Consequences of Institutional Differentiation in Secondary School Systems

Three Empirical Essays in the Economics of Education

Inaugural Dissertation

zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaft
doctor rerum politicarum
(Dr. rer. pol.)

am Fachbereich Wirtschaftswissenschaft
der Freien Universität Berlin

vorgelegt von

Sönke Hendrik Matthewes (M.Sc.)

Berlin, 2022

Erstgutachterin:	Prof. Dr. C. Katharina Spieß *Freie Universität Berlin und BiB*
Zweitgutachter:	Prof. Dr. Jan Marcus *Freie Universität Berlin und DIW Berlin*

Datum der Disputation: 05.07.2022

Acknowledgements

I would like to thank all people who have helped, supported and guided me in writing this dissertation.

First and foremost, I would like to thank my two supervisors C. Katharina Spieß and Jan Marcus. Thank you Katharina for accepting me as an external PhD candidate, providing strategic and academic advice when needed, commenting on my drafts and inviting me to join your monthly doctoral colloquium at the DIW Berlin. Its friendly and welcoming atmosphere there made me feel increasingly part of the unit. The constructive discussions in the colloquium greatly improved my research. Thank you Jan for your invaluable academic guidance and continued support throughout the time of my PhD, despite only being my second supervisor on paper. Without your detailed and constructive feedback, especially during the first years, I would have not been able write much of this dissertation—and certainly not the first paper.

Second, I would like to thank Heike Solga for her guidance and support throughout the years. She and my former colleagues at the WZB Berlin provided a very enjoyable and stimulating working environment. I greatly benefited from the interdisciplinary exchange. A special thank you goes to my former colleagues of the CIDER team.

Third, I would like to thank my co-author Camilla Borgna for the fruitful and fun exchanges. An extra big thank you goes to my second co-author Guglielmo Ventura for the great teamwork, stimulating discussions and friendship throughout the time of my PhD.

Fourth, I would like to thank Sandra McNally for hosting me as a visiting PhD student at the Centre for Economic Performance in London and giving me the opportunity to contribute to CVER's research programme.

Finally, I would like to thank all my friends for their continued support and the healthy distraction and my parents, Michèle and Stefan, and Charlotte for their kind and loving support during the final sprint of this dissertation.

Berlin, March 18, 2022
Sönke Hendrik Matthewes

Contents

List of Tables

List of Figures

Rechtliche Erklärungen

Erklärung gem. §4 Abs. 2 (Promotionsordnung)
Hiermit erkläre ich, dass ich mich noch keinem Promotionsverfahren unterzogen oder um Zulassung zu einem solchen beworben habe, und die Dissertation in der gleichen oder einer anderen Fassung bzw. Überarbeitung einer anderen Fakultät, einem Prüfungsausschuss oder einem Fachvertreter an einer anderen Hochschule nicht bereits zur Überprüfung vorgelegen hat.

Berlin, 18. März 2022
Sönke Hendrik Matthewes

Erklärung gem. §10 Abs. 3 (Promotionsordnung)
Hiermit erkläre ich, dass ich für die Dissertation folgende Hilfsmittel und Hilfen verwendet habe:

- Software:
  - Stata 15, 16 und 17
  - Microsoft Office 2013 und 2019
  - Sublime Text 3 und 4
  - LaTeX
  - R 3.6.1
- Literatur: siehe Literaturverzeichnis

Auf dieser Grundlage habe ich die Arbeit (in Zusammenarbeit mit meinen Ko-AutorInnen) selbstständig verfasst.

Berlin, 18. März 2022
Sönke Hendrik Matthewes

Ko-Autorenschaften und Vorveröffentlichungen

Chapter 2: Better Together? Heterogeneous Effects of Tracking on Student Achievement

- Ko-AutorInnen: keine.
- Vorveröffentlichungen:
  - Eine leicht überarbeitete Version dieses Kapitels ist erschienen als:
    Matthewes, S.H. (2021). 'Better Together? Heterogeneous Effects of Tracking on Student Achievement', *The Economic Journal*, Vol. 131, Issue 635, pp. 1269–1307. https://doi.org/10.1093/ej/ueaa106.
  - Außerdem sind zwei frühere Versionen dieses Kapitels als Arbeitspapiere erschienen:
    Matthewes, S.H. (2020). 'Better Together? Heterogeneous Effects of Tracking on Student Achievement', *CEP Discussion Paper No. 1706*, Centre for Economic Performance, London School of Economics.
    Matthewes, S.H. (2018). 'Better Together? Heterogeneous Effects of Tracking on Student Achievement', *DIW Discussion Paper No. 1775*, German Institute for Economic Research, DIW Berlin.
  - Teilweise basierend auf diesem Kapitel ist folgende Transferpublikation erschienen:
    Matthewes, S.H. (2020). 'Längeres gemeinsames Lernen macht einen Unterschied', *WZBrief Bildung 40*, WZB Berlin Social Science Center.

Chapter 3: De-Tracking at the Margin: Local School Supply and Educational Expansion in Germany

- Ko-Autorin: Camilla Borgna (Collegio Carlo Alberto, Turin)
- Vorveröffentlichungen: keine.

Chapter 4: Labour Market Returns to Vocational Education in the Presence of Multiple Alternatives

- Ko-Autor: Gugliemo Ventura (Centre for Economic Performance, London School of Economics & University College London)

- Vorveröffentlichungen: keine.

Abstract

This dissertation comprises three self-contained research papers in the empirical economics of education, each investigating a different aspect of institutional differentiation in secondary school systems. Across chapters, the focus shifts from *vertical* differentiation between schools due to ability tracking to *horizontal* differentiation between vocational and academic curricula. The chapters respectively study the effects of between-school ability tracking on student achievement (chapter 2), the effects of local school supply on educational attainment (chapter 3) and labour market returns to vocational education (chapter 4). They are preceded by a general introduction of the topic (chapter 1) and succeeded by a general conclusion discussing policy implications (chapter 5).

Chapter 2 studies the effects of early between-school ability tracking on student achievement in Germany. A long-standing and controversial debate on tracked *vs.* comprehensive schooling revolves around a perceived efficiency-equity trade-off: while homogeneous learning environments should make teaching more efficient through better targeting of content and teaching style to median classroom ability, students sorted into lower tracks might systematically lose out through unfavourable motivational and peer effects. However, given the severity of the selection problems involved, compelling empirical evidence on the matter is scant.

My research design exploits institutional differences between German federal states: in all states about 40% of students transition to separate academic-track schools after comprehensive primary school, at about age ten. Depending on the state, the remaining non-academic-track students are either further tracked between separate low- and intermediate-track schools or taught comprehensively for at least another two years. I compare the achievement growth of non-academic-track students during their first two years of secondary school between tracked and comprehensive states, while controlling for state-specific achievement trends using academic-track students in a triple-differences framework. I find evidence for positive average effects of prolonged comprehensive schooling on mathematics and reading achievement. The average effects are almost entirely driven by large effects for low achievers, while effects for high achievers are null. This shows that too early and rigid forms of tracking can impair both the equity and the efficiency of school systems.

Chapter 3 studies the effects of local school supply on upper-secondary attainment in Germany's tracked school system. We consider not only the *number*, but also the *type* of available schools by estimating school-type-specific supply effects on attainment of the university-entrance certificate for traditional academic-track schools, as well as for newer comprehensive schools and vocational high schools. The latter two offer alternative pathways towards university-eligibility,

which traditionally could only be obtained on academic-track schools, and thus constitute a partial de-tracking of upper-secondary schooling in Germany.

Drawing on yearly administrative records that cover the universe of German students, schools, and graduates, we compile a county-level panel of local school supply and upper-secondary attainment for 13 cohorts between 1995 and 2007. We document that, while attainment has substantially expanded, so has regional dispersion, pointing to growing inequality of educational opportunities. Cross-sectionally, we find that the supplies of all three school types correlate positively with attainment, but for comprehensive and academic-track schools this association is largely spurious, i.e., due to regional differences in educational demand. For vocational high schools, in contrast, we find robust evidence for a positive supply-side effect on attainment, confirmed in two-way fixed-effects, difference-in-differences, and event-study models. The hybrid nature of vocational high schools, combining academic and specialised curricula, might attract students who otherwise would be diverted from academic upper-secondary education towards vocational training.

In many countries, decreasing job prospects for non-tertiary-educated students, together with skill shortages in technical occupations, have led to a heightened interest in improving and expanding vocational programmes in secondary education. However, critics fear that doing so might divert students from academic routes whose focus on general, instead of occupation-specific, skills might be better suited for today's rapidly changing labour markets. ***Chapter 4*** aims to contribute to this debate by delivering causal evidence on labour market returns to vocational upper secondary education in England, where at the age of 16, after completing compulsory schooling, students choose between a vocational track, an academic track and no upper secondary education.

Our research design leverages multiple instrumental variables to estimate margin-specific treatment effects, i.e., causal returns to vocational education for students at the margin between vocational and academic education and, separately, for students at the margin between vocational and no post-16 education. Using the fact that vocational and academic education are offered by distinct institutions, identification comes from plausibly exogenous variation in distance to the nearest vocational provider conditional on distance to the nearest academic provider (and *vice-versa*), while controlling for granular student-, school- and neighbourhood-level characteristics. We draw on linked administrative education and earnings data, through which we can follow the full student population into the labour market until age 29. We find that the vast majority of marginal vocational students are at the margin with academic education (instead of no further education), so that the first-order effect of expanding vocational-track access is diversion from the academic track. Diversion leads to losses in earnings at age 29 of about 9% for males and 7% for females. These effects are not driven by employment but due to wages (or working hours). A substantial part of the effect is explained by reduced university degree completion. For the few marginal students at the margin with no further education, we find tentative evidence of positive employment and earnings effects but results are imprecise and insignificant. Our findings caution against an expansion of vocational upper secondary education in England in its current form.

Zusammenfassung

Die vorliegende Dissertation umfasst drei in sich abgeschlossene Forschungsarbeiten im Bereich der empirischen Bildungsökonomie, die jeweils einen anderen Aspekt der institutionellen Differenzierung in Sekundarschulsystemen untersuchen.

Über die einzelnen Kapitel hinweg verlagert sich der Schwerpunkt von der *vertikalen* Leistungsdifferenzierung zwischen Schulformen auf die *horizontale* Differenzierung zwischen beruflichen und akademischen Lehrplänen. In drei Kapiteln werden die Auswirkungen einer frühen Leistungsdifferenzierung zwischen Schulformen (*Tracking*) auf die Kompetenzen der Schülerschaft (***Kapitel 2***), die Auswirkungen des lokalen Schulangebots auf Bildungsabschlüsse (***Kapitel 3***) und zuletzt die Arbeitsmarkterträge aus beruflicher Bildung (***Kapitel 4***) untersucht. Den Kapiteln geht eine allgemeine Einführung voraus (***Kapitel 1***) und ihnen folgt eine allgemeine Schlussfolgerung, in der politische Implikationen erörtert werden (***Kapitel 5***).

Kapitel 2 untersucht die Auswirkungen der frühen Leistungsdifferenzierung im gegliederten Schulsystem Deutschlands auf die Kompetenzen der SchülerInnen. Eine seit langem geführte und kontroverse Debatte über die Vor- und Nachteile eines gegliederten Schulsystems dreht sich um einen vermeintlichen Kompromiss zwischen Effizienz und Chancengleichheit: Während homogene Lernumgebungen den Unterricht durch eine bessere Ausrichtung der Inhalte und des Unterrichtsstils auf die durchschnittliche Leistungsfähigkeit der Klasse effizienter machen sollten, könnten SchülerInnen, die in niedrigere Schulzweige eingeteilt werden, durch ungünstige Motivations- und Peer-Effekte dadurch systematisch benachteiligt werden.

Mein Forschungsdesign nutzt institutionelle Unterschiede zwischen den Bundesländern: In allen Bundesländern wechseln etwa 40% der SchülerInnen nach der Grundschule im Alter von etwa zehn Jahren auf ein Gymnasium. Je nach Bundesland werden die verbleibenden nicht-gymnasialen SchülerInnen entweder weiter zwischen Haupt- und Realschulen aufgeteilt oder in einer Gesamtschule oder Schule mit mehreren Bildungsgängen mindestens zwei weitere Jahre gemeinsam unterrichtet, weil in der Orientierungsstufe innerhalb dieser Schulen nicht nach Leistung differenziert wird. Um den kausalen Effekt der leistungsbezogenen Aufteilung nicht-gymnasialer SchülerInnen auf Haupt- und Realschulen zu bestimmen, wurde anhand eines *Difference-in-Differences* Ansatzes analysiert, wie sich die Leistungsentwicklung der nicht-gymnasialen Schülerschaft in den ersten beiden Jahren der Sekundarstufe zwischen zwei Gruppen von Bundesländern unterscheidet: solchen, in denen es weiterhin ab der 5. Klasse separate Haupt- und Realschulen gibt, und solchen, in denen diese Bildungsgänge in einer Schulform zusammengefasst wurden. Um bundeslandspezifische Leistungstrends herauszurechnen, wurden die Unterschiede in der Leistungsentwicklung der nicht-gymnasialen Schülerschaft anhand eines *Triple-Differences* Ansatzes zusätzlich mit jenen der gymnasialen Schülerschaft verglichen.

Die Befunde zeigen positive Durchschnittseffekte des längeren gemeinsamen Lernens auf die Mathematik- und Lesekompetenz der nicht-gymnasialen Schülerschaft. Diese Durchschnittseffekte sind fast ausschließlich auf starke positive Effekte für SchülerInnen mit geringen Leistungen zurückzuführen, während für SchülerInnen mit höheren Leistungen keine (negativen) Effekte festzustellen sind. Dies zeigt, dass eine zu frühe und starre Form der vertikalen Differenzierung sowohl die Gerechtigkeit als auch die Effizienz von Schulsystemen beeinträchtigen kann.

Kapitel 3 untersucht, wie sich das regionale Schulangebot im gegliederten Schulsystem Deutschlands auf die Wahrscheinlichkeit des Erlangens der allgemeinen Hochschulreife (Abitur) auswirkt. Die Analysen berücksichtigen nicht nur die Anzahl, sondern auch die Art der verfügbaren Schulen, indem die Angebotseffekte auf die Wahrscheinlichkeit des Abiturabschlusses getrennt für traditionelle Gymnasien, Gesamtschulen und berufliche Gymnasien geschätzt werden. Da die beiden letztgenannten Schulformen alternative Wege zur Hochschulreife anbieten, welche traditionell nur an Gymnasien erlangt werden konnte, impliziert ihre Angebotsausweitung eine teilweise Abschwächung der Differenzierung im Sekundarbereich II.

Auf der Grundlage der regionalen Schulstatistik, die die Gesamtheit der deutschen SchülerInnen, Schulen und AbsolventenInnen abdeckt, wurde ein Paneldatensatz des regionalen Schulangebots und regionalen AbiturientInnenquoten für 13 Kohorten zwischen 1995 und 2007 auf der Landkreisebene erstellt. Auf Grundlage dieser Daten zeigt sich, dass nicht nur das Niveau, sondern auch die regionale Streuung der AbiturientInnenquoten erheblich gestiegen ist, was auf eine wachsende Ungleichheit der Bildungschancen hindeutet. Querschnittlich korreliert das Angebot aller drei Schultypen positiv mit der regionalen AbiturientInnenquote, aber für Gesamtschulen und Gymnasien ist dieser Zusammenhang größtenteils auf regionale Unterschiede in der Bildungsnachfrage zurückzuführen. Für die beruflichen Gymnasien zeigt sich dahingegen ein signifikanter und robuster angebotsseitiger Effekt auf den Bildungserfolg, der sich in auch in längsschnittlichen *Two-Way Fixed-Effects*, *Difference-in-Differences* und *Event-Study* Modellen bestätigt. Der Befund, dass das Angebot beruflicher Gymnasien das Niveau der schulischen Bildungsabschlüsse erhöht, lässt sich möglicherweise auf den hybriden Charakter dieser Schulform, die akademische und berufliche Lehrpläne kombiniert, zurückführen: Dies könnte manche SchülerInnen, die andernfalls in die berufliche Bildung abgewandert wären, anziehen und sie somit zum Besuch der allgemeinbildenden Sekundarstufe II bewegen.

In vielen Ländern haben die sinkenden Berufsaussichten für Menschen ohne Hochschulabschluss und der Fachkräftemangel in technischen Berufen zu einem verstärkten Interesse an der Verbesserung und Ausweitung beruflicher Bildung im Sekundarbereich geführt. Kritiker befürchten jedoch, dass dadurch SchülerInnen von akademischen Bildungswegen abgelenkt werden könnten, obwohl ein Schwerpunkt auf allgemeinen anstatt auf berufsspezifischen Fähigkeiten, für die sich schnell verändernden Arbeitsmärkte von heute, besser geeignet wäre. ***Kapitel 4*** leistet einen Beitrag zu dieser Debatte, indem es kausale Evidenz für die Arbeitsmarktrenditen der beruflichen Bildung in England liefert, wo die SchülerInnen im Alter von 16 Jahren nach

Abschluss der Schulpflicht zwischen einer beruflichen, einer allgemeinen oder keiner weiteren Sekundarbildung wählen können.

Das Forschungsdesign nutzt mehrere Instrumentalvariablen, um *alternativenspezifische* Bildungsrenditen zu schätzen, d.h. kausale Erträge aus beruflicher Bildung für SchülerInnen an der Grenze zwischen dem beruflichen und dem allgemeinen Bildungsgang und, separat, für SchülerInnen an der Grenze zwischen beruflicher und keiner weiteren Sekundarbildung. Da die beruflichen und allgemeinen Bildungsgänge von unterschiedlichen Institutionen angeboten werden, lässt sich die Kausalidentifikation dieser Effekte mit Hilfe von plausibel exogener Variation in der Entfernung zum nächstgelegenen beruflichen Bildungsanbieter unter Kontrolle der Entfernung zum nächstgelegenen allgemeinen Bildungsanbieter (und umgekehrt) erzielen, wobei gleichzeitig diverse Merkmale auf Schüler-, Schul- und Nachbarschaftsebene kontrolliert werden. Die Analysen basieren auf verknüpften administrativen Bildungs- und Einkommensdaten, anhand derer sich die Gesamtheit der englischen SchulbsolventInnen dreier Kohorten bis zum Alter von 29 Jahren auf dem Arbeitsmarkt beobachten lässt.

Die Befunde zeigen, dass die überwiegende Mehrheit der marginalen BerufsschülerInnen an der Grenze zum allgemeinen Bildungsgang steht. Eine Ausweitung des Zugangs zu beruflicher Bildung würde also in erster Linie SchülerInnen anziehen, die andernfalls allgemeine Bildung gewählt hätten, und kaum SchülerInnen, die keine Ausbildung wählen. Für SchülerInnen an der Grenze zwischen beruflicher und allgemeiner Bildung gehen mit der Wahl beruflicher Bildung jedoch substanzielle Einkommensverluste einher: Ihr jährliches Einkommen im Alter von 29 Jahren verringert sich durch diese Wahl um etwa 9% für Männer und um 7% für Frauen. Diese Auswirkungen sind nicht auf die Beschäftigung zurückzuführen, sondern auf die Löhne (oder die Arbeitszeiten). Ein erheblicher Teil der Einkommenseinbußen durch berufliche Bildung ist auf die damit einhergehende geringere Wahrscheinlichkeit des Hochschulabschlusses zurückzuführen. Für die wenigen marginalen BerufsschülerInnen an der Grenze zu keiner weiteren Ausbildung zeigen sich Hinweise auf positive Beschäftigungs- und Einkommenseffekte durch berufliche Bildung, die aber ungenau geschätzt und nicht statistisch signifikant sind. Insgesamt sprechen die Befunde klar gegen eine Ausweitung beruflicher Bildung im Sekundarbereich in England—zumindest in ihrer derzeitigen Form.

Chapter 1

Introduction

1.1 Motivation

Education is crucially important for success in life. Across countries, more highly educated individuals are better-off economically, as evidenced by higher average life-time earnings and lower risks of unemployment (Checchi, 2006). A long-standing literature in labour economics has established that these empirical associations are indeed largely to be interpreted causally (e.g., Mincer, 1974; Angrist and Krueger, 1991; Card, 1999, 2001; Psacharopoulos and Patrinos, 2018). Above and beyond its labour market returns, education has been shown to improve other life outcomes, such as health (e.g., Grossman, 2006) and life satisfaction (e.g., Oreopoulos and Salvanes, 2011).

