionalen* Systemeffekte der *Änderungen* der allgemeinen Einflüsse, die Moderation der allgemeinen Effekte in der Sekundarstufe also. Die sollten sich ach dem theoretischen Modell vor allem auf die Schulklasseneffekte der kognitiven Zusammensetzung nach Niveau und Homogenität beziehen, modelliert als Interaktionseffekt mit den jeweiligen Systemeigenschaften.

In Tabelle 6 ist das als *konditionaler* Systemeffekt für die Zweigliedrigkeit und die Stringenz eingetragen, auf den sich die Interaktionseffekte der kognitiven Einflüsse mit den Systemeigenschaften jeweils beziehen. Und dann als die ebenfalls *konditionalen* Effekte der *Interaktion* der Systemmerkmale mit den allgemeinen Bedingungen nach dem MoAbiT aus Abschnitt 4 (Spalte 1 für die Dreigliedrigkeit (D), Spalte 2 für die Stringenz (T).

Tabelle 6: Systemeffekte der Dreigliedrigkeit (D) und der Stringenz (T) auf die Leistungen in der Sekundarstufe, 7. Klasse: lineare Regression; alle Kontrollen, fett: $p<0.05$; Abkürzungen nach Tabelle 1.

	Effekte Zweigliedrigkeit vs. Stringenz Sekundarstufe		
	1		2
Systemeffekte	Dreigliedrigkeit (D)		Stringenz (T)
nicht konditional konditional	-0.04 1.44	nicht konditional konditional	**0.18** -1.00
D*SES	0.30	T*SES	0.04
D*NSES	-2.85	T*NSES	-0.58
D*HSES	-2.31	T*HSES	-0.63
D*NSES*HSES	4.03	T*NSES*HSES	0.11
D*ABL	-0.10	T*ABL	0.02
D*NABL	0.06	T*NABL	**2.94**
D*HABL	0.02	T*HABL	**2.47**
D*NABL*HABL	-0.18	T*NABL*HABL	**-4.17**
D*GYM	0.15	T*GYM	0.20
c	-3.45	c	-2.22
R^2: Level 1	0.14	R^2: Level 1	0.14
R^2: Level 2	0.85	R^2: Level 2	0.87
n (Kinder)	2636	n (Kinder)	2636
N (Schulklassen)	313	N (Schulklassen)	313

Für die Dreigliedrigkeit gibt es keinerlei Systemeffekte im Vergleich zur Zusammenlegung. Von den Effektstärken her gibt zwar einige Abschwächungen der Effekte der sozialen Zusammensetzung der Schulklassen, aber die sind weit unterhalb der Signifikanzgrenze. Und es gibt keinerlei Hinweise auf Effekte für die kognitive Zusammensetzung. Also eigentlich das was man auch sonst meist gefunden hat: *keine* Effekte der Zusammenlegung von Haupt- und Realschulen).

Das ist bei der Stringenz anders (Spalte 2). Hier gibt es – nach allen Kontrollen der anderen Bedingungen (s. dazu Tabelle 1) – schon einen signifikanten nicht konditionalen Systemeffekt (von 0.18). Das ist im Wesentlichen wohl der Antizipationseffekt aus den Leistungssteigerungen in der Grundschule. Bei der Zweigliedrigkeit gibt es den nicht. Daneben

sind es bei der Stringenz vor allem die Moderationseffekte aus der kognitiven Zusammensetzung der Schulen: Deutliche positive Effekte des kognitiven Niveaus (mit 2.94) und der kognitiven Homogenität (mit 2.47), und das als Vorteil gerade in den *unteren* Bereichen der kognitiven Zusammensetzung (mit -4.17), was anzeigt, dass die positiven Effekte von Niveau und Homogenität beiden Kindern in den unteren Leistungsbereichen auftreten. Es wäre nachgerade das Gegenteil der Hypothesen der Integrationsposition, die von Vorteilen der Heterogenität ausgeht und von Schereneffekten nach oben bei Stringenz und Homogenität. Diese Befunde sind bekannt (Esser und Seuring 2020, Esser 2023, Kapitel 7; s dazu auch noch Abschnitt 7 unten)).

Die Konstellation der Interaktion von Mehrgliedrigkeit und Stringenz ist für die Schätzung der Effekte mit den Dreifach-Interaktionen allerdings nicht ohne Risiko und auch recht unübersichtlich (vgl. aber die graphischen Darstellungen zu den Stringenzeffekten bei Esser und Seuring 2020: 296, Esser 2023: 457f.). Stabilere Schätzungen lassen sich in einem nach den Systemvarianten getrennten Vergleich erwarten (Tabelle 7), wobei allerdings die Unterschiede nach dem Leistungsniveau, die nicht-konditionalen Systemeffekte also, nicht mehr vorkommen (die Ergebnisse dazu finden sich in den Tabelle 5 und 6 oben schon).

Tabelle 7: Systemeffekte der Mehrgliedrigkeit und der Stringenz auf die Leistungen in der Sekundarstufe, 7. Klasse; getrennte Analysen; lineare Regression; Kontrolle Geschlecht, Migrationshintergrund, Vorschulbesuch; fett: p<0.05; Abkürzungen nach Tabelle 1.