Importantly, education improves not only the lives of those who receive it, but also of those around them. First-order positive externalities from education are of fiscal nature, concerning higher tax bases from increased labour income and lower benefit spending from reduced unemployment levels. But by generating local productivity spillovers (e.g., Moretti, 2004), by curbing crime (e.g., Lochner and Moretti, 2004) and by fostering civic engagement (e.g., Dee, 2004), education promotes wealthier and saver societies with better functioning democratic institutions, more generally. Note that these benefits extend to future generations as children of more highly educated individuals attain higher levels of education themselves (Black *et al.*, 2005). From a macroeconomic perspective, education has been shown to foster innovation (e.g., van Reenen, 2022) and economic growth (e.g., Hanushek and Woessmann, 2015), thus contributing to the long-term prosperity of nations.

The primary mechanism economists believe to govern the relationship between education and all of these outcomes is *human capital*. The idea of workers' capabilities as inputs to economic production, over and above the mere quantity of labour and physical capital, can be traced back to the writings of Adam Smith (1776) and the very birth of the economics discipline. Yet, the term was most famously coined, and the concept formalised, by Gary Becker (1964). He defined human capital as stock of an individual's knowledge and (cognitive) skills that determine the productivity of their labour. Since, his definition has been broadened to include non-cognitive skills (e.g., Cunha *et al.*, 2006; Cunha and Heckman, 2008) and even health (e.g., Conti and Heckman, 2014; Goldin, 2016), which have become to be recognised as similarly important for productivity.

The central idea in Becker (1964) is that individuals can accumulate human capital through (costly) investments into education and training. During their education students acquire skills and knowledge that correspond directly to specific tasks in the production process (*specific* human capital) and/or it is the abstract process of studying itself, which fosters general cognitive and non-cognitive skills that increase individual's productivity across tasks (*general* human capital). Consequently, more highly educated individuals face better prospects on labour markets. Clearly, general human capital also corresponds to aptitude in other realms of life, thereby causing the range of beneficial outcomes listed above. To name but one example, the health benefits from education are usually attributed to educated individual's heightened understanding of, and responsiveness to, critical health-related information (Huebener, 2020).

Conceptualising education as an investment into human capital highlights that the educational decision can be fraught with market failure. Investments into education require time, effort and money from children and their families. Given that this investment confers large *societal* benefits beyond those that accrue to the individual (see above), the state has a mandate to subsidise education to prevent underinvestment in human capital by students who do not take into account these externalities.

Also from the perspective of students, underinvestment in human capital is likely if it is left to the market. Research spearheaded by James Heckman has shown that skill formation is a cumulative process, marked by the self-productivity of skills (i.e., skills beget further skills) and the dynamic complementarity of investments (i.e., early investments increase returns to later investments) (see, e.g., Cunha and Heckman, 2008). This implies that human capital investments need to start early in life and be resolutely maintained throughout students' youth to yield good outcomes. Accordingly, children are dependent on their parents to make these decisions (and bear their costs) for them because they can hardly make up for missed early investments later in life themselves. Yet, neither are all parents equally altruistic therein, nor are all of them well informed about the returns to education and the importance of early human capital investments. Moreover, parents typically cannot borrow against their children's future income, so many families will face binding credit constraints with respect to educational investments (Lochner and Monge-Naranjo, 2011). As they become older, children may become more independent in these matters but especially adolescents might be neither informed nor forward-looking enough to make these decisions optimally.

Together these arguments make a strong case for publicly funded investments into human capital from an efficiency standpoint. This case is further strengthened by equity considerations, as problems relating to information asymmetries, credit constraints and low parental preferences for education are particularly salient among families with lower socio-economic status (SES). Therefore, a fully unregulated education market would result in a highly unequal distribution of human capital and, given its importance for economic success, grave inequality of opportunity. For these reasons almost all societies publicly provide primary and secondary schooling, and increasingly also early childhood education and care (ECEC), and enforce attendance through

compulsory schooling laws in an effort to ensure continued investments into human capital for all its children.

Recent economic trends further increase the relevance of education for individual and societal welfare, especially in post-industrial societies, such as Germany and the UK. A burgeoning literature in labour economics documents that ongoing processes of technological change and automation exhibit a skill bias, increasing labour demand for highly educated workers at the expense of those with lower levels of education (e.g., Spitz-Oener, 2006; Acemoglu and Autor, 2011; Michaels *et al.*, 2014; Acemoglu and Restrepo, 2018). Indeed, demand for highly-educated workers seems to have outpaced their supply, resulting in steadily increasing returns to education despite substantial educational expansion (Goldin and Katz, 2010; Becker and Blossfeld, 2022). In stark contrast to previous decades, since the 1980s most Western economies saw wages fall for less-educated workers and overall wage inequality increase (see Autor, 2019, for the US; Dustmann *et al.*, 2009, for Germany; Blundell *et al.*, 2018, for the UK). As economies continue to shift from manufacturing to knowledge-based services, education and training also becomes increasingly important for the international competitiveness of countries as a whole (Dustmann *et al.*, 2008). These developments mean that it is increasingly important for students to attain high levels of education to secure a good standard of living, while governments face mounting pressures to improve the quality and inclusiveness of their education systems to secure future growth and combat inequality.

The problem is that it is far from obvious how the state should invest in education to further the 'production' of human capital. Education levels in most Western societies are quite high already, so that the potential of a sole focus on the *quantity* of education individuals receive has its limits. Clearly, increasing ECEC attendance and tertiary enrolment rates is key, but it appears no less important to improve the *quality* of the educational programmes individuals already attend. Given near universal attendance due to compulsory schooling laws, a focus on increasing the productivity of schools in transferring human capital is particularly promising in this context. A large strand in the economics of education is concerned with the 'educational production' in schools: in analogy to classic economic theory on firms, the process of learning is conceptualised as a production process which takes inputs, such as student characteristics, teachers, school resources, but also institutional rules, in order to produce outputs, such as (domain-specific) cognitive skills, non-cognitive skills, degrees and wages. Through empirical modelling of 'education production functions', this research aims to identify the most productive manipulable input factors to guide optimal resource allocation.[1]

Yet, the early literature in this field, starting with the famous Coleman Report (Coleman *et al.*, 1966), identified what has been dubbed the 'resource puzzle': most of the variation in student achievement can be accounted for by family background and only little by factors relating to schools and, despite steady increases in schooling inputs, achievement has not increased commensurately (Hanushek, 1986, 2003). This fuelled widespread scepticism regarding the

[1] Where optimality might not only be defined by *maximisation* of educational output but also by features of its *distribution*.

efficacy of (resource-based) schooling policy and, more broadly, gave rise to the notion that policymakers can do little to influence the distribution of human capital in society.

However, this pessimistic conclusion has since been challenged. On the one hand, a more recent strand of the literature, which—following the much discussed 'credibility revolution' in empirical economics (see Angrist and Pischke, 2010)—employs careful research designs to discern *causal* effects of schooling inputs, generally does confirm their importance for student achievement (e.g., Angrist and Lavy, 1999; Krueger, 1999; Jackson *et al.*, 2016). On the other hand, an internationally comparative literature, made possible by advent of international large-scale educational assessment studies, such as TIMSS, PIRLS and PISA[2], revealed that the effects of schooling inputs on achievement depend crucially on a country's educational *institutions*, such as school autonomy (e.g., regarding budgets or curricula), school accountability (e.g., through central exit exams) and institutional differentiation between schools (e.g., through tracking) (e.g., Bishop, 1997; Lee and Barro, 2001; Woessmann, 2003; Hanushek and Woessmann, 2011).

The institutional framework of the education system is important because it determines how resources are allocated in the first place and what kind of incentives actors face. As such, institutions can themselves be conceptualised as higher-level inputs to educational production. Using PISA data, Woessmann (2016) shows that competences of 15-year-olds vary greatly between countries—both in levels and in distribution—and that institutions of the school system can account for about half of the between-country variation in average test scores—more than student-level characteristics, which include detailed measures of family background. Also associations between student achievement and SES vary greatly across school systems (e.g., Brunello and Checchi, 2007; Waldinger, 2007; Schütz *et al.*, 2008; Jerrim and Micklewright, 2014). This suggests that, through shaping institutions, the potential for school policy to impact on the efficiency and equity of educational production is large. School systems being largely state-controlled and all children passing through school, they also represent a readily available lever for policymakers.

The problem is that, while the cross-country evidence can detect their overall importance, disentangling causal effects of *specific* institutional features is notoriously difficult. Institutions change seldom and, if at all, slowly and gradually. Accordingly, there often is little to no within-country variation that could be used to investigate the effects of specific institutions. Comparisons of students who attend different institutions within a given education system are plagued by selection bias, because students self-select into institutions based on idiosyncratic characteristics which also independently affect their outcomes, but typically cannot be observed by the researcher. Identification from school reforms, which are rare, context-specific and often involve multiple changes at once, or cross-country comparisons, which offer only a small number of observations and likely to suffer from severe omitted variable bias due to unobserved country factors, have their own problems. For these reasons, debates on *how* school systems are to be

[2]The acronyms refer to the three most important international studies of student achievement: Trends In International Mathematics And Science Study (TIMSS), which commenced in 1995; Progress In International Reading Literacy Study (PIRLS), which commenced in 2001; Programme for International Student Assessment (PISA), which commenced in 2000.

reformed to live up to the challenges of the twenty-first century continue to be marked by severe controversy.

Here, this dissertation aims to contribute by employing careful research designs, in the spirit of the above-mentioned credibility revolution, to deliver plausibly causal empirical evidence on a particularly controversial feature of education policy, namely the level and nature of *institutional differentiation* in secondary school systems and its impact on the efficiency and equity of educational production. The empirical analyses focus on Germany and England, exemplifying two countries for which the above-sketched arguments regarding the importance of education are of particular relevance, but which take radically different approaches to institutional differentiation in their secondary school systems.

1.2 Institutional Differentiation in Secondary School Systems

In education, the term 'differentiation' refers to grouping students by some characteristic in order to better target instruction. The aim is to create better matches between students and the education they receive. Differentiation may be institutionalised in different ways and to different degrees. In fact, there is dramatic international variation with respect to institutional differentiation in secondary school systems. While all countries group students *by age* into different grade levels, beyond this point there is little consensus if and how students ought to be grouped. To structure this discussion, it is helpful to distinguish between *vertical* and *horizontal* differentiation.

Vertical differentiation, also known as *ability tracking*, refers to grouping students by their scholastic ability into hierarchically ordered school tracks. The rationale is that schools or classrooms which are homogeneous in terms of ability allow educators to tailor instruction speed, content difficulty and pedagogical approaches for the given set of students (Betts, 2011). Thus, ability tracking is supposed to increase the efficiency of instruction: by teaching every student according to her level, all students are ought to learn more. Vertical differentiation comes in degrees: from least to most rigid, students may be sorted into ability-ranked classes *per subject*, they may be sorted into permanent ability-ranked streams *within schools* or they may be sorted into different ability-ranked *school types*. Next to the rigidity of tracking, the second parameter that describes the amount of vertical differentiation in a school system is the age, i.e., grade level, at which students begin to be tracked, if at all.

Horizontal differentiation, or *curricular tracking*, refers to grouping students by a (self-chosen) subject of specialisation. Conceptually, this is not a distinction between *levels* but between *types* of education. The rationale is that students can choose a curriculum and mode of instruction which best matches their interests and talents and to, thereby, prepare a workforce with a diversity of skills. A weak form of curricular tracking arises when students can choose different high school majors within an academic curriculum. Stronger forms involve a differentiation between vocational and academic education, where the former focuses on occupation-specific knowledge and practical skills (i.e., specific human capital), while the latter

emphasises general knowledge and analytical skills (i.e., general human capital). Implementation varies from students being able to pick single vocational courses as part of their regular secondary curriculum, over sorting between distinct academic and vocational schools to vocational training in the form of firm-based apprenticeships. Given that states generally define a core curriculum every student has to follow first, curricular differentiation usually commences only in upper-secondary education.

While the primary argument for vertical differentiation concerns improved efficiency in the production of human capital, the primary argument against it concerns impaired equity therein. Critics of ability tracking fear that it only benefits high-track students, whereas those assigned to lower tracks lose out compared to a scenario with comprehensive schooling. The reasons for this fear are threefold: first, to the extent that high-performing peers are beneficial to learning (or low-performing ones harmful), tracking mechanically exerts an unequal influence as it deprives lower track students of more able peers (Sacerdote, 2011). Second, there might be motivational consequences of separating students by ability: lower track students, knowing they are deemed to be of lower aptitude, might feel discouraged and reduce their learning efforts. Endogenous teacher expectations can have similar effects (Papageorge *et al.*, 2020). Third, while an argument for tracking is resource differentiation, this has been shown to often play out to the disadvantage of low tracks (Brunello and Checchi, 2007). The gravity of these concerns increases with the degree of vertical differentiation, i.e., the earlier tracking starts, because divergences can accumulate, and in between-school tracking systems as compared to within-school ones, because differences between tracks are starker (Betts, 2011).

While conceptually distinct, horizontal and vertical differentiation are linked in practice, because vocational tracks usually cater to lower-achieving students. Accordingly, many of the arguments from above play a role in the debate surrounding the differentiation between vocational and academic tracks, as well. Yet, given that this affects older upper-secondary students and vocational education explicitly aims to prepare for work, arguments relating to the *type* (rather than the *level*) of skills students acquire and their value in the labour market are more relevant here (Bertrand *et al.*, 2021). A central argument for vocational education is that, by creating links to specific occupations, it offers stable pathways into employment, especially for less academically inclined students who would otherwise struggle with, or drop out from, academic education (Mane, 1999; Hall, 2016). However, the prospect of stable early employment might also divert students from academic to vocational education who would have benefited from an academic track that opens up pathways to higher education and, subsequently, higher-paying graduate jobs. This argument might be particularly cogent for students from lower socio-economic backgrounds, who tend to be more risk averse (Birkelund and van de Werfhorst, 2022). The general skills taught in academic tracks might also face lower risks of becoming obsolete in the face of technological change, thereby offering better labour market prospects in the long run (Hanushek *et al.*, 2017).

The school systems of Germany and England under study in this dissertation inhabit radically different positions in the sketched two-dimensional space of institutional differentiation. Figure

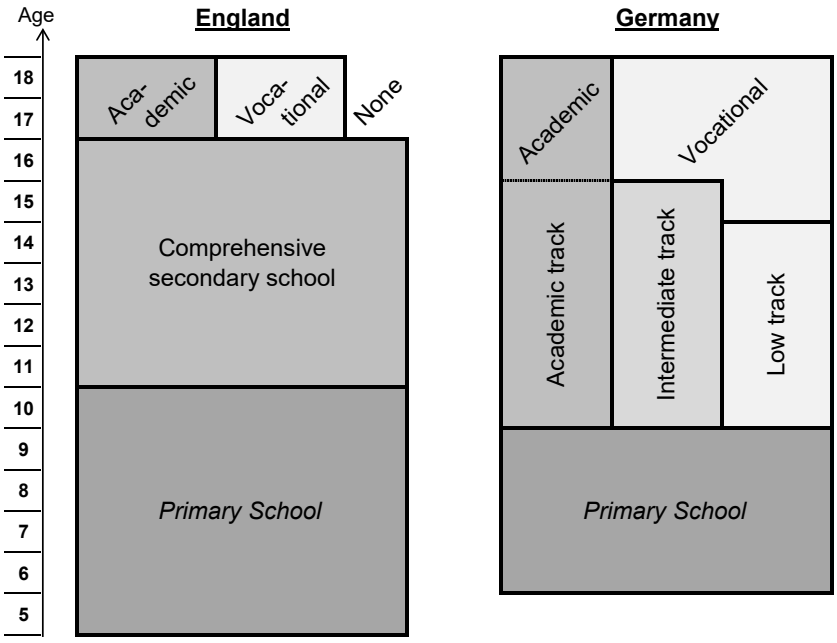


Figure 1.1. Stylized depiction of school systems in England and Germany.

Notes: The schematic depiction abstracts from many details. Notably, in England academic upper secondary education is offered both by schools which also cover the lower-secondary phase and by non-school providers, such as Sixth Form Colleges and Further Education Colleges, which also offer vocational education (see chapter 4). For Germany, the traditional three-tiered system is depicted, yet many states have a two-tiered structure (see chapter 2). Further, it abstracts from the existence of comprehensive schools and vocational high schools as alternative providers of academic upper secondary education (see chapter 3). Further, it abstracts from between-state differences in the length of primary school. Finally, for both countries the figure abstracts from heterogeneity in vocational education provision. *Source:* Own illustration.

1.1 gives a stylized depiction of their institutional structures. Germany has one of the most strongly differentiating secondary school systems worldwide, both along the vertical and the horizontal dimension. The lower-secondary phase is marked by an early and rigid ability tracking of students between school types, and the upper-secondary phase is marked by a strict form of curricular tracking into school-based academic and partially firm-based vocational education. England, in contrast, has a comprehensive secondary school system through the end compulsory schooling (at age 16) and only then differentiates between academic and largely classroom-based vocational education.

Interestingly, until the 1960s England used to have an early between-school tracking system, much like the German one, but the above-sketched arguments concerning the inegalitarian consequences of tracking led a major school reform which established a comprehensive lower-secondary school system (see Kerckhoff *et al.*, 1996). Nowadays, education policy debate in England largely concerns the post-compulsory schooling segment, especially the quality of its vocational education and training (VET) (see, e.g., Wolf, 2011; Independent Panel on Technical Education, 2016; Machin *et al.*, 2020) and the comparatively high share school-leavers who pursue neither vocational nor academic upper-secondary education after age 16 (see, e.g., Hupkau *et al.*, 2017; Boshoff *et al.*, 2019). Notably, Germany's 'dual' VET system, which combines firm-based apprenticeships with school-based instruction, is often cited as a role model in this context. Chapter 4 contributes to the debate on the merits of vocational education in England

by studying its labour market returns *vis-à-vis* either alternative, i.e., academic education and no post-16 education.

Despite having received similarly harsh criticisms over the years, Germany's tracking system has proved remarkably stable (Edelstein and Veith, 2017). Interestingly, this has been attributed to the strength of its VET system: the distinct pathways of the between-school tracking system ensure that a majority of students is channelled into upper-secondary VET eventually, thereby constituting a hard-to-change equilibrium (Thelen, 2004; Busemeyer and Trampusch, 2011). However, while the tracking system's most fundamental rationale of sorting students by ability into academic-track (*Gymnasium*) and non-academic-track schools[3] at an early age remains, the traditional 'tripartite' structure depicted in Figure 1.1 underwent many subtle changes over the years. One important change that several states have implemented, is a switch from the three- to a two-tiered structure through conflation of the lower two tracks. Chapter 2 exploits these between-state differences to discern the effect of early tracking on student achievement. Other changes concern the introduction of alternative school types offering academic upper secondary education, which coexist next to traditional academic-track schools. Chapter 3 investigates the effects of this, in international comparison, rather peculiar and subtle form of de-tracking upper secondary education.

1.3 Overview and Summary

This dissertation comprises three empirical studies that investigate different facets of institutional differentiation in the secondary school systems of Germany and England. Across chapters, the focus shifts from *vertical* differentiation between schools due to ability tracking to *horizontal* differentiation between vocational and academic curricula. Below, Table 1.1 gives a stylized overview of the type of differentiation considered in each chapter, together with other key facts of the analysis. The following briefly summarises each chapter in turn.

Chapter 2 focuses on vertical differentiation in Germany, investigating the effects of early between-school ability tracking on student achievement. As explained above, the long-standing debate on tracked *vs.* comprehensive schooling revolves around a perceived efficiency-equity trade-off: while homogeneous learning environments should make teaching more efficient through better targeting of content and teaching style to median classroom ability, students sorted into lower tracks might systematically lose out through unfavourable motivational and peer effects. However, in the face of severe selection problems—in particular, the endogenous selection of students into tracks and of countries into a tracking or non-tracking approach—the effects of early between-school ability tracking remain contested (e.g., Hanushek and Woessmann, 2006; Betts, 2011; Dustmann *et al.*, 2017).

To make progress on this question, I employ a research design that exploits institutional differences in the degree of tracking between German federal states: in all states about 40% of

[3] Note that despite the convention of calling these tracks 'academic' and 'non-academic', lower-secondary education in Germany is focused on general skills and knowledge across tracks.

students transition to separate academic-track schools after comprehensive primary school, at about age ten. Depending on the state, the remaining non-academic-track students are either further tracked between separate low- and intermediate-track schools or taught comprehensively for at least another two years. I compare the achievement growth of non-academic-track students during their first two years of secondary school between tracked and comprehensive states, while controlling for state-specific achievement trends using academic-track students in a triple-differences framework. I find evidence for positive average effects of prolonged comprehensive schooling on mathematics and reading achievement: point estimates equal 15% of a standard deviation (SD) in maths and 21% of a SD in reading. The average effects are almost entirely driven by large effects for low achievers, while effects for high achievers are null. This shows that too early and rigid forms of vertical differentiation can impair both the equity and the efficiency of school systems.

Both vertical and horizontal differentiation have been identified as obstacles to higher education expansion, which is an important goal for education policy given rising labour demand for high-skilled workers (Powell and Solga, 2011). As explained above, early ability tracking puts many students on non-university-bound routes from a young age onwards, while upper-secondary curricular tracking may divert students away from university-bound academic education into VET. However, while Germany eschewed structural reform of its tracking system, gradual and subtle changes have operated at its margins. Two notable changes were the introduction and gradual expansion of new school types which were added to the schools of the traditional tracking system: comprehensive schools (*integrierte Gesamtschule*) and vocational high schools (*berufliches Gymnasium/Fachgymnasium*). Both offer alternative pathways towards university-eligibility, which traditionally could only be obtained on academic-track schools. Hence, they constitute a partial de-tracking of upper-secondary schooling. Next to reducing vertical differentiation, vocational high schools also impact horizontal differentiation because, while they offer the academic upper-secondary curriculum and award the corresponding university-entrance certificate, they augment the academic curriculum with vocational subjects in a student-chosen field of specialisation.

Because school supply depends on local authorities, there are large regional and temporal disparities in the availability of these new school types. Against this backdrop, ***Chapter 3*** studies the effects of local school supply on upper-secondary attainment in Germany's tracked school system, considering not only the *number* but also the *type* of locally available schools. Drawing on yearly administrative records that cover the universe of German students, schools, and graduates, we compile a county-level panel of local school supply and upper-secondary attainment for 13 cohorts between 1995 and 2007. We document that, while attainment has substantially expanded, so has regional dispersion, pointing to growing inequality of educational opportunities. Cross-sectionally, we find that the supplies of all three school types with (academic) upper-secondary provision correlate positively with attainment, but for comprehensive and academic-track schools this association is largely spurious, i.e., due to regional differences in educational demand. For vocational high schools, in contrast, we find robust evidence for a positive supply-

side effect on attainment, confirmed in two-way fixed-effects, difference-in-differences, and event-study models. Our estimates suggest that, at the extensive margin, the introduction of a single vocational high school, when it was not previously available in a county, increases attainment rates by about 1.5 percentage points. Across extensive and intensive margins, a supply increase of one school per 1,000 students (corresponding to an increase of about one slot per 100 students) increases attainment rates by roughly the same amount. The hybrid nature of vocational high schools, combining academic and specialised curricula, might attract students who otherwise would be diverted from academic upper-secondary education towards vocational training.

Finally, ***Chapter 4*** focuses solely on horizontal differentiation, studying labour market returns to vocational upper-secondary education in England. In England, but also in countries like the US, decreasing job prospects for workers with low levels of education and skill shortages in technical occupations have led to a heightened interest in improving and expanding vocational curricula in secondary education. Often in reference to countries with strong apprenticeship systems, like Germany, VET is heralded as a means to relieve skill shortages, while improving the employment and earnings prospects of non-university-bound students. However, as discussed above, critics fear that doing so might divert students from academic routes that lead to university and hamper the acquisition of general skills, which might be more valuable in labour market in the longer run. The returns to vocational education are particularly unclear in market-oriented economies and education systems where firm involvement in VET is rare, both of which applies to the English setting (and the US). We contribute to this debate by delivering causal evidence on labour market returns to vocational education in England, where at the age of 16, after completing compulsory schooling, students choose between a vocational track, an academic track and no further education.

Our research design leverages multiple instrumental variables to estimate margin-specific treatment effects, i.e., causal returns to vocational education for students at the margin between vocational and academic education and, separately, for students at the margin between vocational and no post-16 education. Using the fact that vocational and academic education are offered by distinct institutions, identification comes from plausibly exogenous variation in distance to the nearest vocational provider conditional on distance to the nearest academic provider (and *vice-versa*), while controlling for granular student-, school- and neighbourhood-level characteristics. We draw on linked administrative education and earnings data, through which we can follow the universe of secondary school graduates through post-compulsory education and into the labour market until age 29.

We find that the vast majority of marginal vocational students are at the margin with academic education (instead of no further education), so that the first-order effect of expanding vocational-track access is diversion from the academic track. Diversion leads to losses in earnings at age 29 of about 9% for males and 7% for females. These effects are not driven by employment but due to wages (or working hours). A substantial part of them can be explained by reduced university degree completion, especially for males. Consistent with comparative advantage, we

Table 1.1. Overview of chapters.

	Chapter 2	**Chapter 3**	**Chapter 4**
Country	Germany	Germany	England
Type of differentiation under study	Ability tracking (*vertical)*	Ability tracking and some curricular tracking (*vertical & horizontal)*	Curricular tracking (*horizontal*)
Affected phase of schooling	Lower secondary	Lower and upper secondary	Upper secondary
Ages during exposure	10–12	10–18/16–18	16–18
Level of treatment variation	States	Counties	Students
Primary outcome	Achievement in school	Attainment at the end of school	Labour market performance
Data type	Student-level survey data (*random sample*)	County-level aggregate administrative statistics (*population*)	Student-level linked administrative education and tax data (*population*)
Data source	National Educational Panel Study (NEPS) & IQB National Assessment Studies	Official statistics of the German federal states ("Regionaldatenbank")	Longitudinal Education Outcomes (i.e., various admin. registries linked by England's Department for Education)
Methodology	Differences-in-differences, triple-differences, value-added	Two-way fixed-effects, differences-in-differences, event-study	Multiple instrumental variables (IVs)

find that returns to vocational education increase with students' preferences for the vocational track, even turning positive for males with the highest preferences. For the few marginal students at the margin with no further education, we find tentative evidence of positive employment and earnings effects but results are imprecise and insignificant. Altogether, our findings caution against an expansion of vocational upper secondary education in England in its current form.

1.4 Contributions

Through providing rigorous empirical evidence on the consequences of institutional differentiation in secondary school systems, this dissertation contributes to the debate on how school systems are to be reformed to become more efficient and equitable in their production of human capital. Each chapter in this dissertation makes specific contributions to the economics of

education literature, which are outlined in more detail in the introduction and conclusion of each individual chapter. This section highlights a few common contributions.

The primary contribution of this dissertation is that, together, the three chapters offer a unified study of vertical and horizontal differentiation in secondary education—two concepts that, though related, are conceptually distinct and have often been conflated in prior research. With respect to the former, the results in chapter 2 provide a cautionary tale about early and rigid forms of ability tracking. They show that there are limits to efficiency gains from classroom homogeneity as other mechanisms, such as peer effects, motivation and aspirations, start to depress achievement at the bottom once a system becomes too vertically differentiated. With respect to Germany, my results show that the newer two-tiered structure improves over the traditional three-tiered one, because it benefits low-achieving students without affecting their higher-achieving peers. At the same time, the results in chapter 3 suggest that comprehensive schools, which often come to replace the low- and intermediate-track schools in the two-tiered system, do relatively little to raise achievement at the top, given that we find no effect on attainment of the highest school-leaving certificate. In this regard, the hybrid model of vocational high schools proves more effective, drawing a population of students into academic upper-secondary education that otherwise would have pursued vocational education.

Chapter 4 suggests that, for such students at the margin between vocational and academic education, the latter does indeed pay off in the labour market. Of course, this evidence comes from England, where VET is arguably of lower quality than in Germany, and therefore we cannot be sure it extrapolates. Yet, as vocational high schools teach vocational courses alongside the academic curriculum, it is hard to imagine why these students would not benefit. Indeed, recent evidence on early career labour market effects of vocational high schools (Zimmermann, 2019) and similar hybrid upper-secondary institutions in England (so-called 'University Technical Colleges') (Machin *et al.*, 2020) looks promising. Generally, the case for curricular tracking in upper-secondary school is stronger than it might seem in light of the negative average effect of vocational *vs.* academic education we find in chapter 4. Remember that we document sorting on comparative advantage, including positive returns for those with the highest preferences for vocational education. So, while the balance in England may be slanted too far towards the vocational track, there likely are many students who do benefit from VET. Consistent with this notion, we estimate positive (yet far from significant) point estimates for the returns for students at the margin with no further education. For Germany, one can reasonably expect benefits from vocational education to be, if anything, more pervasive, given the superior quality of VET.

A methodological focus on the identification of causal effects is the second key contribution of this dissertation. As highlighted above, while the internationally comparative literature has identified the importance of institutions for human capital production, causal effects of *specific* features of school systems are often elusive. With respect to ability and curricular tracking, identification has proven particularly difficult because students who attend different tracks are strongly (self-)selected; different tracking systems are hard to meaningfully compare between

countries; and within-country changes in these institutions are seldom, idiosyncratic and also potentially endogenous.

To still arrive at above conclusions, all three chapters employ research designs that seek to carefully isolate exogenous variation in the educational institutions of interest in order to identify causal effects. In particular, chapter 2 exploits that the first two grades of secondary school in Germany offer a unique window with clear-cut differences in ability tracking between states. Accordingly, achievement differences at the end of primary school, which is comprehensive in all states, can be used to net out confounding between-state differences in student achievement after these two years. Similarly, chapter 3 exploits the differential expansion of different school types between counties across time to net out confounding country- and cohort-specific factors. Chapter 4 seeks out variation in the education choices of English students that is only due to differences in their geographical proximity to different education providers, but unrelated to other characteristics that determine students' earnings potential. On the basis of these 'natural experiments', I estimate effects by means of state-of-the-art econometric techniques, including (non-parametric) differences-in-differences (DD), triple-differences, value-added, two-way fixed-effects, event-study and instrumental variables (IV) models.

A third, connected, contribution is the variety of complementary datasets that are used to implement these research designs, ranging from student-level survey data created for educational research, over aggregate statistics published by the federal states for monitoring purposes, to student-level linked administrative education and earnings data, originally collected for tax purposes (see Table 1.1). The level of detail and the panel structure of the NEPS used in chapter 2 allow me to investigate effect heterogeneity and mechanisms particularly well, but the relatively modest sample size raises concerns about sampling variation. Accordingly, I corroborate my findings in the much larger but less detailed IQB National Assessment Studies. In contrast, chapter 3's aim to study local school supply in Germany cannot be achieved with survey data, which typically is not representative at levels lower than the national or, at best, state level. Accordingly, we build a novel dataset of school supply and attainment at the county level from a variety of aggregate statistics. This allows us to inspect regional differences in a granularity not previously possible. Implementation of the IV-based approach in chapter 4 requires granular information on place of residence at age 16 and labour market outcomes later in life for a very large number of students to obtain sufficient statistical power. The English administrative data meets these demanding requirements, allowing us to uncover rare exogenous variation in upper-secondary choices at the *student-level*.

The fourth contribution concerns the range of outcomes considered. As summarised in Table 1.1, across chapters the 'treatments' under study respectively take place during the lower-secondary, during the lower-secondary (for comprehensive schools) and upper-secondary (for comprehensive and vocational high schools) and during the upper-secondary phase, respectively concerning only ability tracking, both ability and curricular tracking, and only curricular tracking. To decide which outcome to inspect in each chapter I follow theory: the rationale of vertical differentiation is more efficient teaching of general skills. Hence, in chapter 2 I study effects

on students' maths and reading *achievement* during school. A key concern for either type of differentiation is reduced higher education participation. Hence, in chapter 3 I study effects on student *attainment* of the university-entrance certificate. Finally, the rationale of horizontal differentiation is to equip students with valuable skills for the labour market. Hence, in chapter 4 I study effects on students' medium-term *employment and earnings*. Therefore, together the chapters provide a comprehensive evaluation of differentiation in secondary school.

Finally, this dissertation contributes by putting a strong emphasis on identifying effect heterogeneity (in the two chapters which use individual-level data). This is key given my aim to evaluate the consequences of differentiation for both efficiency *and* equity. Under ability tracking students are sorted into different school types by their ability. Accordingly, chapter 2 stratifies the estimates by previous achievement, uncovering that the positive average effect of comprehensive *vs.* tracked schooling is driven by low-achieving students, corresponding to the population that is sorted into low-track schools under tracking. Importantly, I reveal that higher-achieving students are not hurt by prolonged comprehensive schooling, thus rejecting the presence of an efficiency-equity trade-off in this context. This exercise then guides the search for potential mechanisms, which focuses on peer quality and educational aspirations for lower-track students.

In the case of chapter 4, the analysis' very premise is the importance of effect heterogeneity: we start from the conceptual point that a crucial source of variation in returns to vocational education might stem from different counterfactual choices. Students whose alternative to vocational education is drop-out might greatly benefit, whereas students whose alternative is academic education might lose out. Hence, to uncover *alternative-specific* effects of vocational education a novel IV-based identification framework proposed by **?** is applied. Furthermore, as students *self*-select into curricular tracks, the central dimension of sorting in the case of horizontal differentiation is not ability but students' preferences. Accordingly, the paper uses the marginal treatment effects framework by Heckman and Vytlacil (2005) and local regressions to stratify effect estimates along the unobserved preference space. This reveals that, while returns are negative for the average marginal student, students sort on gains and effects are positive for those with the highest preferences, suggesting that vocational education may well be beneficial for a large group of inframarginal students.

Chapter 2

Better Together? Heterogeneous Effects of Tracking on Student Achievement*

Abstract

I study the effects of early between-school ability tracking on student achievement. My research design exploits institutional differences between German federal states: in all states about 40% of students transition to separate academic-track schools after comprehensive primary school, at about age ten. Depending on the state, the remaining non-academic-track students are either further tracked between separate low- and intermediate-track schools or taught comprehensively for at least another two years. I compare the achievement growth of non-academic-track students during their first two years of secondary school between tracked and comprehensive states, while controlling for state-specific achievement trends using academic-track students in a triple-differences framework. I find evidence for positive average effects of prolonged comprehensive schooling on mathematics and reading achievement. The average effects are almost entirely driven by large effects for low achievers, while effects for high achievers are null. This shows that too early and rigid forms of tracking can impair both the equity and the efficiency of school systems.

*This chapter has been published in The Economic Journal, Volume 131, Issue 635, pp. 1269–1307, https://doi.org/10.1093/ej/ueaa106. I benefited from comments and suggestions by the editor Gilat Levy, three anonymous referees, Jan Marcus, Guglielmo Ventura, Heike Solga, Katharina Spieß, Jan Paul Heisig, seminar participants at WZB, DIW and CEP/LSE, as well as participants of the EALE conference 2018, COMPIE conference 2018, DFG SPP 174 workshop 2018, the BeNA labour workshop 2017. I acknowledge financial support by the Jacobs Foundation and German Federal Ministry for Education and Finance (BMBF) through the College for Interdisciplinary Educational Research (CIDER). This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort (SC) Grade 5, doi:10.5157/NEPS:SC3:8.0.1, and SC Kindergarten, doi:10.5157/NEPS:SC2:8.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the BMBF. As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

2.1 Introduction

In the face of decreasing employment opportunities for low-skilled workers, the pressure on education systems to equip students with the necessary skills to succeed in modern labour markets is growing (European Commission, 2014). Woessmann (2016) demonstrates that national school systems differ markedly in how well they live up to this task. This raises the question how the optimal school system should be organised. While the (positive) effect of some institutional features of school systems on student achievement is relatively well-established by now (e.g., central exit exams), others remain fiercely debated. One of the most controversial issues in this regard is the practice of ability tracking. Tracking means grouping students by ability into vertically ordered school tracks. Countries differ widely on the degree to which they track students, and the age at which students begin to be tracked (Betts, 2011). Some countries, like Finland, eschew tracking altogether, relying only on comprehensive compulsory schooling. Others, like Germany, separate students into one of three ranked schools types at an age as early as 10. Between these two extremes lie countries like the US, which stream students into different tracks within schools.

The argument behind grouping students by ability is always one of efficiency.[1] Proponents of tracking posit that lower variance classrooms allow for better tailoring of curricula, instruction speed and pedagogy to students' abilities and should, therefore, benefit learning for all students (Duflo *et al.*, 2011). Critics, in contrast, fear that only high track/ability students benefit from tracking, whereas students assigned to lower tracks are condemned to lower achievement compared to a scenario with comprehensive schooling. Indeed, there are many mechanisms that might make the effects of tracking heterogeneous. First, to the extent that high performing peers are beneficial to learning (or low performing ones harmful), tracking mechanically exerts an unequal influence as it deprives lower track students of more able peers (Sacerdote, 2011). Second, there might be motivational consequences of separating students by ability. Lower track students, knowing they are deemed to be of lower aptitude, might feel discouraged and reduce their learning efforts. Third, if (financial) resources differ between tracks, students of certain tracks might be disadvantaged (Betts, 2011). Additionally, even if ability tracking is theoretically Pareto efficient, practical implementation is likely to be error-prone as ability is not directly observable and proxies like teacher assessments and tests are noisy and can be socio-economically biased (Brunello *et al.*, 2007; van Ewijk, 2011).

Given these opposing mechanisms, the net effect of tracking on student achievement is theoretically ambiguous and ultimately an empirical question. If proponents are right and homogeneous classrooms increase the effectiveness of teaching, tracking should raise average achievement by benefiting students of all ability levels. If the hypothesised negative effects are at work, tracked school systems should depress student achievement at the bottom. In terms of efficiency, the net effect of tracking then depends on scope and relative strength of these ef-

[1] The debate on tracking being a long-standing one, there is a vast social-scientific literature that discusses its pros and cons. For seminal contributions see, e.g., Oakes (1985), Gamoran and Mare (1989) and Slavin (1990).

fects.[2] In terms of equity, tracking might thus translate small performance differentials at young ages into substantial inequalities in later life. These dynamics should be more pronounced the earlier tracking starts, as divergences can accumulate, and in between-school tracking systems as compared to within-school ones, as the vertical differentiation between tracks is stronger (Betts, 2011).

Indeed, achievement differences between students of different tracks are large and well-documented (e.g., Dustmann, 2004) and countries with more rigid tracking systems tend to exhibit higher levels of educational inequality (Waldinger, 2007). The problem is that such correlational findings, whether at the individual or the country level, are likely to suffer from severe endogeneity. Students are not randomly allocated to school tracks but explicitly selected on ability. Similarly, countries' educational systems are affected by historical factors that also directly influence student outcomes. In the face of these selection problems no clear consensus on the effect of early between-school tracking has emerged in the empirical literature. While, in line with theory, effect estimates for high-ability students seem to vary between positive and null, the evidence on how tracking affects low-ability students is more mixed.

This paper exploits unique within-country between-state variation in tracking practices in Germany to isolate the effect of early between-school tracking on the achievement of students in lower tracks. While in all German states primary school is comprehensive, the grouping of students in secondary school, which commences in fifth grade when students are about ten years of age, differs between states: some states have a three-tiered and others a two-tiered secondary school system. In the former, students are tracked between low-, intermediate- and academic-track schools based on their performance in primary school. In states with the two-tiered regime, low- and intermediate-track schools were conflated, so that students are only tracked between academic and non-academic-track schools. Also these combined non-academic-track schools sort students by ability eventually, but in the first two years of secondary school (i.e., in grades 5 and 6) classes are formed disregarding previous performance or ability. Academic-track schools, called *Gymnasium*, do not differ between states and cater to about 40% of students in either regime. Accordingly, between-state differences in tracking are relevant only for the non-academic part of the student body: after comprehensive primary school, non-academic-track students are either directly tracked into low- and intermediate-track schools or taught comprehensively for another two years.[3] Note that these between-state differences pertain to the ability grouping of students only, as curricula are fully general in the first years of secondary school everywhere.

My research design exploits this variation in tracking in a difference-in-differences (DD) framework: I estimate how the achievement of one cohort of non-academic-track students develops differently over the first two years of secondary school depending on whether students are tracked or taught comprehensively. This strategy controls for grade-constant heterogeneity between states and general achievement trends between grades. Because the DD estimate might

[2] It appears that the costs of tracked and untracked school systems are roughly comparable (Hanushek and Woessmann, 2006). Following the literature, I therefore loosely refer to differences in mean outcomes as efficiency differences.

[3] This refers to all 12 federal states under investigation (out of 16 in total); see section 2.2.1 and footnote 18.

still be confounded by state-specific achievement trends, additionally, I compare the between-state differences for non-academic-track students to those for academic-track students, for whom there is no difference in tracking between states (who are thus 'untreated' no matter the state). This is implemented via a triple-differences (DDD) estimator. After having thus established the mean effect, I explore the distributional consequences of tracking. First, I provide non-parametric density estimates of the impact of tracking on the overall achievement distribution. Second, I explore how the effect of tracking depends on students' position in the pre-tracking achievement distribution.