	Zweigliedrigkeit		Dreigliedrigkeit	
	1	2	3	4
	Stringenz niedrig	Stringenz hoch	Stringenz niedrig	Stringenz hoch
SES	0.20		**0.42**	**0.41**
NSES	**4.12**		-0.47	0.49
HSES	**2.36**		-0.88	-029
NSES*HSES	**-4.99**		1.64	-0.64
ABL	**1.48**	empirisch nicht besetzt	**1.38**	**1.42**
NABL	0.28		0.24	**3.20**
HABL	-0.23		-0.49	**1.99**
NABL*HABL	0.84		-0,49	**-3.47**
GYM	**0.42**		**0.54**	**0.69**
c	-3.43		-1.32	-3.03
R^2: Level 1	0.15		0.15	0.12
R^2: Level 2	0.90		0.82	0.93
n (Kinder)	367		1265	1004
N (Schulklassen)	56		142	115

Nun werden die Unterschiede in den entscheidenden Aspekten der Effekte der kognitiven Zusammensetzung der Schulklassen in der Sekundarstufe deutlich erkennbar: Es gibt diese Effekte *nur* in der Kombination von Stringenz und *Drei*gliedrigkeit (Spalte 4). Es ist die Bestätigung des theoretischen Modells, wonach ohne die Stringenz die Dreigliedrigkeit keinen weiteren Effekt in der Sekundarstufe hat.

Bemerkenswert sind insbesondere noch die Befunde zur sozialen Herkunft. Mit der Dreigliedrigkeit bleiben, unabhängig von der Stringenz, nur noch Effekte der individuellen sozialen Herkunft (mit 0.43). Das entspricht dem, was man über die Herkunftseffkte allgemein kennt. Das ist anders bei der Zweigliedrigkeit. Hier gibt es *keine* Effekte der *individuellen* sozialen Herkunft, wohl aber erhebliche Einflüsse der *sozialen Segregation* in den Schulklassen (mit 4.12 für das soziale Niveau, 2.36 für die soziale Homogenität) und eine Umkehrung der Effekte zu *un*gunsten der Kinder in Schulklassen mit einem *geringen* sozialen Niveau und einer *höheren* sozialen Heterogenität (um -4.99). Das aber heißt, dass die Zweigliedrigkeit die

Effekte der sozialen Segregation gerade bei den Schulklassen eher verstärkt als mildert oder gar abschafft. Und zwar deutlich.

Es sieht demnach so aus, dass die *individuellen* sozialen Ungleichheiten mit der Zusammenlegung nicht wie erhofft verschwinden, sondern sich auf die *kontextuelle* Ebene verlagern. Das entspräche den o.a. Befunden von Dollmann und Rudolphi (2020: 229f.) und von Engzell und Raabe (2023) und für einen Vergleich von Deutschland Schweden in Bezug auf die Netzwerkstrukturen: In Deutschland ist die Drehscheibe für die Leistungen die soziale Homogenität der *Schulklasse*, in Schweden ist es die soziale Homogenität der *peer-Netzwerke innerhalb* der bei Integration zunächst einmal sozial heterogeneren Schulklassen. Aber es würde ja nicht viel helfen, wenn so wäre: Man könnte zwar die Zusammensetzung der Schulklassen durch Regelungen ändern, aber nicht die Netzwerke, die sich dann alsbald wieder nach der sozialen Homophilie bilden würden. Auch im NEPS fehlen die Daten dazu, um dem nachzugehen.

In der Tabelle ist das Feld für die Kombination von Stringenz und Zweigliedrigkeit leer (Spalte 2): Es gibt zwar dreigliedrige nicht stringente Systeme bei den Bundesländern (wie Hessen oder Nordrhein-Westfalen), aber *keine* zweigliedrigen Systeme *mit* Stringenz). Wenn überhaupt waren nach der Logik der Integrationsposition Öffnung und Lockerung des Zugangs und der Kontrollen der erste Schritt, dem dann die Zusammenlegung folgte, wie in Baden-Württemberg 2012/13 mit der Abschaffung der Verbindlichkeit und dann 2025/26 mit der „Neuen Sekundarschule“ und der Zusammenlegung im im unteren Bereich. Den umgekehrten Fall gab es dagegen empirisch nicht: Die Verstärkung der Stringenz bei schon bestehender Zweigliedrigkeit. Es „passt“ in die gängigen Vorstellungen von Integration und „Gemeinschaftsschulen“ nicht so recht, dass man in den zusammengelegten Schulformen jetzt plötzlich mehr auf Leistung, kognitive Fähigkeiten, Verbindlichkeit und Kontrolle setzen sollte. In einigen Bundesländern versucht man das jetzt nach dem Desaster mit PISA 2022.

7. Limitationen

Das NEPS bietet im Vergleich zu den herkömmlichen Studien, besonders mit Blick auf die Vollständigkeit der nötigen Konstrukte, eine vergleichsweise tragfähige Grundlage, teilweise auch mit der Möglichkeit empirischer Analysen, wie es sie bisher in diesem Feld nicht gegeben hat. Insoweit müsste man eigentlich nach den Begrenzungen von so gut wie allen anderen Untersuchungen fragen. Zwei Limitationen allerdings sind zu benennen: Die Reichweite der