The analysis is based on individual-level panel data for mathematical and reading competence from the German National Educational Panel Study (NEPS), which followed one cohort of students over their school career. The NEPS provides measures of student achievement right before and after the first two years of secondary school (i.e., right before and after the grade window with clear-cut between-state differences in tracking), as well as detailed information on students' family backgrounds and schooling inputs. In addition, I draw on the Institute for Educational Quality Improvement's (IQB) National Assessment Studies to corroborate my findings in larger samples and to assess the persistence of effects through the end of lower secondary schooling.

In sharp contrast with the predictions of tracking advocates, my results suggest that early between-school tracking *decreases* student achievement. The average effect of continued comprehensive schooling in grades 5 and 6 on seventh-grade test scores is estimated to be 0.17 standard deviations (SD) in mathematics and 0.24 SD in reading. Even though these effects are not very precisely estimated they are statistically significant and remarkably stable across specifications that flexibly control for student and school characteristics, as well as the inclusion of academic-track students as an additional control group in the DDD model. Robustness checks, such as comparing achievement trends in primary school and excluding outlier states, lend further credence to the causal interpretation of, at least, the direction of the effect estimates. Finally, the analysis with the IQB data shows that, while there is some fade-out over time, comprehensively taught non-academic-track students are still significantly better off towards the end of ninth grade.

The heterogeneity analysis reveals that these results are driven by the lower tail of the initial achievement distribution: for low-achievers effects are large and persistent, whereas for high-achievers effect estimates are insignificantly different from zero (yet, strictly non-negative). Consequently, comprehensive schooling has an equalising effect on the distribution of test scores. Delaying tracking does not trade off efficiency against equity, but seems to enhance both. I provide a discussion of the channels through which the effect might operate and find empirical support for the importance of peer effects and socio-emotional mechanisms, like improved school-related motivation and educational aspirations.

Note that the treatment effect identified in this paper pertains to a population of students that excludes the group of highest achievers in academic-track schools. Hence, one cannot directly extrapolate from these results to the effects of fully comprehensive school systems.

Still, they prove wrong the premise that there is a monotonously positive relationship between classroom homogeneity and student learning. Early between-school tracking appears to impose large costs on low-achieving students. Accordingly, more dispersed achievement distributions in more tracked systems do not appear to be a mere artefact of selection and the oft-voiced equity concerns in this context seem warranted.

This paper contributes to the literature on the systemic impact of between-school tracking, which, in the face of the endogeneity issues involved, could only produce tentative evidence so far. The most credible results stem from two strands of the literature.[4] The first exploits temporal *within-country* variation in tracking practices induced by de-tracking reforms. The second leverages the large variation in tracking practices *between countries* in different ways.

A number of prominent studies of the first strand analyse de-tracking reforms in the Nordic countries. Similar to my findings, they find reform-induced achievement gains for students from lower socio-economic backgrounds (see Meghir and Palme, 2005, for Sweden; Aakvik *et al.*, 2010, for Norway; and Kerr *et al.*, 2013, for Finland). Given that these reforms simultaneously changed other features of the school system, like the minimum school-leaving age, the effects cannot be unequivocally attributed to tracking, however. Analyses of Britain's de-tracking reform, which all use the fact that implementation was staggered across regions, have generated more mixed results.[5] Pischke and Manning (2006) argue that this is due to unobserved regional heterogeneity that cannot sufficiently be controlled for with existing data sets.

An arguably cleaner natural experiment, yet more narrow in scope, is the experience of Northern Ireland, which maintained its tracking system but increased the share of students admitted to the high track. Interestingly, the findings concerning the top end of the achievement distribution (medium high performers joining high performers) mirror mine for the bottom end (low performers joining medium performers): weaker students' gains from entering higher track environments are large, whereas losses for the stronger students are small or absent (Guyon *et al.*, 2012). Similarly, Piopiunik (2014) finds that a reform-induced increase in tracking in the German state of Bavaria led to achievement losses at the bottom.[6] A potential explanation for these results (and mine) is offered by Garlick (2018) who shows that low-achieving students are more sensitive to peer group composition than high-achievers, explaining the negative net effect of a residential tracking policy in South Africa.

Studies of the second strand have employed different strategies to circumvent the potentially severe endogeneity problems that come with between-country comparisons. One rather descriptive strategy limits attention to inequality, comparing only family background effects between tracked and untracked countries. These studies generally find that early between-school tracking is associated with steeper socio-economic gradients for student achievement (see, e.g., Brunello and Checchi, 2007; Schütz *et al.*, 2008).

[4] This brief literature review focuses on papers analysing the *systemic* effects of *between-school* tracking. The discussion of a large related literature on the effects of within-school streaming is deferred to the conclusion.

[5] See Kerckhoff *et al.* (1996); Galindo-Rueda and Vignoles (2004); Pischke and Manning (2006).

[6] The Bavarian pre-post differences analysed by Piopiunik (2014) closely resemble the (contemporaneous) differences in tracking analysed in this paper. Reassuringly, his findings based on a single state's reform are confirmed in this study for the whole of Germany.

A second strategy, introduced in a seminal paper by Hanushek and Woessmann (2006), is based on the observation that primary school is comprehensive everywhere, regardless of how the secondary school system looks. These studies use DD to estimate how test scores change differently from primary to secondary school between countries with tracked and comprehensive schooling. Most results indicate that tracking increases inequality in student achievement,[7] though Waldinger (2007) argues that these results are sensitive to the way countries are categorised into tracked and untracked ones. This highlights a major problem of the cross-country literature: when classifying countries as comprehensive or tracked, a range of quite heterogeneous between-school tracking systems are lumped together and compared to an even more diverse group that includes both comprehensive and within-school streaming systems. Hence, the treatment (and the counterfactual) is not clearly defined. Other problems include that also *changes* in outcomes might be related to unobserved differences between tracked and untracked countries (Betts, 2011) and the pooling of incomparable test scores (Contini and Cugnata, 2016).

My study merges the approaches of the within- and the cross-country literatures. I adopt the logic of Hanushek and Woessmann's (2006) DD approach in comparing changes in test scores between elementary and secondary school for the identification of the effect of tracking. Yet, the fact that I exploit within- instead of cross-country differences allows me to improve on a number of important points. First, apart from differences in tracking, school systems are strongly harmonised between German states such that the treatment is clearly defined in my case. Therefore, second, the common trends assumption necessary for DD is much more plausible in my setting than in previous studies. Crucially, I can directly assess its plausibility *ex post* using academic-track students for whom there is no difference in tracking. Third, individual-level panel data allows me go beyond mean effects and estimate how the effect of tracking depends on students' position in the initial achievement distribution. This is key given that the debate on tracking revolves around a perceived efficiency-equity trade-off.

Finally, it is important to highlight that I estimate a systemic (state-level) effect of tracking, which contrasts with a literature on marginal effects. For example, Dustmann *et al.* (2017) also study the German context but, using an individual-level instrumental variables strategy, find no effect of track placement on educational attainment or earnings for students *at the margin between two tracks*.[8] Though important for understanding the consequences of (mis)allocation of hard-to-assign students to tracks given an early between-school tracking system, their estimate tells us little about whether tracking is desirable in the first place. My results suggest that the separation of students into different schools at an age as early as 10 depresses achievement for a sizeable group of (non-marginal) low-achievers, thus putting them at a double disadvantage.

The paper is structured as follows: section 2.2 lays out the institutional background and the identification framework. Section 2.3 describes my data sources and presents descriptive findings. Section 2.4 present the estimation results, including robustness checks and a discussion of potential mechanisms. Finally, section 2.5 discusses implications and concludes.

[7] Next to Hanushek and Woessmann (2006), see Ammermüller (2013) and Schwerdt and Ruhose (2016).

[8] This is in the spirit of a larger literature on the benefits of entering selective schools (e.g., Abdulkadiroğlu *et al.*, 2014).

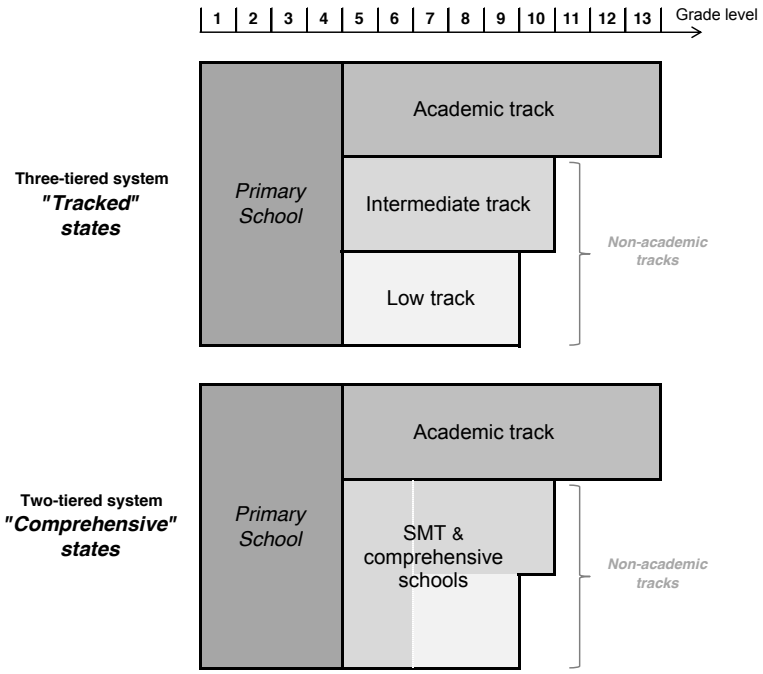


Figure 2.1. Schematic overview of the two tracking regimes in Germany.

Notes: For illustrative purposes the figure abstracts from the fact that in some of the three-tiered Tracked states there are some comprehensive schools (see text and Table 2.1). Academic track = *Gymnasium*, Intermediate track = *Realschule*, Low track = *Hauptschule*, School with multiple tracks (SMT) = *Schule mit mehreren Bildungsgängen*, Comprehensive schools = *Integrierte Gesamtschule*.

2.2 Institutional Background and Research Design

2.2.1 The German School System and Heterogeneity Therein

In Germany, sovereignty over education policy lies with the state governments. In order to ensure the comparability of educational standards and degrees, however, the federal Standing Conference of the Ministers of Education and Cultural Affairs of the States (*Kultusministerkonferenz*) harmonises education policies between states considerably (Kultusministerkonferenz, 2014). Within this unique situation of educational federalism, a school system has developed that is fairly homogeneous across Germany in terms of basic structure, teaching methods and curricula, but exhibits fine differences within some areas of schooling policy—especially, school structure and, thus, tracking practices. It is this heterogeneity within a context of general comparability that I exploit to shed light on the impact of tracking on student achievement.

Throughout Germany, compulsory schooling starts at the age of 6 with primary school, which covers the first four grade levels and is taught comprehensively with no ability grouping of students within or between schools.[9] Differences between states emerge after the end of comprehensive primary school. They are summarised schematically in Figure 2.1.

[9] In the two states of Berlin and Brandenburg, primary school lasts six years. For this reason, they are not part of the analysis.

The traditional (West) German secondary school system is three-tiered: upon leaving primary school after fourth grade, i.e., around the age of 10, students are tracked into one of three vertically ordered school types—*Hauptschule*, *Realschule* and *Gymnasium*, representing low, intermediate and academic track—based on their previous performance.[10] These tracks lead to different school-leaving certificates and differ substantially in terms of years of schooling, curriculum, teacher certification and peer composition. The academic track (i.e., *Gymnasium*) has the most demanding curriculum, lasts eight or nine years and is the only track leading directly to a school-leaving certificate that entitles to entry into university. This makes for a clear divide between the academic and the non-academic segments of the school system (also in reputation). The intermediate track (i.e., *Realschule*) provides general knowledge, lasts six years and is supposed to prepare students for advanced vocational and professional education. If students complete the intermediate track successfully and meet state-specific requirements they may upgrade to the academic track after grade 10. The low track (i.e., *Hauptschule*) provides a more basic general education, lasts five or six years and prepares students for technical vocational education. Also here, after completion, upgrading to intermediate-track schools is possible under specific conditions.

Currently, five states still have the traditional three-tiered system (see Figure 2.2).[11] The rest has a two-tiered secondary school structure that distinguishes between academic- and non-academic-track schools only. This group consists of three East German states, which never adopted the three-tiered system, and four West German states that reformed their system.[12]

The East German states had to align their (comprehensive) school system with that of the West after German reunification. This led to a compromise where the East adopted the *three-tiered* differentiation in school-leaving certificates but opted for a *two-tiered* school structure (Edelstein and Nikolai, 2013).[13] Instead of separate low- and intermediate-track schools there is only one non-academic school type, labelled 'School with Multiple Tracks' (*Schule mit mehreren Bildungsgängen*; henceforth SMT). Here, all students not attending an academic-track *Gymnasium* school are taught together. If a student leaves an SMT after five years (without failing the year, of course) she receives the low degree. If she stays on for another year and attains the necessary grades she earns the intermediate degree. Hence, the difference between the two systems is one of tracking only.

In many Western states low-track schools have become stigmatised due to falling student numbers and a lack of prospects for its graduates (Helbig and Nikolai, 2015). This led several states to reform their school system along the lines of the two-tiered system. Like in the East, the three different school-leaving certificates, as well as a distinct academic track, were retained,

[10] In all states students receive a track recommendation by their teacher based on their performance in primary school. Whether it is binding depends on the state. All results are fully robust to the inclusion of an indicator variable for binding teacher recommendations (and that indicator variable always turns out to be insignificant itself; see Appendix Table B2.3). Therefore, all that follows abstracts from this difference between states.

[11] These are Bavaria, Baden-Württemberg, Hesse, *Lower Saxony* and *North Rhine-Westphalia*.

[12] East German states: Saxony, Saxony-Anhalt and Thuringia. West German states: Bremen, Hamburg, Saarland and Schleswig-Holstein.

[13] Except for Mecklenburg-Vorpommern, which initially adopted and briefly maintained a three-tiered system. For reasons discussed below, this state is not part of the analysis.

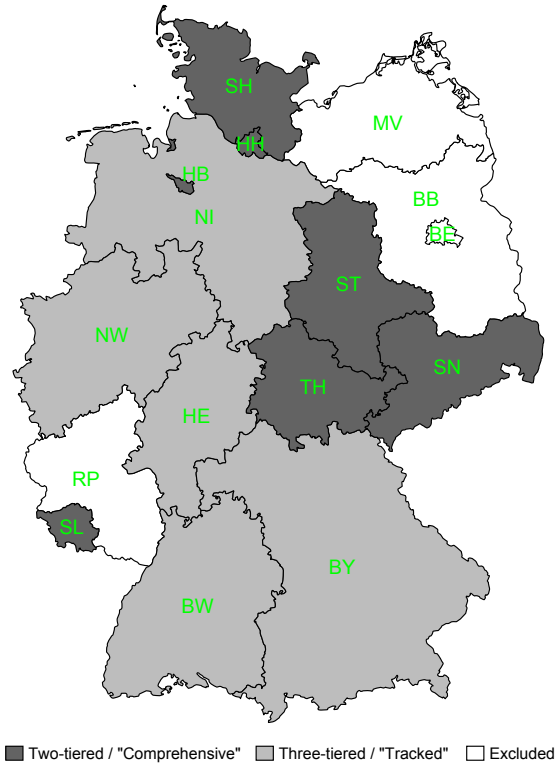


Figure 2.2. German federal states coloured by tracking regime.

Notes: BW = Baden-Württemberg; BY = Bavaria; BE = Berlin; BB = Brandenburg; HB = Bremen; HH = Hamburg; HE = Hesse; MV = Mecklenburg-Vorpommern; NI = Lower Saxony; NW = North Rhine-Westphalia; RP = Rheinland-Pfalz; SL = Saarland; SN = Saxony; ST = Saxony-Anhalt; SH = Schleswig-Holstein; TH = Thuringia.

but separate low- and intermediate-track schools were abolished and replaced with so-called 'comprehensive schools' (*Gesamtschule*). Thus, just like SMTs in the East, these schools comprise all non-academic-track students.[14]

While both SMTs and comprehensive schools track students internally in higher grades, they are prohibited from doing so in the first two years of secondary schooling (i.e., in grades 5 and 6) (Leschinsky, 2008). Instead, in these two grades classes continue to be formed disregarding ability or previous performance. Only from grade 7 onwards these schools may track students by forming track-specific classes (i.e., separate low- and intermediate-track classes) or by applying subject-specific ability sorting.[15] In most states, it is up to schools to decide if and how to group students starting in grade 7 and, unfortunately, this information is not centrally collected. Accordingly, between-state (and school) differences in higher grades are blurry. The first two years of secondary school, however, provide a time window where institutional differences regarding the tracking of students are clear-cut.

[14]The difference between the East German SMTs and the West German comprehensive schools is that in the former only the basic and intermediate degrees can be obtained, while in the latter, mostly, all three degrees can be earned (Helbig and Nikolai, 2015). In practice, this difference is only relevant in much later grades than those studied here.

[15]Schools that use degree-specific within-school streaming from grade 7 onwards are labelled 'cooperative' while those that generally continue to teach comprehensively, except for streaming in particular subjects, are called 'integrative'.

Comparability is further bolstered by the fact that the first two years in non-academic secondary schools are strongly harmonised. Official information of the Kultusministerkonferenz (2014) shows that curricula and learning goals for grades 5 and 6 in the non-academic tracks focus on the acquisition of a standard set of basic general knowledge that is virtually indistinguishable between states. By way of example, average weekly instruction hours in mathematics do not differ between to the two-tiered (4.3 hours/week) and three-tiered states (4.4 hours/week).[16] In most three-tiered states, curricula for grades 5 and 6 do not even differ between low- and intermediate-track schools, though the level of detail in which the material is treated might be higher in intermediate-track schools (due to students' higher ability levels) (Bald, 2011). States ensure the compatibility of curricula during the first two years of secondary school because they formally allow for the possibility to switch between tracks (Bellenberg, 2005). In practice, this happens quite rarely (only about 5% of students switch according to Bellenberg, 2012).

The analysis below focuses on the cohort of students that entered fifth grade/secondary school in 2010. The following summarises the previous discussion's key points: In the four years prior to the transition, all of these students attended comprehensive primary school. Moreover, in all states the highest achieving students transition to separate academic-track *Gymnasium* schools. The remaining non-academic-track students, however, are either further *tracked* between two different school types (in states with the three-tiered regime) or taught *comprehensively* for another two years (in states with the two-tiered regime).[17] For ease of exposition, I will refer to the five states with a three-tiered system as the 'Tracked' states and to the seven states with a two-tiered system as the 'Comprehensive' states (see Figure 2.2). Four states with school systems that do not fit this classification had to be excluded from the analysis.[18]

2.2.2 Identification Strategy

The idea of this paper is to use the institutional differences in the non-academic segment of the school system between Tracked and Comprehensive to learn about the effects of early between-school ability tracking. The main challenge for this endeavour is that states' tracking policies

[16] These numbers are based on official time-table regulations for grades 5 and 6 in non-academic-track students reported in Pant *et al.* (2013).

[17] For completeness, it should be mentioned that in some of the Tracked states municipalities are allowed to offer comprehensive schools, where all three degrees can be earned, next to the ordinary schools of three-tiered system. For the purposes of this paper this can be thought of as non-compliance with regards to the treatment of comprehensive schooling (see section 2.3.4).

[18] Berlin, Brandenburg and Mecklenburg-Vorpommern are excluded because the tracking decision is made after grade 6 instead of after grade 4. *Rheinland-Palatinate* is excluded because the state was transitioning from a three-tiered to a two-tiered system during the period under investigation and, hence, its treatment status is ambiguous. While *de jure* all separate low- and intermediate-track schools should have been closed by 2010, both the official statistics and the current data set show some students entering such schools in 2010, indicating that *de facto* the fade-out took longer. It seems that these schools were closed in the following years and students re-assigned. Administrative records show that the cohort's share of students in a low- or intermediate-track school declined from 8% in 2010 to 6% in 2011 to 3% in 2012 (Statistisches Bundesamt, 2001a). A robustness check where *Rheinland-Palatinate* is assigned the Tracked states (as initially there was some tracking) leaves all results unchanged (see Appendix Table B2.3).

might correlate with a whole range of other, potentially unobserved, factors determining student achievement, such as student body composition or early childhood education policies.[19]

To account for such unobserved differences between states, my identification strategy uses test scores taken at two points in the educational career of students. The first achievement test is administered right after primary school, at the beginning of grade 5, and the second two years later, at the beginning of grade 7. All students are taught comprehensively in primary school. Hence, grade 5 scores should be unaffected by tracking and capture achievement differences between states unrelated to the tracking system. Grade 7 scores measure achievement right after exposure to either treatment condition and thus reflect both causal effects from tracking and permanent between-state differences. To purge the seventh-grade comparison between Comprehensive and Tracked states of grade-constant confounders I thus propose the following difference-in-differences (DD) design:

$$Y_{isg} = \delta_0 + \delta_1 \, Compr_s + \delta_2 \, Grade7_g + \beta_{DD} \, (Compr \times Grade7)_{sg} + \psi \, \mathrm{X}_{isg} + \theta_s + u_{isg}, \tag{2.1}$$

where Y_{isg} is the test score of non-academic-track student i in state s and grade level $g \in \{5, 7\}$, $Compr_s$ and $Grade7_g$ are indicator variables for the Comprehensive states and grade 7 scores, respectively, X_{isg} is a row vector of predetermined student and school covariates, discussed in further detail below, and θ_s are state fixed effects.

In equation (2.1), δ_1 absorbs level achievement differences between the two state groups at the end of primary school, while δ_2 absorbs general achievement trends between grades. Accordingly, the DD estimate β_{DD} captures the differential development of non-academic-track students in the two-tiered system compared to the three-tiered system. To interpret β_{DD} as the causal effect of comprehensive *versus* tracked schooling we require two assumptions: first, that primary school achievement is indeed unaffected by the structure of the secondary school system and, second, that in the absence of differences in tracking, non-academic-track achievement would have developed in parallel between Comprehensive and Tracked states.[20]

A threat to the first assumption are incentive effects: knowing that they will be placed in different tracks depending on their performance in primary school, students might increase their study efforts already prior to the start of tracking in more tracked regimes (Eisenkopf, 2007).

[19] Formally, let Y^1_{isg} denote the (potential) achievement of student i in state s in grade g under tracking and Y^0_{isg} (potential) achievement under comprehensive schooling. The identification challenge is that the average treatment effect is not generally equal to the observed mean difference between Tracked ($Compr_s = 0$) and Comprehensive ($Compr_s = 1$) states: $\tau_{ATE} = \mathbb{E}[Y^1_{is7} - Y^0_{is7}] \neq \mathbb{E}[Y_{is7}|Compr_s = 0] - \mathbb{E}[Y_{is7}|Compr_s = 1]$.

[20] Continuing the potential outcomes notation from above, formally, we require the following two assumptions:

$$\begin{cases} Y_{is7} = (1 - Compr_s) * Y^1_{is7} + Compr_s * Y^0_{is7} \\ Y_{is5} = Y^0_{is5} \end{cases} \qquad \text{(Observation rule)}$$

$$\mathbb{E}[Y^0_{is7}|Compr_s = 0] - \mathbb{E}[Y^0_{is5}|Compr_s = 0] = \mathbb{E}[Y^0_{is7}|Compr_s = 1] - \mathbb{E}[Y^0_{is5}|Compr_s = 1] \qquad \text{(Common trends)}$$

Then, DD identifies the average treatment effect on the treated, i.e., the effect of tracked *versus* comprehensive schooling for (non-academic-track) students from the Tracked states: $\tau_{ATT} = \mathbb{E}[Y^1_{is7} - Y^0_{is7}|Compr_s = 0] = \{\mathbb{E}[Y_{is7}|Compr_s = 0] - \mathbb{E}[Y_{is5}|Compr_s = 0]\} - \{\mathbb{E}[Y_{is7}|Compr_s = 1] - \mathbb{E}[Y_{is5}|Compr_s = 1]\} = -\beta_{DD}$. (I define comprehensive schooling as the treatment and tracking as the control condition in the regression formulation because in Germany it is more intuitive to think of the newer two-tiered system as the treatment. This is without loss of generality.)

Below I explore the importance of this mechanism directly by comparing the two state groups' achievement trends in primary school. For now, however, note that the test scores used in this paper are not used for students' track placement, ruling out immediate incentive effects related to the tests. Further, note that the presence of the academic track creates strong performance incentives for ambitious students (and their parents) in both regimes, dramatically limiting the importance of this mechanism compared to previous applications.

With respect to the second, 'common trends' assumption, a standard concern in DD designs is sample compositions changing differentially between treatment and control groups between periods. Given that I am comparing one cohort across grade levels this is unlikely to play an important role: students would need to strategically move to another state or from academic to non-academic tracks (or vice versa) *after having started secondary school*. I will confirm that this is not the case by means of balance tests on an array of observed predetermined student covariates and, additionally, condition on these covariates, X_{isg}, in the DD regression.

In the current setting, the more serious threats to the common trends assumption come from two sources: systematic differences between the two state groups in *student composition* and in *schooling inputs*. Regarding the former, there are, indeed, non-negligible differences in student bodies between Comprehensive and Tracked states (see Table 2.1). If, by the end of primary school, the achievement of different types of students not only differs in levels but also continues to develop differently, DD is biased because it merely removes grade-constant achievement differences. To address this concern, I increase the flexibility of the (conditional) DD model by adding interactions between predetermined student characteristics, X_{isg}, and grade level. This allows for different development trajectories for different types of students. The sensitivity of the DD estimate to this exercise is informative of the extent to which such confounding might play a role.

Turning to the latter, note that different schooling inputs in lower secondary school can be considered 'co-treatments': factors other than tracking that change differently between states between primary and secondary school. In that case, the DD estimate would no longer represent the sole effect of tracking but include the effect of other features of the school environment. To see if differently equipped secondary schools between Comprehensive and Tracked states play an important confounding role, we can proceed as before and inspect the sensitivity of β_{DD} to the addition of school input measures to the control set.

Even if these exercises leave the DD estimate unchanged, concerns about unobserved grade-specific differences between states that violate the common trends assumption might remain. Fortunately, the current setting offers the unique opportunity for an additional test. As explained in the previous section, the distinction between the Comprehensive and Tracked States is only meaningful for students in the non-academic tracks. For academic-track students, there is no difference between the two regimes as they enter *Gymnasium* schools after fourth grade everywhere. Under the assumption that selection into the academic track does not differ between the

two state groups, they can, therefore, be used as a control group to test for potential regime-specific trends in achievement that the DD model does not pick up.[21]

This additional control group comparison is easily implemented by the following difference-in-difference-in-differences (DDD), or triple-differences, model, which is estimated over all students and hence adds the subscript $t \in \{academic,\ non\text{-}academic\}$ for track:

$$Y_{istg} = \lambda_{sg} + \phi_{tg} + \mu_{st} + \beta_{DDD}\,(Compr \times Grade7 \times NonAcad)_{stg} + \psi\,\mathrm{X}_{istg} + e_{istg}, \tag{2.2}$$

where $NonAcad_t$ is an indicator variable for non-academic-track students, λ_{sg}, ϕ_{tg} and μ_{st} are state-grade, track-grade and state-track fixed effects, respectively, and the remaining variables are defined as before. The triple interaction takes value one for grade 7 observations of non-academic-track students in the Comprehensive states. Accordingly, the DDD estimate β_{DDD} measures how comprehensively taught non-academic-track students progress differently in the first two years of secondary school net of state-specific achievement trends as approximated by academic-track students.

If the estimates for β_{DDD} and β_{DD} are roughly identical this indicates that achievement trends in the academic track are roughly identical in Tracked and Comprehensive states. This should increase one's confidence that there are no state-specific trends confounding the DD estimate from above and that the assumptions for it to be interpreted causally hold. If the two estimates differ, then there are divergent trends in the academic track. Causal interpretation of DDD then hinges on the assumption that academic-track students provide an good approximation of non-academic-track students' counterfactual achievement trends.

In terms of inference, the group-level treatment variable means that I need to account for clustering at the state level when estimating the above regression models (Bertrand *et al.*, 2004; Abadie *et al.*, 2017). As in the current setting there are only twelve states, the large sample assumptions necessary for a conventional cluster robust variance estimator are unlikely to hold (Mackinnon and Webb, 2017). Therefore, throughout this paper, inference is based on a wild cluster bootstrap (Cameron *et al.*, 2008), which has been shown to perform well in settings with few clusters (e.g., Mackinnon and Webb, 2017).[22]

[21] The crucial assumption that selection into the academic track is identical between the two state groups is discussed in further detail and tested below.

[22] The wild cluster bootstrap permutes the outcome variable based on 'restricted' residuals (i.e., those stemming from coefficient estimates that impose the null hypothesis to be tested) and weights from a Rademacher distribution. Webb (2014) shows that with 12 or less clusters, a specific six-point distribution is preferable over the Rademacher distribution. Hence, I implement the latter. However, results do not substantially differ between the standard (Cameron *et al.*, 2008), an unrestricted (Mackinnon and Webb, 2017) or a schools-as-'sub-clusters'-of-states (MacKinnon and Webb, 2018) version of the bootstrap.

2.3 Data, Descriptive Statistics and Preliminaries

2.3.1 Data Sources and Analysis Samples

In this section, I present a brief overview of the data used in this study. For a more detailed discussion of the data sets used, the construction of my samples, as well as sample diagnostics the interested reader is referred to Appendix 2.5.

2.3.1.1 National Educational Panel Study (NEPS)

The main empirical analysis is based on data from Starting Cohort 3 (SC3) of the multi-cohort German National Educational Panel Study (NEPS) (Blossfeld *et al.*, 2011). The NEPS-SC3 survey randomly sampled from the population of newly minted fifth-graders in the school year 2010/11 and, thereafter, followed this cohort over time as it progressed through grade levels of the German school system.

Student achievement is measured using the NEPS-SC3's competence tests in mathematics and reading.[23] The first round of tests was administered in autumn of 2010, two to four months into students' first year of secondary school. Hence, the grade 5 scores should not yet be severely affected by students' secondary school environment and can be conceptualised as the pre-tracking measure of achievement required by the DD design. Note that the DD estimate is attenuated towards zero to the extent that grade 5 scores are already affected by tracking.[24] Accordingly, my estimates should be conservative. Restricting the sample to students on regular schools in one of the twelve states under investigation, the NEPS-SC3 grade 5 cross-section comprises 4,448 students with non-missing test scores, of whom 2,303 are in the non-academic tracks, of whom 330 are from the Comprehensive states.[25]

Students were tested again two years later, at the beginning of the 2012/13 school year, when the cohort in question had just entered seventh grade. The NEPS-SC3 grade 7 cross-section comprises 5,316 students with non-missing test scores, of whom 2,771 are in the non-academic tracks, of whom 552 are from the Comprehensive states. It consists of students who were already part of the survey in fifth grade and a large randomly drawn refreshment sample to counteract attrition, explaining the larger sample sizes.

As repeated cross-sections suffice for estimating the main DD and DDD models, for these regressions I simply pool the NEPS-SC3 grade 5 and grade 7 cross-sections. The resulting NEPS

[23] While, in principle, the NEPS also assesses competencies in other domains, only in maths and reading testing commenced in the first wave of the survey. As the DD design requires pre-treatment outcomes, my analysis restricts attention to maths and reading achievement.

[24] This is because any differences in achievement between states caused by the first couple of months of tracking are absorbed in the baseline and thus cancelled out in the calculation of the double difference. Note that this unless the effect of tracking reverses within the first couple of months of exposure: if the very short-term effects of tracking (i.e., the effects on achievement after 3 months of exposure) are *opposite* to the longer-term effects of tracking I am trying to estimate (i.e., the effects on achievement after 2 years of exposure), the DD estimate could theoretically by biased upwards. As the effects reproduce in the IQB data, which measures pre-tracking achievement at the end of fourth grade, this does not seem to be the case.

[25] Note that smaller number of observations in Comprehensive states simply reflects that these states are smaller and less populous than the Tracked states (also see the map in Figure 2.2).

DD sample includes only non-academic-track students and comprises 5,074 student×grade observations (882 of which are from the Comprehensive states). The NEPS DDD sample adds academic-track students as an additional control group for, in total, 9,764 student×grade observations (1,711 of which are from the Comprehensive states). I standardise the maths and reading test scores to have mean zero and standard deviation one in the group of Tracked states' non-academic-track students (i.e., the 'control group'), separately by grade level. Accordingly, all treatment effects in this paper can be interpreted in standard deviations of test scores.[26]

In contrast to estimation based only on the panel sample, my use of repeated cross-sections retains students who drop out between waves and includes the refreshment sample. This has two advantages: First, it maximises sample sizes and, hence, precision in the estimation of my main models—a key concern given the NEPS' humble sample sizes in the Comprehensive states. Second, the use of the refreshment sample reduces sample selection bias due to attrition. This is because panel non-response is negatively related to achievement and, hence, attrition is substantially higher in the non-academic tracks (29% compared to 13% in the academic track). The majority of observations in the refreshment sample are from the non-academic tracks (63%), thus restoring the representativeness of my sample. Note that there are no significant differences in student-level attrition between the Tracked and Comprehensive states.[27]

Nevertheless, for a number of heterogeneity analyses I leverage the panel structure of the NEPS and use the sub-sample of students for whom I observe both grade 5 and grade 7 test scores. This NEPS 5-to-7 panel sample comprises 3,521 students, of whom 1,646 are in the non-academic tracks, of whom 269 are from the Comprehensive states.

Finally, a third round of achievement tests was administered in the 2014/15 school year, when the cohort in question was in ninth grade. The grade 9 tests are used to assess effect persistence.

To probe the robustness of the DD estimates to the inclusion of controls for student characteristics, I draw on detailed information from the NEPS' student and parent questionnaires. In particular, I construct the following student-level control variables: age, sex, migration background, single parent household, foreign language spoken at home, highest level of parental education measured in four categories, monthly household income, receipt of unemployment benefits and a standardised index for home possessions.

Additionally, I use information from the principal and teacher questionnaires to construct the following school-level controls, aimed to capture school quality independent of tracking: average teacher age, days of further training received by teachers over the past year, school size measured by the number of students per cohort, student-teacher ratio and four composite indices for schools' facilities; extracurricular programmes; educational support offers and quality

[26] Note that the NEPS competence tests are designed to measure the progress of students *on one scale* across grade levels. This 'linking' of scales is achieved through the recurrence of certain anchor items in each wave of the test (see Fischer *et al.*, 2016, for details). For simplicity I nonetheless standardise scores within each grade level. As the DD design identifies a seventh-grade-specific treatment effect it is more intuitive to interpret the estimates in seventh-grade standard deviations. Results are virtually unchanged when using a cross-grade standardisation scheme, however.

[27] See Appendix 2.5 for a detailed analysis of panel attrition in the NEPS-SC3.

control measures.[28] Note that pre-treatment achievement (i.e., grade 5 scores) is a function of primary school inputs, whereas post-treatment achievement (i.e., grade 7 scores) is a function of secondary school inputs. However, only the secondary school environment is observed in the NEPS-SC3 as it commenced in fifth grade. Thus, to impute the missing primary school inputs in the DD sample, I use data from the NEPS' primary school cohort, Starting Cohort 2 (NEPS-SC2). In particular, I calculate state-level averages for all school-level controls in the primary school data and assign each grade 5 observations in the DD sample its state-level average.

Moreover, I use the NEPS' primary school cohort to investigate pre-tracking achievement trends. For this, I apply the DD model of equation (2.1) to grade 2 and grade 4 mathematics test scores from the NEPS-SC2.[29] The NEPS-SC2 follows a later cohort than the NEPS-SC3 but, given that there were no major changes to primary education in Germany in this time period, their trends should be similar.

2.3.1.2 IQB National Assessment Studies

Next to the panel structure that can be exploited for heterogeneity analyses, the main advantage of the NEPS is that it measures student achievement in seventh grade, right after exposure to either treatment condition and when between-state differences in tracking and other school policies are still clear-cut. Its main downside is the modest number of observations in the Comprehensive states, raising concerns about sampling variation. As a robustness check, I therefore double check my results using two large cross-sectional student assessments carried out by the Institute for Educational Quality Improvement (IQB).

The IQB studies do not randomly sample from the population of all students in a particular grade level like the NEPS but, instead, draw separate random samples of roughly similar sizes within each state. Hence, in the IQB data I achieve much larger samples with rough parity in the number of observations between Tracked and Comprehensive states. Due to this sampling design all analyses with the IQB data use student sampling weights to obtain estimates representative of Germany. The IQB data's main downside is that post-treatment outcomes are measured in ninth grade, meaning that the estimates represent a mixture of effect persistence and effects from continued (but somewhat unclear) differences in tracking and other schooling inputs.

The IQB National Assessment Study 2011 (IQB11) tested fourth-graders in maths, reading and listening at the end of the 2010/11 school year, when students were at the end of their primary school time (see Stanat *et al.*, 2012, for details). This is one cohort later than that of the main analysis (see Appendix Figure A2.1), so that my analysis with the IQB data operates under the assumption that these two consecutive cohorts' primary school experiences match. Fourth-grade students are not yet assigned to academic- or non-academic tracks, but testing happened late enough in the school year for students' secondary school and, hence, track to be determined already. This allows me to classify students as non-academic or academic using

[28] For more information on these indices see Appendix 2.5.

[29] Unfortunately, reading scores are not available for these grades.

information provided by parents and teachers.[30] The IQB11 grade 4 cross-section comprises 18,904 students on regular schools with non-missing test scores, of whom 11,158 are assigned the non-academic tracks, of whom 6,573 are from the Comprehensive states.

The IQB National Assessment Study 2015 (IQB15) tested ninth-graders in reading and listening at the end of the 2014/15 school year, which is the same cohort as in the main analysis (see Stanat *et al.*, 2016, for details). All analyses with the IQB15 data restrict attention to students on regular non-academic-track schools with non-missing test scores. The non-academic-track IQB15 grade 9 cross-section comprises 13,742 students, of whom 7,009 are from the Comprehensive states. Analogously to above, I pool the non-academic parts of the IQB11 grade 4 and the IQB15 grade 9 cross-sections to construct the IQB DD sample, which comprises 24,900 student×grade observations (13,582 of which are from the Comprehensive states).

2.3.2 Descriptives and Balance Tests

The first two columns of Table 2.1 compare Comprehensive and Tracked states in terms of the distribution of students over tracks (panel A), pre-tracking achievement (panel B) and student characteristics (panel C) in the non-academic-track NEPS DD sample. For reference, column 5 describes academic-track students.

I discuss panel A below. First, note that pre-tracking achievement in panel B is extremely well balanced between the two state groups, as indicated by small and insignificant differences in test scores at the beginning of secondary school. Though pre-treatment balance in outcomes (or covariates) is not technically required by the DD design, this should raise one's confidence that student achievement is generally comparable between Comprehensive and Tracked states. Stark differences in mean scores between non-academic- and academic-track students of (more than) one standard deviation indicate that track assignment is very much a function of achievement despite the absence of strict cut-off rules. Consequently, the treatment variation in the non-academic segment of the school system analysed in this paper concerns a negatively selected group of students. Still, test score distributions of academic- and non-academic-track students overlap substantially (see next section).

In terms of student characteristics, panel C of Table 2.1 shows moderate compositional differences between the two state groups, highlighting that simple cross-sectional comparisons between states might well be confounded. On the one hand, the Comprehensive states, composed mostly of the poorer East German states and city states, score slightly worse on socio-economic variables like household income and unemployment. On the other hand, they have lower shares of students with migration background, mainly reflecting the different migration histories of West and East Germany. Importantly, however, column 4 shows that there is no significant double difference in any of the covariates, indicating that these differences in sample composition stay roughly constant between grades. Appendix Table B2.1 provides a detailed comparison of school characteristics in primary and secondary school in both the NEPS and the IQB data, which has

[30] The details of the classification procedure are described in Appendix 2.5.

Table 2.1. Descriptive statistics and balance tests in the NEPS data.