Aussagen und die Breite und Repräsentativität der Daten, insbesondere auch für die Analyse von Effekten der strukturellen Bedingungen der Schulklassen und der Schuleffekte (Heisig und Matthewes 2022, Lorenz et al. 2023). Dazu gibt es ausführliche Erwiderungen (Esser 2023, Kapitel 9, Esser und Seuring 2023, Esser 2024b), die an dieser Stelle, wieder aus Platzgründen, nicht noch einmal dokumentiert werden müssen. Antworten darauf hat es (bisher) nicht gegeben. Abgesehen davon, dass die Fallzahlen in der Tat an der Grenze des Vertretbaren liegen, ergaben sich, auch in weiteren Re-Analysen, keine Hinweise auf systematische Verzerrungen, eher im Gegenteil. Der Eindruck darauf entstand vielmehr erst darüber, dass in den Replikationen entweder eine ausreichende Vergleichbarkeit nicht gegeben war oder den ursprünglichen Ansatz verfälschende Modifikationen bei der Bestimmung der empirischen Estimanden vorgenommen wurden, wie Überkontrollen in den Analysen zur Mediation und unzulässige Vereinfachungen bei den Interaktionseffekten in den Analysen zur Moderation.

8. Zusammenfassung und Bewertung

Ausgangspunkt des Beitrags war eine nicht nur auf den ersten Blick bemerkenswerte Entwicklung der Positionierung des deutschen Bildungssystems mit seiner frühen und rigiden Differenzierung: Entgegen dem Eindruck aus den frühen internationalen Vergleichsstudien gab es von Beginn eine deutliche Differenzierung der Befunde bei der Differenzierung zwischen den Bundesländern: Die besten Leistungen und eine eher auch höhere soziale Durchlässigkeit hatten, von Anfang an, die Bundesländer mit den strengsten Regelungen in Zugang und Organisation einerseits und der traditionellen Dreiteilung nach Hauptschule, Realschule und Gymnasium andererseits. Das waren auch jene Bundesländer, in denen es über eine längere Zeit keine Reformen gegeben hatte. Am schlechtesten waren dagegen, ebenfalls von Anfang an, die Bundesländer mit Öffnungen, Lockerungen Zweigliedrigkeit ohne Haupt- und Realschulen unterhalb des Gymnasiums und beständigen Reformversuchen, um dem Niedergang beizukommen. Das war so auch bei den Grundschulen schon.

Nahezu alle gingen jedoch nach PISA 2000 wie selbstverständlich davon aus, dass die Umstellung auf die Integration das Ziel sein müsse. Aber es geschah, weitgehend unbemerkt, etwas anderes: Nach dem PISA-Schock hatten sich die Leistungen verbessert und auch die soziale Durchlässigkeit, sehr zum Erstaunen mancher der Kommentatoren (vgl. Klieme et al. 2010: 281f.). Gründe dafür waren wohl die hohe öffentliche Aufmerksamkeit nach dem PISA-

Schock und die besondere Bekümmerung um den Ausgleich von Benachteiligungen, etwa beim Vorschulbesuch. Und so hätte es weiter gehen können. Es kam indessen anders: Seit etwa 2012 ging es wieder bergab. Das jedoch wieder nicht überall. Zwei der drei Spitzenreiter nach PISA 2000, Bayern und Sachsen, behielten ihren Kurs bei und zogen weiter nach oben davon. Bei den anderen gab es Stagnation oder einen weiteren Rückgang, speziell in Baden-Württemberg, wo 2012/13 die Verbindlichkeit abgeschafft und die sog. Sekundarschulen eingeführt worden waren. Man hätte also auf den Gedanken kommen können, den – aus vielen Gründen gespeisten – Tendenzen zur weiteren Ent-Differenzierung und Öffnung entgegenzutreten. Das war aber nicht so, und insbesondere die Umstellung der unteren Bildungswege auf die Zweigliedrigkeit hat sich eher noch verstärkt.

Im Beitrag wurde untersucht, was denn dran ist an der Hoffnung, mit Zweigliedrigkeit und Öffnung die Probleme endlich (besser) in den Griff bekommen zu können. Die Untersuchung hat gezeigt, dass es dazu keinen Anlass gibt. Ganz im Gegenteil: Im internationalen Vergleich sind die beiden verblieben Bundesländer mit Dreigliedrigkeit und Stringenz, Sachsen und Bayern, an der Spitze, noch vor Kanada, Finnland, Schweden, Norwegen und auch Estland. In der vergleichenden Analyse mit den Daten des NEPS zeigt sich, dass die Mehrgliedrigkeit so gut wie keinen Einfluss hat und dass es den stärksten Effekt auf die Verbesserung der Leistungen in der *Kombination* von Stringenz und Dreigliedrigkeit gibt.

Die in den Abschnitten 2 und 4 oben angesprochene Untersuchung von Matthewes (2021) zum Vergleich der Bundesländer, die die gleichen Daten verwendete, scheint dem zu widersprechen. Dort worden positive Effekte der Zweigliedrigkeit und das speziell zugunsten der Kinder mit schwachen Leistungen in der Grundschule gefunden. Es kann aber bezweifelt werden, ob der dort verwendete Ansatz den Schluss auf belastbare (kausale) Effekte überhaupt erlaubt: Mit dem dort gewählten Ansatz zur Kontrolle der unbeobachteten Heterogenität der Bundesländer und der Einwicklungen insgesamt (über das *difference-in-difference*-Verfahren) sind gravierende Probleme zur korrekten Schätzung der Effekte verbunden: Collider-Probleme, die Selektivität bei der Aufteilung in die Bildungswege, die Berücksichtigung der Leistungen vorher mit Antezipationen und Anreizen für Anstrengungen vor (und auch danach) und die Variabilität der Beziehungen mit den jeweiligen Systemen genau danach – die Aspekte also, die den Kern der Ergebnisse in der Untersuchung hier ausgemacht haben. Zudem gibt es nur eine isolierte Betrachtung der Mehrgliedrigkeit ohne Bezug zur Stringenz, dem deutlich

stärkeren Systemeffekt, die, wie es aussieht, eine *Bedingung* für evtl. positive Effekte der Dreigliedrigkeit darstellt. [4]