	Non-academic tracks				Academic track
	Compr. states	Tracked states	*p*-value (1)=(2)	*p*-value DD=0	(both)
	(1)	(2)	(3)	(4)	(5)
Panel A: Distribution over tracks/school types					
Share non-academic track	0.52	0.52	(0.96)	(0.44)	–
Of those:					
Low-track school	0.00	0.32	–	–	–
Intermediate-track school	0.00	0.55	–	–	–
Comprehensive/SMT school	1.00	0.12	–	–	–
Panel B: Pre-treatment outcomes					
Grade 5 maths score	0.06	0.00	(0.66)	–	1.28
Grade 5 reading score	0.01	-0.00	(0.93)	–	0.99
Panel C: Student characteristics					
Female (binary)	0.47	0.50	(0.02)	(0.92)	0.51
Age in fifth grade (in years)	11.00	10.96	(0.46)	(0.10)	10.67
Single parent household (binary)	0.20	0.13	(0.07)	(0.56)	0.08
Migration background (binary)	0.18	0.32	(0.05)	(0.72)	0.23
Foreign language at home (binary)	0.09	0.14	(0.17)	(0.90)	0.10
Highest parental education level:					
None, low, intermediate w/o appr.	0.16	0.28	(0.07)	(0.84)	0.06
Intermediate w/ apprenticeship	0.48	0.38	(0.31)	(0.59)	0.24
Academic track, some college	0.27	0.25	(0.77)	(0.62)	0.36
University degree	0.09	0.09	(0.91)	(0.13)	0.34
Monthly household income (in Euros)	2781	3072	(0.12)	(0.73)	4070
Unemployment benefits (binary)	0.20	0.11	(0.07)	(0.79)	0.03
Home possessions (index)	-0.31	-0.26	(0.39)	(0.55)	0.24
Observations	882	4192			4690

Notes: This table reports the distribution of students over tracks, as well as variable means for pre-treatment test scores and student covariates in the pooled grade 5 and grade 7 NEPS data. The first two columns describe the NEPS DD sample of non-academic-track students, separately by state group. Academic-track students, who are added as an additional control group in the DDD model, are described in column 5 (for brevity not split by state group). Corresponding to the later regressions, the shares in panel A and grade 5 test scores in panel B are unweighted. The remainder of the table uses student sampling weights to reflect the underlying populations as accurately as possible. Column 3 reports *p*-values from testing whether covariate means are equal in the Comprehensive and Tracked states. Column 4 reports *p*-values from testing for zero double differences (i.e. the second difference between the Comprehensive and Tracked states between grade 5 and grade 7). This tests whether the parallel trends assumption holds for the respective covariate. All tests are based on 999 wild cluster bootstrap iterations, clustering at the state level and using Webb weights.

considerably more power for inference at the school level, to show that the same holds true for these. As, especially among the school-level covariates, some of the level differences are not small the analysis below will pay close attention to the sensitivity of the DD estimates to the inclusion of these controls, despite the insignificance of the presented balance tests.

2.3.3 Selection into the Academic Track

The treatment a student receives depends on her state of residence (Tracked or Comprehensive) and whether she is assigned the academic track or not. My identification strategy requires that selection into the academic track does not differ between the two state groups. Otherwise, neither the academic-track nor the non-academic-track student bodies would be comparable.

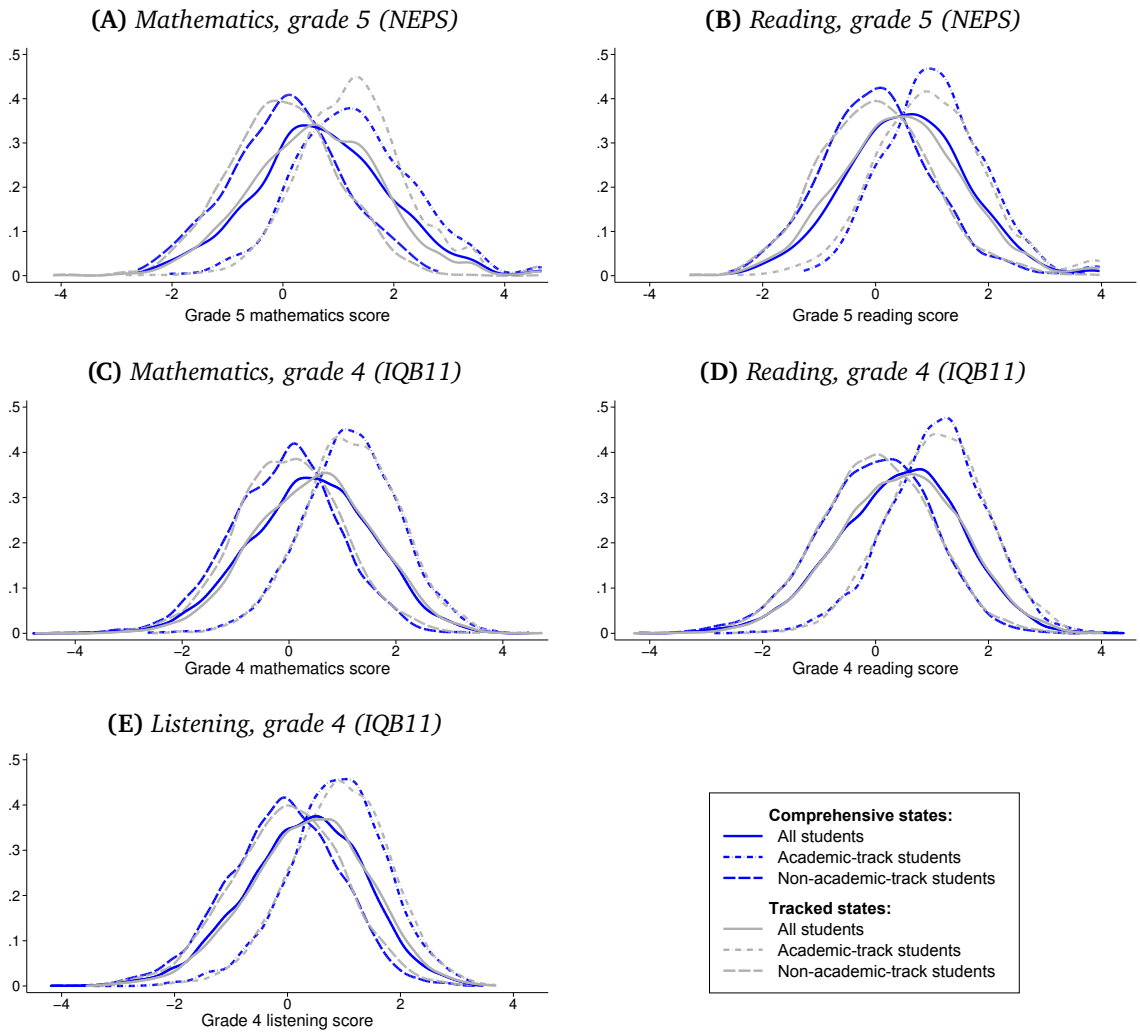


Figure 2.3. Pre-tracking test score distributions by track and tracking regime.

Notes: This figure shows kernel density estimates of different test score distributions for all, only non-academic-track and only academic-track students, separately for Comprehensive and Tracked states. All density estimates use a Gaussian kernel with optimal bandwidth for normally distributed variables (Silverman, 1986). Panels A and B are based on the NEPS-SC3 grade 5 cross-section. Panels C, D and E are based on the IQB11 grade 4 cross-section, using the first plausible value of each test score and student sampling weights.

Panel A of Table 2.1 shows that the non-academic-track sample shares are 52% in both state groups. Academic-track students appear to be slightly overrepresented in the NEPS, as according to administrative records the true non-academic shares for the cohort in question are 60% and 57% in Tracked and Comprehensive states, respectively (Statistisches Bundesamt, 2001a). Reassuringly, the shares are very similar both in the population and in my sample.

Equal shares leave open the possibility of compositional differences, however. For example, it is conceivable that competition for the academic track is stronger when there are only two tracks, because the alternative school type necessarily comprises all low-achievers. This might amplify average ability differences between academic and non-academic tracks in two-tiered *versus* three-tiered systems. To test for the presence of such differences in selection, Figure 2.3 plots pre-tracking test score distributions by state group, both overall and for academic and

non-academic-track students separately. Panels A and B refer to the (beginning of) grade 5 maths and reading scores from the NEPS and panels C through D refer to the (end of) grade 4 maths, reading and listening scores from the IQB11 data. Across achievement domains and data sets, the distributions look very similar in Tracked and Comprehensive states; in particular, the gaps between the academic- and non-academic-track distributions do not seem to differ between states. Correspondingly, the mean gap between non-academic- and academic-track students does not significantly differ between the two states groups for any of the five scores.[31] Therefore, I conclude that selection into the academic track does not meaningfully differ between the two state groups.[32]

2.3.4 Distribution over Non-Academic Tracks

Panel A of Table 2.1 reports the distribution of non-academic-track students over different school types by state group. In the Tracked states one third of students attend low-track schools and about half attend intermediate-track schools. In the counterfactual scenario of a two-tiered school system these two groups would be taught together instead of being separated into different tracks. Note that a small percentage of students in the Tracked states (12%) attends comprehensive schools where all three degrees can be obtained and there might or might not be within-school streaming.[33] In the language of the treatment effects literature, these students can be thought of as 'always-takers', slightly attenuating my 'intent-to-treat' effect estimates towards zero.

In the Comprehensive states, there are no low- and intermediate-track schools. All non-academic-track students in these states attend SMT or comprehensive schools. As explained above, there is no within-school streaming in grades 5 and 6 in these schools. It is the effect of this comprehensive schooling for non-academic track students in the two-tiered regime, as compared to the between-school tracking in the three tiered-regime, that I aim to estimate.

2.3.5 Peer Group Composition

To get a better idea of what these differences in tracking mean from the perspective of students, Figure 2.4 explores how tracking affects students' peer group composition. Using the IQB11 (end of) grade 4 test scores in maths (panel A) and reading (panel B), the left-hand panels compare the relationship between students' own achievement and that of their classroom peers between Comprehensive and Tracked states in the final year of primary school, right before tracking commences. Using the NEPS (beginning of) grade 5 scores, the right-hand panels depict the

[31] The wild cluster bootstrapped p-values for these 'differences in differences' are 0.82 for maths and 0.60 for reading in the NEPS and 0.58 for maths, 0.55 for reading and 0.18 for listening in the IQB data.

[32] It might seem puzzling that alternative choice options are largely irrelevant for selection into the academic track. It likely is explained by the special status of the academic-track *Gymnasium* in Germany: virtually all ambitious and high-SES students will aspire to the academic track regardless of what other school forms are present because of its reputation and academic focus (Paulus and Blossfeld, 2007). For example, in my sample 78% of students with college-educated parents attend the academic track.

[33] These sample shares are very close to the true population shares: 32% low-track, 51% intermediate-track and 17% comprehensive schools (Statistisches Bundesamt, 2001a).

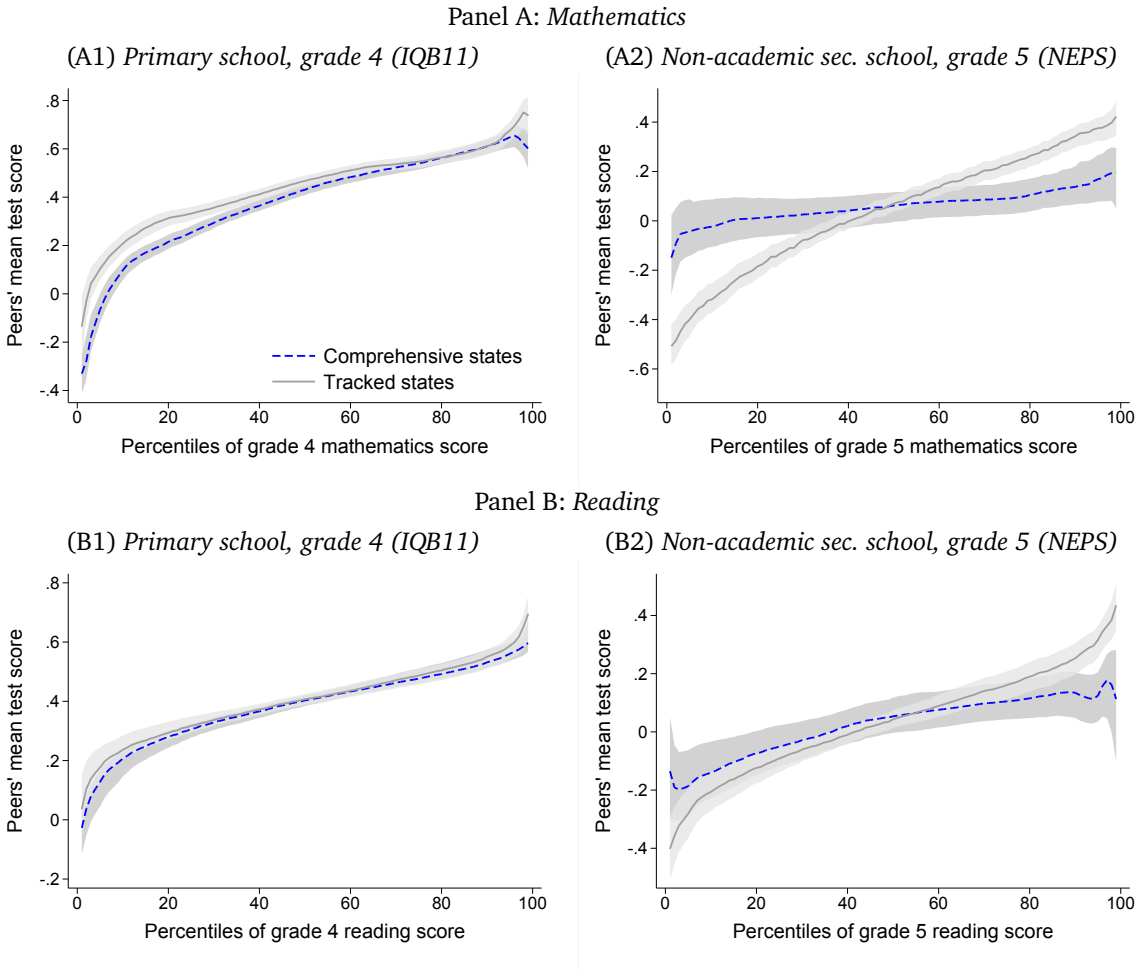


Figure 2.4. Effect of tracking on peer group composition.

Notes: All curves show fitted values from student-level local constant regressions of mean classroom test scores in maths (Panel A) and reading (Panel B) on students' own test score, separately for the Tracked and Comprehensive states. The fitted values are evaluated at each percentile of the respective test score distributions. The left-hand side figures are based on the IQB11 grade 4 cross-section (restricted to students classified as non-academic), thus describing the relation between own and classroom peers' performance at the end of primary school, right before tracking commences. The right-hand side figures are based on the NEPS-SC3 grade 5 cross-section (restricted to non-academic-track students), thus describing the same relation a couple of months later, when students have been tracked according to the state-specific rules. The shaded areas show pointwise 95% confidence intervals from 999 iterations of a percentile bootstrap, clustering at the classroom level.

same relationship a couple of months later,[34] when students have been tracked according to the state-specific rules.

Despite the absence of tracking, the relationship between students' own achievement and that of their peers is clearly positive already in primary school, representing residential sorting.[35] Importantly, however, this relation looks very similar in both state groups. If anything, the gradient for maths is slightly steeper in the Comprehensive states. With the transition to secondary school, this relationship changes quite dramatically between the two state groups. Now, the gradient is much steeper in the Tracked states, where non-academic-track students are assigned

[34]Note that this is under the assumption that this relation stayed constant between the two consecutive cohorts that are tested in IQB11 and NEPS-SC3.

[35]There are much more primary than secondary schools in Germany, such that students generally attend schools closer to home and residential sorting plays a larger role.

to different schools based on their previous performance, than in the Comprehensive states, where classes continue to be formed disregarding previous performance.[36] These differences in classroom heterogeneity form the core of the (composite) treatment of comprehensive *versus* tracked schooling.

2.4 Results

2.4.1 Level Effects of Comprehensive *versus* Tracked Schooling

This section presents my findings on the average effect of comprehensive, as compared to tracked, schooling in the first two years of lower secondary school for non-academic-track students. For illustrative purposes, I begin by comparing students' progress between Tracked and Comprehensive states graphically. Using the NEPS 5-to-7 panel sample, Figure 2.5 plots kernel density estimates of gain-score ($\Delta_7 Y_{is} = Y_{is7} - Y_{is5}$) distributions in maths and reading for academic- and non-academic-track students in both tracking regimes. A striking picture emerges: whereas academic-track students' progress is very similar between regimes in both domains—if anything, those in the Tracked states progress slightly more in maths—for non-academic-track students the Comprehensive states' distribution of gains appears to stochastically dominate that of the Tracked states. These graphs provide strong initial evidence for the existence of efficiency gains from *comprehensive* schooling. In the following, I assess the significance and robustness of this descriptive finding more formally by estimating the DD and DDD models.

Column 1 of Table 2.2 displays the regression results for the simple (unsaturated) DD model, corresponding to equation (2.1) without control variables and state fixed effects, estimated using the non-academic NEPS DD sample. The results for maths are presented in panel A and those for reading in panel B. Next to point estimates, in parentheses I present p-values and in brackets 95% studentised bootstrap confidence intervals from 999 wild cluster bootstrap iterations, clustering at the state level.[37] As indicated by small and insignificant coefficients on the Comprehensive states indicator, there seem to be no substantial level achievement differences between the two state groups prior to tracking. The DD coefficients, equal to the (double) difference between Comprehensive and Tracked states' achievement changes between grades 5 and 7, indicate that comprehensively taught non-academic-track students progress about 0.18 standard deviations (SD) more in maths and 0.26 SD more in reading, confirming the graphical finding from above.

The next columns probe the robustness of this result, following the steps outlined in section 2.2.2: Column 2 presents the saturated DD model, which replaces the Comprehensive states indicator with state fixed effects for increased flexibility and precision. Column 3 adds the

[36] Note that one cannot meaningfully compare the slopes between primary and secondary school without very strong assumptions, as IQB and NEPS differ in their test design and thus do not measure achievement on the same scale. Therefore, I restrict attention to between-state comparisons within (and not across) data sets.

[37] Note that in my case inference based on the wild cluster bootstrap is strictly more conservative than inference based on conventional cluster-robust standard errors, clustered at the state level (which in turn is more conservative than clustering at the state-track level, which in turn is more conservative than clustering at the school level). For details on the wild cluster bootstrap implementation used in this paper see Roodman *et al.* (2019).

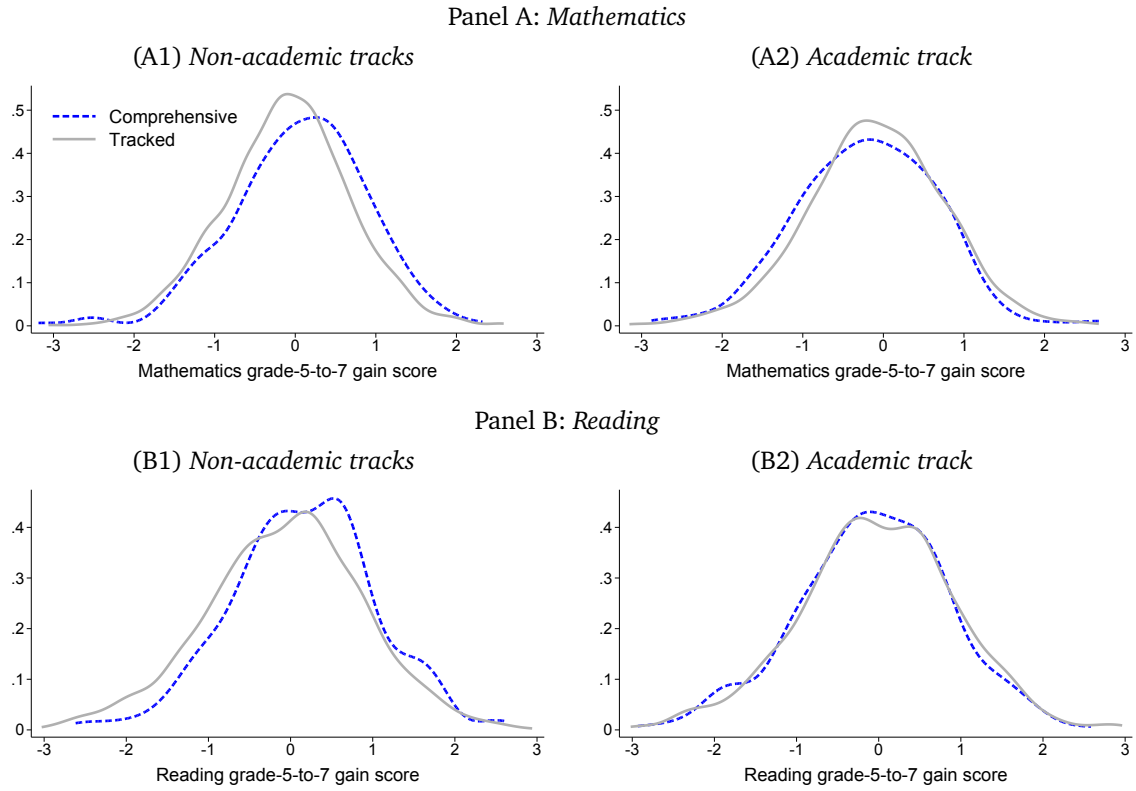


Figure 2.5. Distribution of grade-5-to-7 gain-scores by track and tracking regime.

Notes: This figure shows kernel density estimates of the grade-5-to-7 gain score distribution in maths (Panel A) and reading (Panel B), separately for Comprehensive and Tracked states, for non-academic-track (left) and academic-track students (right). All density estimates use a Gaussian kernel with optimal bandwidth for normally distributed variables (Silverman, 1986). Estimates are based on the NEPS 5-to-7 panel sample ($N = 1,646$).

complete set of student covariates, described in Table 2.1, to correct for potential compositional changes in the sample. Column 4 interacts the vector of student covariates with the grade 7 indicator to allow for different development trajectories for different types of students. Finally, column 5 adds the complete set of school covariates to control for potential differences in schooling inputs.