Insgesamt bestätigen die Ergebnisse den seit Beginn an zu beobachtenden Trend einer Konvergenz und der Zunahme von integrativen Differenzierungen und Differenzierungen in der Integration der Länder und Bildungssysteme mit allerlei Konditionalisierungen und funktionalen Äquivalenten mit Folgen für die Einebnung auch der Unterschiede in den Leistungen und der sozialen Durchlässigkeit (vgl. so schon Jackson und Jonsson 2013. 325f, 329f. für die primären Effekte und die Regelungen der Schulwahl; s. aktuell dazu Schindler et al. 2024).

Eigentlich wäre es, nach allem was sich abgezeichnet hat, also wohl besser gewesen, nichts weiter zu ändern, schon gar nicht bei den immer schon gut funktionierenden Bundesländern: „Never Change a Winning System!“ könnte man meinen Und gerade aktuell wäre die Rückkehr zu dem traditionellen System der Stringenz der Differenzierung und der Dreigliedrigkeit angeraten, jedenfalls dann, wenn es um Niveau, Leistungsgerechtigkeit und soziale Chancengleichheit geht, ergänz natürlich um alle die nur zu begrüßenden allgemeinen Massnahmen nach dem erneuten PISA-Schock: Sanierung der Infrasturktur, Frühdiagose und Frühförderung, Ganztag, auch „multiprofessionelle Teams“ oder „KI“, wenn das dann hilft. Es gehört freilich nicht viel Phantasie dazu, sich auszumalen wie realistisch das ist und welche Invektiven zu vergegenwärtigen sind, wenn auf die Befunde hingewiesen wird. Eines könnte man sich aber wohl immer noch vornehmen, vielleicht klammheimlich und einem geschickteren Framing vorerst: Jene Strukturen *nicht* weiter auszuhöhlen, von denen man *weiß*, dass sie besser (gewesen) sind. Aber selbst das, so steht zu befürchten, wird nicht mehr lange so gehen: Sachsen, der Spitzenreiter nicht erst zuletzt, hat 2017, auf Druck von Eltern und Gerichten, die Verbindlichkeit abgeschafft. Bisher galt bei solchen und anderen Lockerungen eher auch ein anderer Grundsatz: „They Never Come Back!“. Man wird sehen.

[4] Die ebenfalls in Abschnitt 2 angesprochene Untersuchung von Piopiunik (2014) bezog sich auf eine Reform in einem Bundesland mit der Einführung der Dreigliedrigkeit, ist also eigentlich nicht vergleichbar. Sie teilt die genannten Probleme, aber nur teilweise: das Fehlen der kognitiven Einflüsse, womöglich dann verbunden mit Verzerrungen über die damit nach hinten offene Collider-Struktur.

Literatur

Autor*inengruppe Neue Sekundarschule. 2024. Neue Sekundarschule in Baden-Württemberg. Begründung, Ausgestaltung und Einführung. Ein Vorschlag zur Neugestaltung der Schulstruktur im Kontext der derzeitigen Diskussion um die Einführung des G9.

Bach, M., und M. Fischer. 2020. Understanding the Response to High Stakes Incentives in Primary Education. IZA-Discussion Paper Nr. 13845. Bonn: Forschungsinstitut Zukunft der Arbeit. Institute for the Study of Labor.

Baumert, J., P. Stanat und R. Watermann. 2006. Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus. In: J. Baumert, P. Stanat und R. Watermann, Hrsg., Herkunftsbedingte Disparitäten im Bildungswesen. Vertiefende Analyse im Rahmen von PISA 2000. Wiesbaden: VS Verlag für Sozialwissenschaften: 95-188.

Baumert, J., M. Becker, M. Neumann und R. Nikolova. 2010. Besondere Förderung von Kernkompetenzen an Spezialgymnasien? Der Frühübergang in grundständige Gymnasien in Berlin. Zeitschrift für Pädagogische Psychologie, 24: 5-22.

Baumert, J., K. Maaz, M. Neumann, M. Becker, H. Dumont, S. Böse und M. Kropf. Die Berliner Schulstrukturreform – Zusammenfassung und Ausblick. 2013. In: M. Neumann, M. Becker, J. Baumert, K. Maaz und O. Köller, Hrsg., Die Berliner Schulstrukturreform. Bewertung durch die beteiligten Akteure und Konsequenzen des neuen Übergangsverfahrens von der Grundschule in die weiterführenden Schulen. Münster und New York: Waxmann: 263-281.

Baumert, J., O. Köller, M. Neumann und K. Maaz. 2017. Kompetenzarmut im mehr- und zweigliedrigen System. In: M. Neumann, M. Becker, J. Baumert, K. Maaz und O. Köller, Hrsg., Zweigliedrigkeit im deutschen Schulsystem. Potentiale und Herausforderungen in Berlin. Münster und New York: Waxmann: 189-226.

Becker, M., M. Neumann, S. Radmann, M. Janson, G. Nagy, Ch. Borzikowski, M. Leucht, K. Maaz, O. Köller und J. Baumert. 2017. Schulleistungen vor und nach der Berliner Strukturreform. In: M. Neumann, M. Becker, J. Baumert, K. Maaz und O. Köller. Zweigliedrigkeit im deutschen Schulsystem. Potentiale und Herausforderungen in Berlin. Münster und New York: Waxmann: 155-188.

Betts, J.R., und J. L. Shkolnik. 2000. Key difficulties in identifying the effects of ability grouping on student Achievement. In: Economic Education Review, 19, 21–26.

Blossfeld, H. P., H.G. Roßbach und J. von Maurice. 2011. Education as a Lifelong Process – The German National Educational Panel Study (NEPS). Wiesbaden: Springer VS.

Bohl, Th., J. Budde und M. Rieder-Ladich. 2017. Einleitung. In: Bohl, Th., J. Budde, M. Rieger-Ladich, Hrsg., Umgang mit Heterogenität in Schule und Unterricht. Grundlagentheoretische Beiträge, empirische Befunde und didaktische Reflexionen. Bad Heilbrunn: Verlag Julius Klinkhardt: 7-12.

Bol, Th., J. Witschge, H. G. van de Werfhorst und J. Dronkers. 2014. Curricular tracking and central examinations: Counterbalancing the impact of social background on student achievement in 36 countries. In: Social Forces 92: 1545–1572.

Chmielewski, A. K. 2014. An International Comparison of Achievement Inequality in Within- and Between-School Tracking Systems. In: American Journal of Education, 120: 293–324; http:// www.jstor.org/stable/10.1086/675529.

Clarke, P., C. Crawford, F. Steele und A. Vignoles. 2010. The Choice Between Fixed and Random Effects Models: Some Considerations for Educational Research. IZA-Discussion Paper Nr. 5287. Bonn: Forschungsinstitut Zukunft der Arbeit. Institute for the Study of Labor.

Cord, D. & L. Giuliano. 2016. Can Tracking Raise the Test Scores of High-Ability Minority Students? American Economic Review, 106: 2783-2816.

Cummins, J. R. 2016. Heterogeneous treatment effects in the low track: Revisiting the Kenyan primary school experiment. In Economics of Education Review, 56: 40–51.

Deutscher Bundestag. 2006. Vor- und Nachteile der Gesamtschule bzw. des dreigliedrigen Schulsystems. Wissenschaftliche Dienste des Deutschen Bundestages. Ausarbeitung WD 8-231/2006. Fachbereich WD 8: Umwelt, Naturschutz, Reaktorsicherheit, Bildung und Forschung. Berlin

Dollmann, J., und F. Rudolphi. 2019. Classroom composition and language skills: The role of school class and friend characteristics. In: British Journal of the Sociology of Education. https://doi.org/10.1080/01425692.2020.1799754.

Domina, T., A. Penner und E. Penner. 2017. Categorical Inequality: Schools as Sorting Machines. In: Annual Review of Sociology, 43: 311–330.

Domina, T., A. McEachin, P. Hanselman, P. Agarwal, N. Hwang und R. W. Lewis. 2019. Beyond Tracking and Detracking: Dimensions of Organizational Differentiation in Schools. In: Sociology of Education, 92: 293–322.

Dräger, J., Th. Schneider, M. Olczyk, A. Solaz, A. Sheridan, E. Washbrook, V. Perinetti Casoni, S. J. Kwon und J. Waldfogel. 2003. The relevance of tracking and social school composition for growing achievement gaps by parental education in lower secondary school: a longitudinal analysis in France, Germany, the United States, and England. In: European Sociological Review, 8, 1–17. https://doi.org/10.1093/esr/jcad076.

Dronkers, J., und Jan Skopek. 2015. Performance in Secondary School in German States – A Longitudinal Three-Level Approach. Working Paper. European University Institute und ROA Maastricht University.

Duflo, E., P. Dupas und M. Kremer. 2011. Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. In: American Economic Review, 101: 1739–1774.

Dunne, A. 2010. Dividing lines. Examining the relative importance of between- and within-school differentiation during lower secondary education. PhD-Thesis, Department of Political and Social Sciences. Florence: European University Institute.

Elwert, F. und Ch. Winship. 2014. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. In: Annual Review of Sociology, 40: 31–53.

Engzell, P., und I. J. Raabe. 2023. Within-School Achievement. Sorting in Comprehensive and Tracked Systems. Sociology of Education, 96: 324-343.

Esser, H. 2016a. Bildungssysteme und ethnische Bildungsungleichheit. In: C. Diehl, Ch. Hunkler und C. Kristen, Hrsg., Ethnische Ungleichheiten im Bildungsverlauf. Mechanismen, Befunde, Debatten. Wiesbaden: Springer VS: 331–396.

Esser, H. 2016b. Educational Systems and Educational Inequality. The Model Ability Tracking and Empirical Findings. In: H. P. Blossfeld, S. Buchholz, J. Skopek und M. Triventi, Hrsg., Models of Secondary Education and Social Inequality – An International Comparison. eduLIFE Lifelong Learning Series, Band 3. Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing: 25–44.

Esser, H. 2021. „Wie kaum in einem anderen Land?“ Die Differenzierung der Bildungswege und ihre Wirkung auf Bildungserfolg, -ungleichheit und -gerechtigkeit. Band 1: Theoretische Grundlagen. Frankfurt/M. und New York: Campus.