Considering the overall level of imprecision in the estimates due to the moderate number of observations in the Comprehensive states, the DD estimate stays remarkably stable across all specifications. As the student control variables in the NEPS are very detailed, substantially increasing the model's explanatory power as evidenced by the sharp increase in R^2 between columns 2 and 3, it is very unlikely that between-state differences in student body composition explain the advantage for comprehensively taught students. Despite my controls for schooling inputs being somewhat less detailed, the fact that the DD estimate *increases* upon their inclusion provides strong evidence against the (null) hypothesis that the effects are driven by differences in schooling inputs. Appendix Table B2.3 further corroborates the robustness of the DD results to alternative model specifications and the inclusion of further potential confounders at the

Table 2.2. Level effect of comprehensive schooling on seventh-grade achievement.

Model specification:	Double differences					Triple differences			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Mathematics									
Comprehensive schooling	0.176***	0.207***	0.167***	0.165***	0.259**	0.257**	0.214*	0.199*	0.187*
	(p = 0.01)	(0.01)	(0.00)	(0.01)	(0.04)	(0.03)	(0.07)	(0.10)	(0.10)
	[0.07, 0.40]	[0.10, 0.49]	[0.10, 0.29]	[0.09, 0.28]	[0.02, 0.60]	[0.03, 0.54]	[-0.01, 0.51]	[-0.04, 0.51]	[-0.03, 0.46]
Indicator Compr. states	0.061								
	(0.65)								
	[-0.33, 0.31]								
R^2	0.005	0.027	0.172	0.177	0.199	0.296	0.378	0.380	0.394
Panel B: Reading									
Comprehensive schooling	0.264**	0.277**	0.241**	0.243**	0.363***	0.293	0.235	0.231	0.286**
	(0.02)	(0.02)	(0.04)	(0.02)	(0.00)	(0.12)	(0.14)	(0.14)	(0.03)
	[0.04, 0.59]	[0.05, 0.64]	[0.02, 0.53]	[0.04, 0.48]	[0.20, 0.60]	[-0.06, 0.73]	[-0.07, 0.57]	[-0.06, 0.54]	[0.02, 0.59]
Indicator Compr. states	0.012								
	(0.94)								
	[-0.34, 0.32]								
R^2	0.007	0.021	0.108	0.113	0.131	0.220	0.273	0.276	0.285
Individual controls			✓				✓		
Grade × Ind. controls				✓	✓			✓	✓
School controls					✓				✓
State FE		✓	✓	✓	✓				
State×grade FE						✓	✓	✓	✓
State×track FE						✓	✓	✓	✓
Track×grade FE						✓	✓	✓	✓
N state clusters	12	12	12	12	12	12	12	12	12
N Compr. state students	882	882	882	882	882	1711	1711	1711	1711
N Tracked state students	4192	4192	4192	4192	4192	8053	8053	8053	8053

Notes: This table reports OLS regression results for the double- (DD) and triple-differences (DDD) models for fifth- and seventh-grade maths and reading test scores. The DD models in columns 1–5 are estimated using the NEPS DD sample of non-academic-track students. For the DDD models in columns 6–9 academic-track students are added to the sample. Column 1 reports results for the unsaturated DD model, i.e. from regressing test scores on an intercept, an indicator for the Comprehensive states, an indicator for grade 7 observations and their interaction. Column 2 report results for the saturated DD model, which replaces the Comprehensive state indicator with state fixed effects. Column 3 adds student covariates: sex, age, age squared, migration background, foreign language spoken at home, single parent household, household income, parental unemployment, parental education and an index for home possessions (incl. missing data indicators). Column 4 adds interactions between the grade 7 indicator and all student covariates. Column 5 adds school covariates: average teacher age, average days of further training received by teachers, school size, student-teacher ratio and indices for schools' equipment, extracurricular programmes, educational support offers and quality control measures (incl. missing data indicators). Column 6 reports the results for the saturated DDD model, i.e. from regressing all students' test scores on an indicator for non-academic-track grade 7 observations from the Comprehensive states and state-grade, state-track and grade-track fixed effects. Columns 7–9 add covariates to the DDD model analogously to before. p-values in parentheses and 95%-confidence sets in brackets stem from 999 wild cluster bootstrap iterations using Webb weights, clustering at the state level. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

school (e.g., private schools and class size) and state level (e.g., binding track recommendations and school funding).[38]

Despite the stability of the DD estimate across different control sets of varying flexibility, concerns about non-parallel (counterfactual) achievement trends might remain. Hence, as a more design- and less covariate-based test for the common trends assumption, column 6 presents the uncontrolled DDD model that uses academic-track students, for whom there are no differences in tracking between states, as an additional control group. Columns 7–9 add control variables analogously to before. As to be expected the DDD estimates are less precise than the ones from DD, but, reassuringly, they are very similar in magnitude—if anything, slightly larger. The similarity between the double- and triple-difference estimates implies that there are no divergent achievement trends between the Comprehensive and Tracked states in the academic track, which suggests that the different trends in the non-academic tracks are indeed due to differences in tracking.

Given the small number of states it is important to check that the results are not driven by any particular outlier state whose performance diverged extremely from the others. To this end I perform a simple leave-one-out analysis. Figure 2.6 plots coefficient estimates against p-values from repeatedly re-estimating the DD and DDD models eat time leaving out one state. While the precision is slightly affected when some (larger) states are dropped, the results appear robust to the exclusion of any particular state.

Given that differences in tracking between states only emerge with students' transition to secondary school, a natural requirement for interpreting my results causally is that Comprehensive and Tracked states exhibit parallel achievement trends in primary school, prior to tracking. Inspecting such pre-trends also allows to test for the presence of the discussed anticipation effects of tracking in secondary school on pre-tracking achievement. To this end, I first apply the unsaturated DD model from equation (2.1) to grade 2 and 4 maths scores in the NEPS primary school cohort and find no differences in second-grade achievement ($\delta_1 = 0.043$; $p = 0.77$) or achievement growth in the two years prior to tracking ($\beta_{DD} = -0.015$; $p = 0.77$).[39]

[38] In particular, I first add all school controls separately to ensure that coefficient movements are largely homogeneous across variables (columns 2–9). Then, I show robustness to applying student sampling weights (column 11) and using the *unsaturated* DD specification (column 12). Next, I interact the school controls with grade level just like the student controls (column 13). This drastically increases imprecision, as the data lacks power to allow this degree of flexibility at the school level—especially since in grade 5 these variables only vary at the state level. Still, point estimates are rather similar. Finally, in columns 14–19 I show that results are fully robust to controlling for a school-level indicator for private schools (which is excluded in the main regressions because there are so few private schools in the NEPS that the imputed primary school state averages are highly unrepresentative); a school-level measure of average class size (which is excluded because it is arguably a 'bad control': class sizes are likely to be a function of how students are sorted—e.g., low-track schools tend to have smaller class sizes); the time students are back in school since the end of the summer break at the day of testing; a state-level measure of per pupil public expenditure for schools (which is excluded because its comparability across states is somewhat doubtful, mainly because both living expenses and teacher salary-unrelated expenditures vary greatly between states); a state-level indicator for binding teacher recommendations (which is excluded due to its irrelevance—see section 2.2.1); and adding *Rheinland-Palatinate* to the sample as a Tracked state.

[39] See Appendix Table B2.2 for the complete regression results. Note that I cannot inspect pre-trends for reading, as reading scores are only available for grade 4.

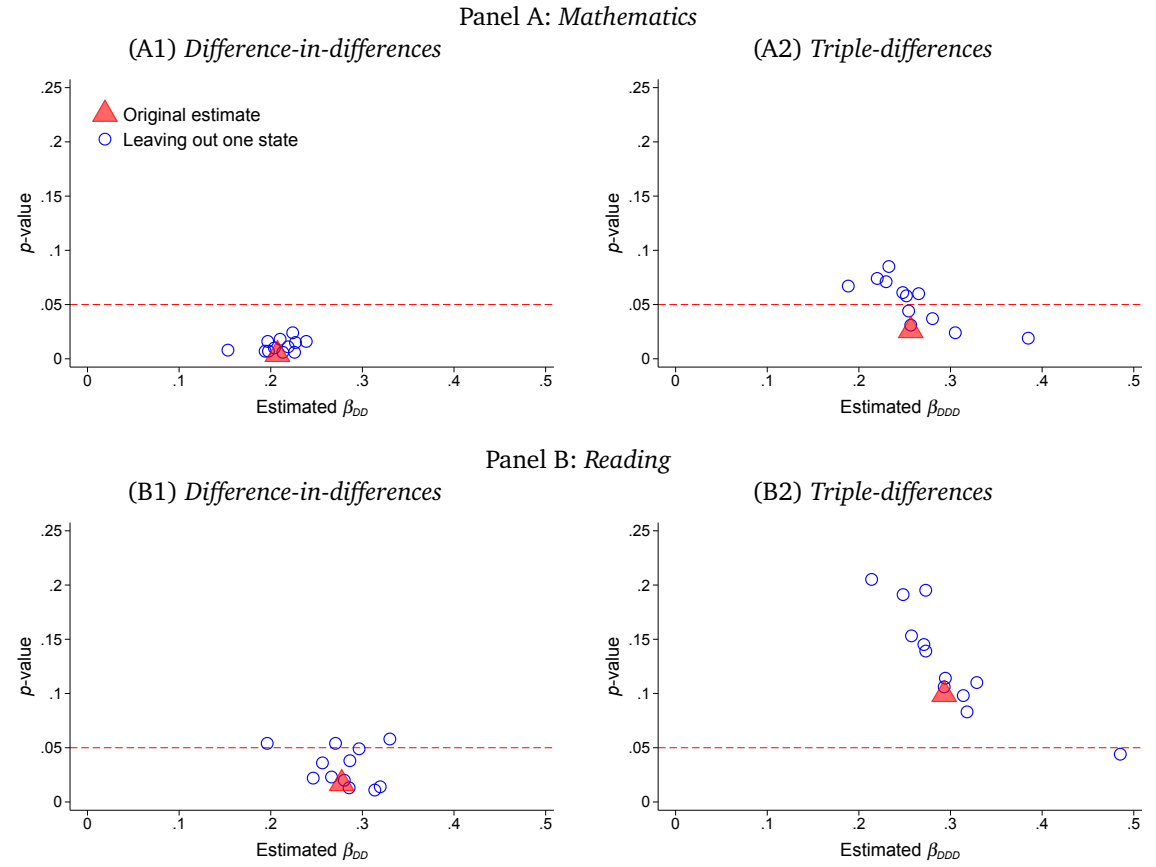


Figure 2.6. Leave-one-state-out DD and DDD estimates.

Notes: This figure compares the DD and DDD estimates for the effect of comprehensive schooling on seventh-grade maths (Panel A) and reading (Panel B) scores in the original sample with those obtained when excluding single states. Point estimates on the horizontal axis are plotted against *p*-values testing the null of no effect on the vertical axis, obtained from 999 wild cluster bootstrap iterations using Webb weights, clustering at the state level. Each panel has 13 data points: one triangle showing the original estimate and 12 circles showing the estimates when excluding each state. State names cannot be revealed for data confidentiality reasons.

This result concerns all and not only non-academic-track students: as students are not yet assigned to tracks in primary school I cannot restrict the sample accordingly. To investigate pre-trends specifically for lower achieving students, who are more likely to attend non-academic-track schools later on, second, I leverage the NEPS' panel structure and investigate achievement growth *by previous achievement*. In particular, using the NEPS 2-to-4 panel sample comprising all students for whom I observe both second and fourth grade maths scores, I non-parametrically estimate the difference in grade 2-to-4 gain scores between Comprehensive and Tracked states at different percentile of the grade 2 test score distribution.[40] The results, presented in Figure 2.7, indicate that achievement growth in the two years prior to tracking is indeed roughly parallel across the entire achievement distribution.

In summary, this section showed that achievement levels in the non-academic tracks diverge between Comprehensive and Tracked states during the first two years of secondary school. The

[40]Formally, I estimate the two conditional expectations $\mathbb{E}[\Delta_4 Y_{is} | Y_{is2}, Compr_s = k]$, for $k \in \{0, 1\}$, non-parametrically and evaluate their difference at every fourth percentile of Y_{is2}.

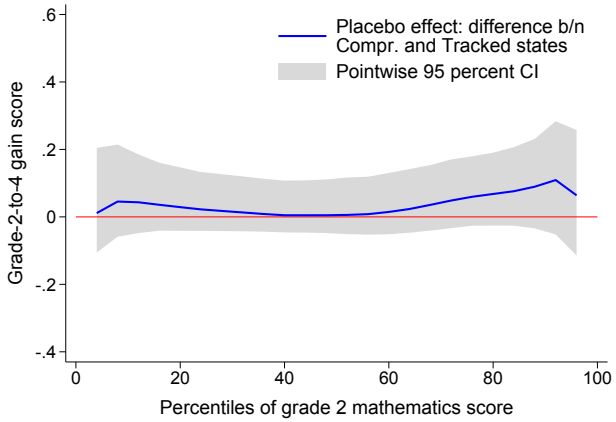


Figure 2.7. Pre-tracking achievement trends by previous performance.

Notes: This figure shows the difference in average second to fourth grade achievement growth between Comprehensive and Tracked states across the grade 2 maths score distribution. Estimation is based on the primary school NEPS-SC2 2-to-4 panel sample that includes all students for whom both grade 2 and grade 4 maths scores are observed ($N = 4,676$). The curve is constructed as follows: First, separately for Comprehensive and Tracked states, I estimate a student-level local constant regression of grade-2-to-4 gain scores on grade 2 test scores. Second, I calculate the difference between Comprehensive and Tracked states' fitted values at every fourth percentile of the grade 2 distribution. Third, I construct pointwise 95% confidence intervals from a percentile bootstrap with 999 iterations, clustering at the state level, stratifying by tracking regime and holding the bandwidth constant across bootstrap iteration (Hall and Kang, 2001).

presented evidence suggests that this divergence is caused by differences in ability grouping: comprehensive, instead of between-school tracked, schooling at the ages 10 through 13 appears to boost achievement for non-academic-track students—a group comprised of low and, as is visible from the pre-tracking achievement distributions displayed in Figure 2.3, also a considerable share of medium to high achievers. My preferred specification is the DD model in column 4 of Table 2.2, which flexibly controls for student characteristics but omits school-level controls that, despite being chosen carefully to only include inputs that are independent of a state's tracking policy, might raise concerns of controlling for mechanisms. Here, 95% confidence sets for the effect of comprehensive schooling are [0.09, 0.28] in maths, with a point estimate of 0.17 SD, and [0.04, 0.48] in reading, with a point estimate of 0.24 SD. Whilst the small sample size prohibits a more precise estimation of the average effect, these results strongly reject tracking proponents' claim that comprehensive schooling impedes achievement. As comprehensive systems reduce the homogeneity of classrooms in terms of ability, these findings are at odds with the notion that there is a monotonously positive relation between classroom homogeneity and performance.

How large are these estimates? The point estimate of 0.17 SD in maths is roughly half the female-male achievement gap in maths (0.35), roughly one-third of the migration-native gap (0.50) and roughly one-fifth of the gap between children of parents from the lowest and the highest education category (0.94). The point estimate of 0.24 SD in reading is roughly double the male-female achievement gap in reading (0.11), roughly half the migration-native gap (0.42) and roughly one-fourth of the parental education gap (0.92).[41] Note that the effect sizes are measured in non-academic-track standard deviations. They are marginally smaller

[41] These figures refer to NEPS grade 5 test score gaps for non-academic-track students.

when measured in terms of the overall student population at 0.15 SD in maths and 0.21 in reading. Still, they are larger than the zero effect found by Hanushek and Woessmann (2006) at the age of 15. However, most of the tracked countries in their sample start tracking students at much later ages than considered here, when effects can generally be expected to be smaller. Importantly, they find that comprehensive schooling decreases the dispersion of test scores, indicating that weaker students benefit, which can reconcile these findings. Somewhat similarly, Kerr *et al.* (2013) find very small average effects of a Finnish comprehensive schooling reform and larger positive effects for disadvantaged students. My estimates are similar in magnitude to the one found by Garlick (2018) on South-African college students' GPA who are either (ability) tracked or randomly assigned to student dormitories. Also those are driven by low-achievers. Hence, in the following I will explore the heterogeneity of these average effects.

2.4.2 Effect Heterogeneity

As a first step to go beyond average effects, I extend the logic of the DD estimator and, instead of limiting attention to the mean, inspect how the whole achievement distribution changes differently between Comprehensive and Tracked states from grade 5 to 7.[42] Let $f_g^C(\cdot)$ be the density of non-academic-track students' grade g test scores for the Comprehensive states. The difference $f_7^C(y) - f_5^C(y)$ measures the change in the density at level $Y_{isg} = y$ between grades 5 and 7 for this group. $f_7^T(y) - f_5^T(y)$ is the equivalent change for the Tracked states. Comparing these two quantities across the support of the test score distribution allows me to map out the distributional consequences of comprehensive schooling:

$$\{f_7^C(y) - f_5^C(y)\} - \{f_7^T(y) - f_5^T(y)\} \tag{2.3}$$

Figure 2.8 presents the results from this exercise. Panels A and B plot density estimates for grade 5 and grade 7 maths scores respectively for Comprehensive and Tracked states. In the former the distribution appears to tighten slightly between grades, whereas in the latter it stays relatively constant. As the differences between the densities are small relative to the scale, panel C plots the vertical distances between the grade 5 and 7 densities by state group. Thus, these lines describe how the shape of the test score distributions changes between grades. Finally, panel D plots the vertical distance between these two lines, corresponding to the expression in equation (2.3). It appears that Comprehensive schooling shifts probability mass from the bottom end of the distribution (approximately from the range [-2.5, -0.5]) to the middle part (approximately to the range [-0.5, 1.5]). The picture for reading scores is very similar (see Appendix Figure B2.1). This means that, next to a positive average effect, comprehensive schooling has an equalising effect on the achievement distribution.

Figure 2.4 revealed that a state's tracking regime strongly affects peer group composition—but differently for students at different positions in the previous achievement distribution: in

[42] Neumark *et al.* (2004) proposed this method to estimate the effect of minimum wages on the distribution of family income.

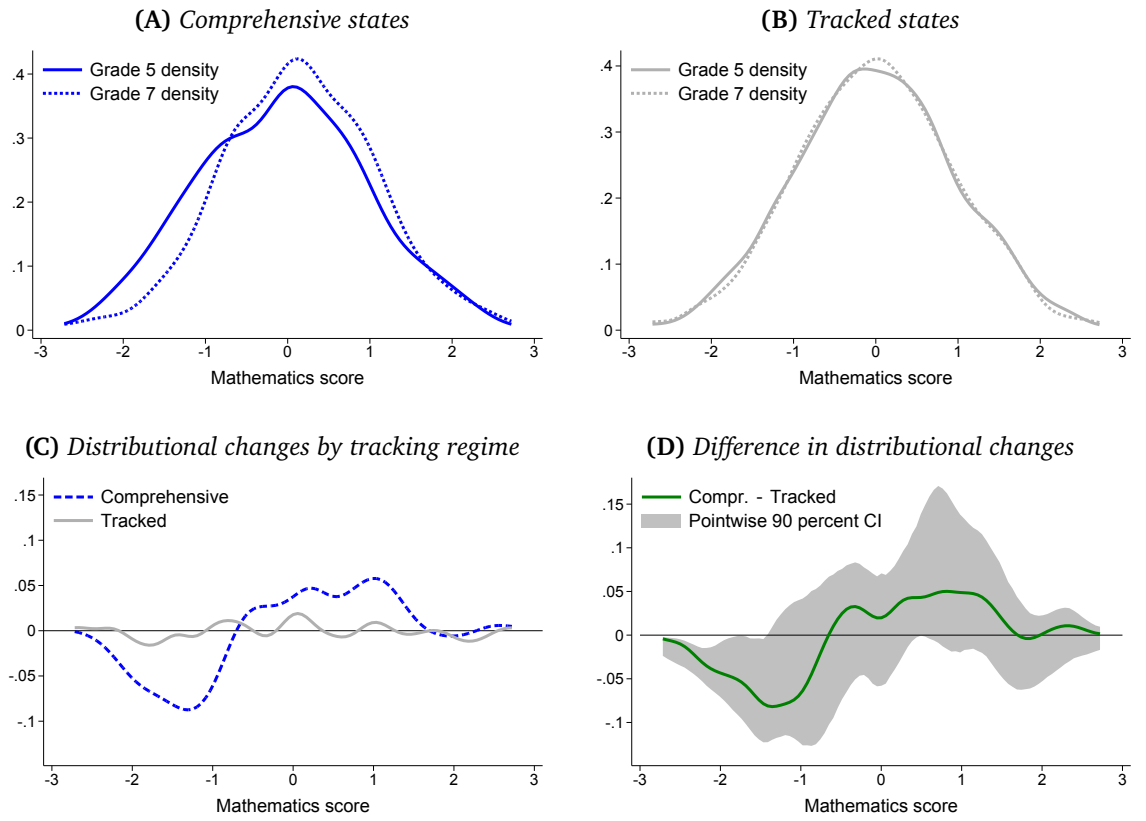


Figure 2.8. Maths score distributions before and after treatment exposure.