Esser, H. 2023. „Wie kaum in einem anderen Land?“ Die Differenzierung der Bildungswege und ihre Wirkung auf Bildungserfolg, -ungleichheit und -gerechtigkeit. Band 2: Empirische Zusammenhänge. Frankfurt/M. und New York: Campus.

Esser, H., 2024a. Heterogenität und Diversität. Stellungnahme für die Enquetekommission I des Landtags NRW zu „Chancengleichheit in der Bildung“. Düsseldorf: Landtag NRW.

Esser, H. (2024b). Kein Vergleich! Zur „kritischen Betrachtung des Model of Ability-Tracking (MoAbiT)“ von Lorenz, Lenz und Rjosk (2023) in der Zeitschrift für Soziologie. Zeitschrift für Soziologie, 2024, 53: 419-425. https://doi.org/10.1515/zfsoz-2024-2024.

Esser, H. und I. Relikowski. 2015. Is Ability Tracking (Really) Responsible for Educational Inequalities in Achievement? A Comparison between the Country States Bavaria and Hesse in Germany. IZA-Discussion Paper Nr. 9082. Bonn: Forschungsinstitut Zukunft der Arbeit. Institute for the Study of Labor.

Esser, H., und K. Hoenig. 2018. Leistungsgerechtigkeit und Bildungsungleichheit. Effekte der Verbindlichkeit der Grundschulempfehlungen beim Übergang auf das Gymnasium. Ein Vergleich der deutschen Bundesländer mit den Daten der „National Educational Panel Study“ (NEPS). In: Kölner Zeitschrift für Soziologie und Sozialpsychologie, 70, 419-447.

Esser, H. und J. Seuring. 2020. Kognitive Homogenisierung, schulische Leistungen und soziale Bildungsungleichheit. Theoretische Modellierung und empirische Analyse der Effekte einer strikten Differenzierung nach den kognitiven Fähigkeiten auf die Leistungen in der Sekundarstufe und den Einfluss der sozialen Herkunft in den deutschen Bundesländern mit den Daten der „National Educational Panel Study“ (NEPS). In: Zeitschrift für Soziologie, 49: 277–301.

Esser, H., und J. Seuring. 2023. Was ist Dein Replicandum? Eine Antwort auf die Replik von Heisig und Matthewes (2022) zum Beitrag von Esser und Seuring (2020) über „Kognitive Homogenisierung, schulische Leistungen und soziale Bildungsungleichheit“. https://doi.org/10.1515/zfsoz-2023-2021.

Fend, H., 2017. Bildungsgerechtigkeit – eine Illusion? Eine Nachbetrachtung zur LifE-Studie. In: S. Lin-Klitzing, D. Di Fuccia und Th. Gaube, Hrsg., Bildungsgerechtigkeit und Gymnasium. Bad Heilbrunn: Klinckhardt: 93-111.

Figlio, D. N., und M. E. Page. 2002. School Choice and the Distributional Effects of Ability Tracking: Does Separation Increase Inequality? In: Journal of Urban Economics, 51: 497–514.

Galindo-Rueda, F., und A. Vignoles. 2007. The heterogeneous effect of selection in UK secondary schools. In: L. Wößmann und P.E. Peterson, Hrsg., Schools and the equal opportunity problem. Massachusetts: MIT Press: 103–128.

Gamoran, A. 2009. Tracking and Inequality: New Directions for Research and Practice. WCER Working Paper Nr. 2009–6. Madison: University of Wisconsin-Madison: Wisconsin Centre for Education Research.

Guyon, N., E. Maurin und S. McNally. 2012. The Effect of Tracking Students by Ability into Different Schools: A Natural Experiment. In: The Journal of Human Resources, 47: 684–721.

Grewenig, E. 2021. School Track Decisions and Teacher Recommendations. In: E. Grewenig. Human Capital and Education Policy: Evidence from Survey Data. Ifo Beiträge zur Wirtschaftsforschung, 96, Kapitel 4: 113–161.

Guyon, N., E. Maurin und S. McNally. 2012. The Effect of Tracking Students by Ability into Different Schools: A Natural Experiment. In: The Journal of Human Resources, 47: 684–721.

Hallinan, M.T. 1994. Tracking: From Theory to Practice. In: Sociology of Education, 67: 79–84.

Hanushek, E.A., und L. Wößmann. 2006. Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. In: The Economic Journal, 116: C63-C76.

Hattie, J. A. C. 2009. Invisible Learning. A Synthesis of Over 800 Meta-Analyses Relating to Achievement. London und New York: Routledge.

Hattie, J. A. C. 2023. Visible Learning: The Sequel. A Synthesis of Over 2,100 Meta-analyses Relating to Achievement. London and New York: Routledge.

Heisig, J. P., und S. H. Matthewes. 2022. Keine Belege für leistungsfördernde Effekte von strikter Leistungsdifferenzierung und kognitiver Homogenisierung: Eine kritische Reanalyse von Esser und Seuring (2022). Zeitschrift für Soziologie, 51, 99-111.

Helbig, M., und R. Nikolai. 2015. Die Unvergleichbaren. Der Wandel der Schulsysteme in den deutschen Bundesländern seit 1949. Bad Heilbrunn: Klinkhardt.

Horn, D. 2013. Diverging performances: the detrimental effects of early educational selection on equality of opportunity in Hungary, In: Research in Social Stratification and Mobility, 32: 25–43.

IQB-Berichte 2009, 2012, 2015, 2016, 2018, 2021, 2022.

Jackson, M., und J. O. Jonsson. 2013. Why does inequality of educational opportunity vary across countries? Primary and secondary effects in comparative context. In: M. Jackson, Hrsg., Determined to Succeed? Performance versus Choice in Educational Attainment, Stanford, CA: Stanford University Press: 306–337.

Jakubowski, M., H. A. Patrinos, E. E. Porta und J. Wisniewski. 2016. The effects of delaying tracking in secondary school: evidence from the 1999 education reform in Poland. In: Education Economics, DOI: 10.1080/09645292.2016.1149548.

Kerr, S.K., T. Pekkarinen und R. Uusitalo. 2013. School tracking and development of cognitive skills. Journal of Labor Economics, 31: 577 – 602.

Korthals, R. A., und J. Dronkers. 2016. Selection on performance and tracking. In: Applied Economics. DOI: 10.1080/00036846.2015.1130789.

Klieme, E., N. Jude, J. Baumert und M. Prenzel. 2010. PISA 2000–2009: Bilanz der Veränderungen im Schulsystem. In: E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider und P. Stanat, Hrsg., PISA 2009. Bilanz nach einem Jahrzehnt. Münster: Waxmann: 277–300.

Korthals, R. A., und J. Dronkers. 2016. Selection on performance and tracking. In: Applied Economics. DOI: 10.1080/00036846.2015.1130789.

Lauterbach, W., und H. Fend. 2016. Educational Mobility and equal opportunity in different German tracking systems – Findings of the LiFE study. In: H. P. Blossfeld, S. Buchholz, J. Skopek und M. Triventi, Hrsg., Models of Secondary Education and Social Inequality – An International Comparison. eduLIFE Lifelong Learning Series, Band 3. Cheltenham, UK und Northampton, MA, USA: Edward Elgar Publishing: 93-109.

Lorenz, G., S. Lenz und C. Rjosk. 2023. Effizienz und soziale Ungleichheit in strikt leistungsdifferenzierenden Bildungssystemen. Eine kritische Betrachtung des Model of Ability Tracking (MoAbiT). In: Zeitschrift für Soziologie: https://doi.org/10.1515/zfsoz-2023-2028.

Lucas, S.R. 1999. Tracking Inequality: Stratification and Mobility in American High Schools. New York: Teachers College Press.

Maaz, K. 2017. Dreigliedrigkeit ade. Die Entwicklung des Zwei-Säulen-Modells als zukunftsweisende Struktur des deutschen Sekundarschulsystems. In: Schulverwaltung: Fachzeitschrift für Schulentwicklung und Schulmanagement. Hessen. Rheinland-Pfalz, 22: 199-201.

Maaz, M., M. Hasselhorn, T. S. Idel, E. Klieme, B. Lütje-Klose, P. Stanat, M. Neumann, A. Bachsleitner, J. Lühe und S. Schipolowski. 2019. Zentrale Befunde und Empfehlungen. In: K. Maaz, K., M. Hasselhorn, T. S. Idel, E. Klieme, B. Lütje-Klose, P. Stanat, M. Neumann, A. Bachsleitner, J. Lühe und St. Schipolowski, Hrsg., Zweigliedrigkeit und Inklusion im

empirischen Fokus. Ergebnisse der Evaluation der Bremer Schulreform. Münster und New York: Waxmann: 217-228.

Maaz, K., und M. Lörz. 2024. Nachhaltiges und gerechtes Bildungssystem. Herausforderungen und Ansatzpunkte. In: Bildung und Wissenschaft, 12: 15-19.

Malamud, O., und C. Pop-Eleches. 2011. School tracking and access to higher education among disadvantaged groups. In: Journal of Public Economics, 95: 1538–1549.

Marx, A., und K. Maaz. 2023. Wie lassen sich Bildungsungleichheiten effektiv verringern? Ein Forschungsüberblick zur Schulentwicklung in herausfordernden Lagen. In: Die Deutsche Schule, 115: 189-200.

Matthewes, S. H., 2021. Better Together? Heterogeneous Effects of Tracking on Student Achievement. The Economic Journal, 131: 1269–1307.

Meghir, C., und M. Palme. 2005. Educational reform, ability, and family background. American Economic Review, 95, 414–424.

Münch, R. 2024. Bildungsungleichheit wie in kaum einem anderen Land? Hartmut Essers Frontalangriff auf die Standardposition und das Integrationsmodell in der Bildungsforschung. Soziologische Revue, 47: 285-295.

NEPS-Netzwerk. (2021). Nationales Bildungspanel, Scientific Use File der Startkohorte Klasse 5. Leibniz-Institut für Bildungsverläufe (LIfBi), Bamberg. https://doi.org/10.5157/NEPS:SC3:12.0.0.

Neumann, M., Becker, J. Baumert, K. Maaz, O. Köller und M. Jansen. 2017. Das zweigliedrige Berliner Sekundarschulsystem auf dem Prüfstand: Ein Zwischenresumee. In: J. Baumert, K. Maaz, O. Köller und M. Jansen, Hrsg., Zweigliedrigkeit im deutschen Schulsystem. Potentiale und Herausforderungen in Berlin. Münster und New York: Waxmann: 469-501.

Oakes, J. 1985. How Schools Structure Inequality. New Haven und London: Yale University Press.

Osikominu, A., G. Pfeifer und K. Strohmaier. 2021. The Effects of Free Secondary School Track Choice: A Disaggregated Synthetic Control Approach. CESifo Working Paper No. 8879. http://dx.doi.org/10.2139/ssrn.3784390..

OECD OECD. PISA database 2000, 2003, 2006, 2009, 2012, 2015, 2018, 2022; http://www.oecd.org/pisa/ data/.

Piopiunik, M. 2014. The effects of early tracking on student performance: Evidence from a school reform in Bavaria. Economics of Education Review 42, 12–33

Roller, M., und D. Steinberg. .2020. The distributional effects of early school stratification - non-parametric evidence from Germany. In: European Economic Review, 125. https://doi.org/10.1016/j.euroecorev.2020.103422.

Schindler, St., E. Bar-Haim, C. Barone, J. F. Birkelund, V. Boliver, Q. Capsada-Munsech, J. Erola, M. Facchini, Y. Feniger, L. Heiskala, E. Herbaut, M. Ichou, K. B. Karlson, C. Kleinert, D. Reimer, C. Traini, M. Triventi, und L.-A.Vallet, 2024. Educational tracking and social inequalities in long-term labor market outcomes: Six countries in comparison. In: International Journal of Comparative Sociology, 65: 39–62.https://doi.org/10.1177/00207152231151390.

Skopek, J., M. Triventi und S. Buchholz. (2019). How do educational systems affect social inequality of educational opportunities? The role of tracking in comparative perspective. In R. Becker, Hrsg., Research Handbook on the Sociology of Education. Cheltenham: Edward Elgar Publishing Limited: 214–232.

Sörensen, A. B. 1970. Organizational Differentiation of Students and Educational Opportunity. In: Sociology of Education, 43, 355-376.

Sörensen, A. B., und M. T. Hallinan. 1977. A Reconceptualization of School Effects. In: Sociology of Education, 50, 273-289.

Strello, R., Strietholt, I. Steinmann und Ch. Siepmann. 2021. Early tracking and different types of inequalities in achievement: difference-in-differences evidence from 20 years of large-scale assessments. In: Educational Assessment, Evaluation and Accountability. https://doi.org/10.1007/s11092-020-09346-4.

Terrin, E., und M. Triventi. 2022. The Effect of School Tracking on Student Achievement and Inequality: A Meta-Analysis. In: Review of Educational Research, 20: 1–39.

van Ackeren, I., und S. M. Kühn. 2017. Homogenität und Heterogenität im Schulsystem. In: Bohl, Th., J. Budde, M. Rieger-Ladich, Hrsg., Umgang mit Heterogenität in Schule und Unterricht. Grundlagentheoretische Beiträge, empirische Befunde und didaktische Reflexionen. Bad Heilbrunn: Verlag Julius Klinkhardt: 175-190.

van de Werfhorst, H. G. 2018. Early tracking and socioeconomic inequality in academic achievement: Studying reforms in nine countries. In: Research in Social Stratification and Mobility, 58: 22–32.

van de Werfhorst, H. G. 2022. Sorting or mixing? Multi-track and single-track schools and social inequalities in a differentiated educational system. British Educational Research Journal. https://doi.org/10.1002/berj.3722

van de Werfhorst, H.G., und J.J.B. Mijs. 2010. Achievement inequality and the institutional structure of educational systems: A comparative perspective. In: Annual Review of Sociology, 36: 407–428.

von Below, S. 2011. Bildungssysteme im historischen und internationalen Vergleich. In: R. Becker, Hrsg., Lehrbuch der Bildungssoziologie, 2. Aufl. Wiesbaden: Springer VS: 139–162.

Wacker, A. 2017. Schulstruktur und Zweigliedrigkeit: Umbau des Bildungssystems. In: Th. Bohl, J. Budde, und M. Rieger-Ladich, M., Hrsg., Umgang mit Heterogenität in Schule und Unterricht. Grundlagentheoretische Beiträge, empirische Befunde und didaktische Reflexionen. Bad Heilbrunn: Klinkhardt, 191-206.

Waldinger, F., 2007. Does tracking affect the importance of family background on students' test scores? Unpublished manuscript. London: LSE.

Wößmann, L., E. Lüdemann, G. Schütz, und M.R. West. 2009. School Accountability, Autonomy and Choice around the World. Celtenham: Edward Elgar Publishing.

Wößmann, L. 2010. Institutional Determinants of School Efficiency and Equity: German States as Microcosm for OECD Countries. In: Jahrbücher für Nationalökonomie und Statistik, 230: 234–270.

Wößmann, L. 2016. The Importance of School Systems: Evidence from International Differences in Student Achievement. In: Journal of Economic Perspectives, 30: 3–32.

Wößmann, L. 2023. Erkenntnisse aus aktuellen Schulleistungsstudien zur Evaluation des Bildungssystems. Eine bildungsökonomische Perspektive. In: N. McElvany, M. Becker, H. Gaspard, F. Lauermann und A. Ohle-Peters, Hrsg,, Evaluation des Bildungssystems. Welche Erkenntnisse liefern die Schulleistungsstudien? Münster und New York: Waxmann: 9-32.

Zimmer, R. W., und E. F. Toma. 2000. Peer effects and educational vouchers: evidence across countries. In: Journal of Policy Analysis and Management, 19: 75–79.