Notes: This figure describes how the test score distribution in maths changes differently between fifth and seventh grade depending on the tracking regime. Panels A and B, respectively for Comprehensive and Tracked states, display kernel density estimates for non-academic-track students' grade 5 and grade 7 scores at 100 equally spaced points between the 0.5th and 99.5th percentiles of the (cross-grade) maths score distribution. Estimation is based on the NEPS DD sample and a Gaussian kernel with optimal bandwidth for normally distributed variables (Silverman, 1986). Panel C plots the between-grade differences in the estimated densities at each point, separately for Comprehensive and Tracked states. Panel D plots the difference between Comprehensive and Tracked states in the between-grade density differences, including pointwise 90% confidence intervals from a percentile bootstrap with 999 iterations, clustering at the state level and stratifying by tracking regime. The bandwidth of the kernel estimator is held constant at its optimal level for the original sample in each bootstrap iteration (Hall and Kang, 2001).

the Tracked states, high-achieving students study together with higher achieving peers and low-achieving students study together lower achieving peers, whereas in the Comprehensive states a student's peer group depends much less on her own performance. If peer quality matters, the effect of comprehensive schooling is therefore likely to vary with students' previous achievement. In fact, it might well be that certain groups of students lose out from being taught comprehensively, but that these losses are compensated by the gains of other groups, resulting in a positive net effect.

To explore this possibility I use the NEPS 5-to-7 panel sample of non-academic-track students, which allows me to match students on previous achievement. Note that in the panel sample the average effects are marginally smaller: a simple 'value-added model' that regresses grade-5-to-7 gain scores on grade 5 scores and an indicator for the Comprehensive states gives

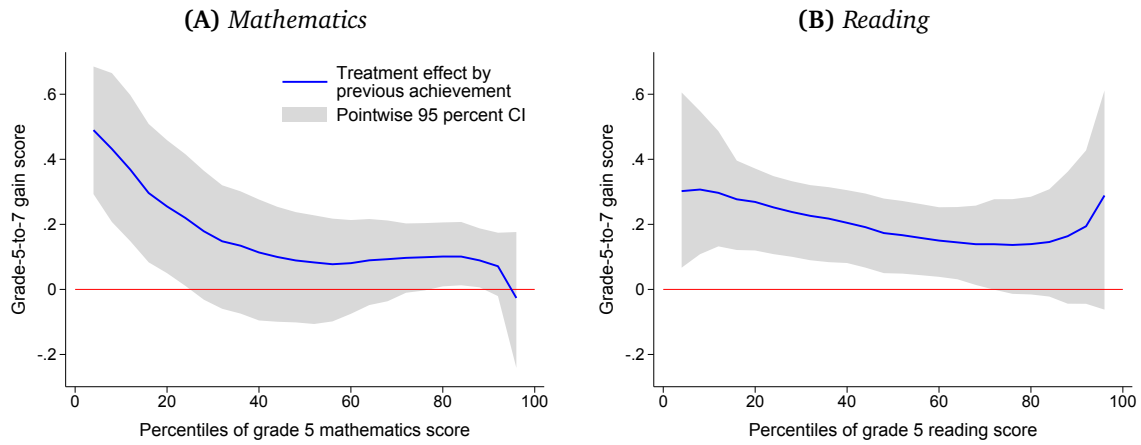


Figure 2.9. Effect heterogeneity by previous achievement.

Notes: This figure shows the difference in average fifth to seventh grade achievement growth between Comprehensive and Tracked states across the grade 5 test score distribution in maths (Panel A) and reading (Panel B). Estimation is based on the non-academic-track NEPS 5-to-7 panel sample ($N = 1,646$). The curves are constructed as follows: First, separately for Comprehensive and Tracked states, I estimate a student-level local constant regression of students' grade-5-to-7 gain scores on grade 5 test scores. Second, I calculate the difference between Comprehensive and Tracked states' fitted values at every fourth percentile of the grade 5 test score distribution. Third, I construct pointwise 95% confidence intervals from a percentile bootstrap with 999 iterations, clustering at the state level, stratifying by tracking regime and holding the bandwidth constant across bootstrap iteration (Hall and Kang, 2001).

effect estimates of 0.15 SD for maths ($p = 0.04$) and 0.23 SD for reading ($p < 0.01$).[43] If it is low-achieving students who benefit from comprehensive schooling, then the slightly smaller estimates for the average effect might be explained by the fact that low-achievers are more likely to drop out between waves and thus are slightly under-represented in the panel sample. To assess effect heterogeneity by previous achievement explicitly, analogously to above, I estimate the two conditional expectations $\mathbb{E}[\Delta_7 Y_{is} | \, Y_{is5}, \, Compr_s = k]$, for $k \in \{0, 1\}$, non-parametrically and evaluate their difference at different percentiles of Y_{is5}. This identifies the effects of comprehensive schooling throughout the pre-tracking achievement distribution.

The results for maths in panel A of Figure 2.9 reveal that the effect exhibits a steep gradient with respect to previous achievement: effects appear to be monotonically decreasing in grade 5 test scores in the first half of the distribution before flattening out, with large and significant effects from 0.5 to 0.2 SD in the bottom quartile, smaller and insignificant effects from 0.2 to 0 SD in the second quartile and roughly zero effects for all remaining students. In the results for reading in panel B the gradient by previous achievement is also visible but less pronounced: effects are significant and positive from the first through the 65th percentile of the grade 5

[43] The DD model estimated on the NEPS 5-to-7 panel sample gives virtually identical results to the value-added (VA) model. Appendix Table B2.4 presents both the VA model (in column 6) and the DD model (in column 5). The latter is labelled 'first-differenced' (FD) model because with individual-level panel data the DD model can be rewritten in FD form: $\Delta_7 Y_{is} = Y_{is7} - Y_{is5} = \delta_2 + \beta_{DD} \, Compr_s + \Delta u_{is}$. Note that the DD/FD model and the VA model are non-nested: The former controls for grade-constant heterogeneity that correlates with both treatment and outcomes, whereas the latter controls for selection into treatment based on (previous) outcomes (Angrist and Pischke, 2009). As my goal is to control for unobserved differences *between states*, instead of controlling for selection at the *individual level* the former seems more appropriate for the context at hand and is used in the main models. The VA model is presented here to show the robustness of the results to this alternative modelling of the selection process and to provide an average effect benchmark for the VA-style heterogeneity analysis below.

distribution, monotonically decreasing from about 0.3 to 0.1 SD, and larger but very imprecisely estimated and insignificantly different from zero thereafter.

These results imply that it is low achievers—and, to the extent that grade 5 achievement measures ability, low-ability students—who drive the positive level effects found before. They seem to benefit immensely from studying together with their higher achieving peers in a more demanding scholastic environment for another two years, especially in maths. Importantly, I do not find a negative effect at any point of the achievement distribution, meaning that higher achievers do not seem to lose out from learning together with their lower achieving peers. Remember, while non-academic-track students are a negatively selected group with substantially lower test scores than academic-track students on average, Figure 2.3 shows that the distributions of these groups overlap substantially. The top 25% non-academic-track students would be above-median students even in the academic track.

To investigate effect heterogeneity along other dimensions, in Appendix Table B2.2 I present results from fully interacting the DD model with indicators for female, low socio-economic status (SES) and migration background students. All treatment-covariate interactions are insignificant and without clear directional pattern across maths and reading scores but, clearly, the analysis is underpowered to detect smaller effect heterogeneities. Regardless, the striking pattern found above suggests that previous achievement is the most important dimension for effect heterogeneity in the current context.[44] Several studies found that especially low-SES students benefit from comprehensive schooling (e.g., Kerr *et al.*, 2013) but in the selected group of non-academic-track students investigated here SES differences are not very pronounced to begin with. For instance, in my sample, only 22% of students with college-educated parents even attend a non-academic-track school. There is a much more salient socio-economic divide between academic- and non-academic tracks than between different school types in the non-academic segment. Accordingly, the first-order effect of the treatment of comprehensive schooling in my setting is the mingling of students of different abilities rather than of different socio-economic backgrounds.

2.4.3 Effect Persistence

In this section I present estimation results for ninth-grade outcomes—the grade level after which students can leave school with a low-track degree (conditional on obtaining the required grades). Note that interpretation of these results is complicated by the fact that, from seventh grade onwards, non-academic-track school in the Comprehensive states may sort students by ability but there is no reliable information on the incidence and exact implementation of this within-school streaming. More generally, the harmonisation of schooling policies between states decreases with grade level (Kultusministerkonferenz, 2014). Accordingly, estimates represent a mixture of effect persistence and effects from continued (but somewhat unclear) differences

[44] This conclusion is corroborated by the fact that some sizeable (but insignificant) interaction effects for reading decrease in magnitude when repeating this exercise in the value-added model that controls for previous achievement (see Appendix Table B2.2).

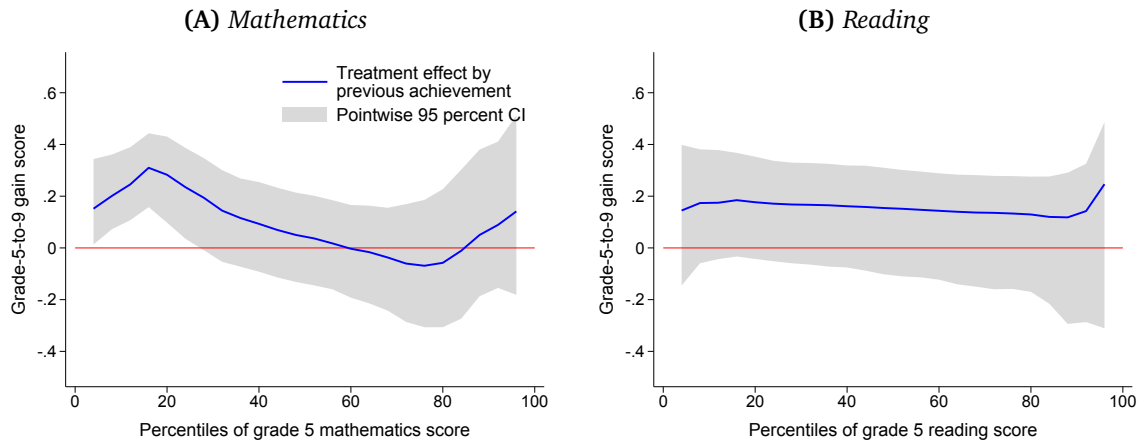


Figure 2.10. Effect persistence by previous achievement.

Notes: This figure0 shows the difference in average fifth to ninth grade achievement growth between Comprehensive and Tracked states across the grade 5 test score distribution in maths (Panel A) and reading (Panel B). The curves are based on the non-academic-track NEPS 5-to-9 panel sample comprising all students for whom I observe test scores in both grades ($N = 1,286$ for maths and $N = 1,255$ for reading). They are constructed analogously to those in Figure 2.9.

in tracking and other schooling inputs. With these caveats in mind, the purpose of this section is two-fold: to obtain a rough idea of effect persistence and to see whether the patterns found until now replicate in the IQB data.

First, I repeat the above analysis for ninth-grade test scores in the NEPS data. While the DD estimates in Appendix Table B2.5 continue to show an advantage for students taught comprehensively in grades 5 and 6, they are smaller than before and far from significant as smaller samples and increased interference from other between-state differences seem to take their toll on precision.[45] As average effects might mask persistence for low-achieving students, Figure 2.10 repeats the heterogeneity analysis from above for grade-5-to-9 gain scores using the NEPS 5-to-9 panel sample. Indeed, in maths effects for the lowest quartile of students are both economically and statistically significant, ranging from 0.15 to 0.3 SD, thus indicating persistent benefits from deferring between-school tracking for low-achievers. In reading, point estimates are positive but insignificantly different from zero across the previous achievement distribution.

Second, I re-estimate the DD model using the larger IQB DD sample.[46] Note that I can only report results for reading and listening, as maths was not tested in ninth grade. Table 2.3 presents the estimation results.[47] Column 1 presents the unsaturated DD model, which estimates the raw

[45]The differences between grade 7 and grade 9 results are not driven by sample differences. Using the smaller grade 9 sample for the grade 7 regressions reproduces the previous results quite precisely.

[46]Summary statistics for the IQB DD sample are presented in Appendix Table B2.6.

[47]The IQB results pool across 15 so-called 'plausible values' (PVs): Students answer different subsets of the total pool of IQB assessment questions ('multi-matrix design'). In order to deal with the missing information on questions outside their subset, each student is imputed 15 PVs per test score domain. Standard practice is to run regressions for each and combine point estimates and standard errors according to the rules in Rubin (1987). These state that the variance of a statistic based on m imputations is the sum of the average within-imputation variance and the between-imputation variance: $Var^{total} = m^{-1} \sum_m Var_m^{within} + (1+m^{-1}) Var^{between}$. Problematically, the wild cluster bootstrap does not produce within-variance estimates (i.e., standard errors), but only a distribution of t-statistics from which p-values are computed. Instead of reverting to standard clustered standard errors (which are likely to underestimate the within-imputation variance due to the few-cluster problem) to use Rubin's rule, I decided to ignore the between-imputation variance and simply pool the wild cluster bootstrapped p-values across imputations

Table 2.3. DD regressions for ninth-grade achievement in the IQB DD sample.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Reading							
Comprehensive schooling	0.136 (p = 0.18)	0.155 (0.12)	0.130 (0.30)	0.175 (0.19)	0.066 (0.37)	0.108 (0.35)	0.026 (0.16)
× Female				-0.036 (0.63)			
× Low SES					0.143* (0.10)		
× Migration background						-0.010 (0.68)	
Compr. state indicator	-0.018 (0.90)						
Classroom peers' mean score							0.844*** (0.00)
R^2	0.001	0.021	0.164	0.036	0.069	0.059	0.207
Panel B: Listening							
Comprehensive schooling	0.146* (0.06)	0.155** (0.04)	0.152** (0.04)	0.134 (0.15)	0.088 (0.15)	0.111* (0.09)	0.019 (0.18)
× Female				0.045 (0.53)			
× Low SES					0.101 (0.37)		
× Migration background						0.010 (0.71)	
Compr. state indicator	-0.088 (0.32)						
Classroom peers' mean score							0.887*** (0.00)
R^2	0.001	0.017	0.199	0.023	0.078	0.098	0.260
(Interacted) controls			✓				
State FE		✓	✓	✓	✓	✓	✓
N state clusters	12	12	12	12	12	12	12
N Compr. state students	13526	13526	13526	13526	13526	11854	13526
N Tracked state students	11276	11276	11276	11276	11276	10186	11276

Notes: This table reports OLS regression results for the DD model for grade 4 and 9 reading (Panel A) and listening (Panel B) test scores of non-academic-track students using the IQB DD sample. Column 1 report results for the unsaturated DD model. Column 2 reports results for the saturated DD model, which replaces the Comprehensive state indicator with state fixed effects. Column 3 adds the following student and school covariates, incl. their interaction with the grade 9 indicator: sex, age, age squared, migration background, foreign language at home, highest parental level of education (HISCED), highest parental occupational status (HISEI), teacher experience, teacher further training, no. days all-day schedule/week, private school, homework support and extracurricular learning offers. Columns 4–6 fully interact the saturated DD model without controls with indicators for female, below-median socio-economic status (SES) and migration background students, respectively. The SES score is the first principal component of the following variables and their missing dummies: student-reported number of books at home, highest parental level of education, highest parental occupational status. Column 7 presents results for the saturated DD model controlling for classroom peers' mean test scores in the respective subject. p-values in parentheses stem from 999 wild cluster bootstrap iterations, clustering at the state level and using Webb weights. Estimations apply student sampling weights and results are pooled across the 15 plausible values per test score domain for each student (see footnote 47 for details). Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

(pooling means to convert them into t-values, average those across imputations and convert back into a p-value). Differences between PVs are so small that ignoring the between-imputation variance is innocuous in this context. A back-of-the-envelope calculation supports this choice: Appendix Table B2.7 shows the effect estimates in the saturated DD model for each PV/imputation, from which I calculate the between-imputation variance. For each imputation, I then (under-)estimate the within-imputation variance using a standard cluster-robust variance estimator. The between-variance is only 2.8% of the *underestimated* within-variance for reading and 7.4% for listening. So, applying Rubin's formula, in the case of listening, one would need to scale the within-variance by 1.08 to get the total variance, which would reduce the t-statistic by a factor of $1/\sqrt{1.08} = 0.96$. For reading this factor is even closer to one. Hence, even in this overly conservative calculation (due to underestimated within-variances), ignoring the between-variance

double difference between Comprehensive and Tracked states between grades 4 and 9. In line with the NEPS data, I find no significant achievement differences at the end of primary school but an advantage for non-academic-track students from the Comprehensive states in secondary school of 0.14 SD in reading and 0.15 SD in listening. This result is robust to including state fixed effects (column 2) and flexibly controlling for student and school characteristics (column 3), but only statistically significant for listening. In columns 4–6 I fully interact the DD model with indicators for female, low-SES and migration background students to test for effect heterogeneities along observable student characteristics in the larger IQB samples. Of those three, only the SES interaction seems to substantially reduce the main effect, reaching marginal significance in the case of reading. Given that low-SES students are more likely to be low achieving, this is at least qualitatively in line with the heterogeneity by previous achievement found above, which I cannot directly investigate without panel data. Instead, I repeat the distributional analysis, which confirms that comprehensive schooling shifts probability mass from the bottom to the middle of the test score distribution in both reading and listening (see Appendix Figures B2.2 and B2.3). In sum, the IQB results confirm those based on the NEPS data.

2.4.4 Mechanisms

A large part of the preceding analysis has been devoted to understanding whether the estimated effects are indeed due to between-state differences in tracking. However, even if I can rule out confounding from school resources and student body composition, it is unclear what are the precise mechanisms underlying my results. The effect of between-school tracking on student achievement might operate through various channels.

First, and most problematically for my purposes, the effects might be driven by logistical implications of running a two- *versus* a three-tiered school system: in the former states need to maintain two distinct school types and in the latter three. This may impact local school supply, i.e., the size of schools and students' travelling time to school (and thereby time left for homework and other educational investments). For the sake of generalisability, I would like to rule out these channels and isolate the portion of the effect that is solely due to the sorting of students by ability between schools. Therefore, school size has been included in the list of school controls in the main regressions. The results in Table 2.2 and Appendix Table B2.3 suggest that it can be ruled out as a relevant channel. Due to lack of data I cannot directly control for students' travelling times. However, in fifth and seventh grade the NEPS questionnaires asked students to report their weekly time spent on homework, allowing me to investigate students' educational time investments directly. Column 1 of Table 2.4 presents results for the DD model applied to time spent on homework and shows that non-academic-track students from the Comprehensive state spent *less* time on homework and that this difference is constant across grades. As this is the group experiencing higher achievement gains between grades, time constraints are unlikely to play an important role in this context.

is close to irrelevant for the coefficients' significance. I abstain from further assumptions to pool the bootstrapped confidence sets and only report averaged coefficients and pooled p-values.

Second, the results might be driven by incentives to exert effort and invest for students and their parents. In contrast to the Tracked states, in the Comprehensive states students are not 'locked' into (low and intermediate) tracks in the first two years of secondary school. This might give students the impression that they need to work hard continually to reach their aspired degree; especially since some non-academic-track schools sort students into low- and intermediate-track classes starting in grade 7. The same applies to parents, who might thus be incentivised to invest more in their children's education during these two years. However, these conjectures do not seem to square with the evidence presented in Table 2.4: as mentioned before, column 1 shows that students in the Comprehensive states spend less time on homework and, on top of that, columns 2 and 3 show that they are neither more likely to receive help from their parents with with school work, nor to receive private tutoring.[48]

Third, the effects might operate through the taught curriculum: for students who would be assigned the low track in the three-tiered system, comprehensive schooling is likely to increase academic standards, whereas for students who would be assigned the intermediate-track standards might decrease. *A priori* it is unclear how this might affect achievement, as low-achievers could lose out from being held to excessive academic standards or benefit if they grow with the demands. Regardless, given that curricular differences between low- and intermediate-track schools are relatively small in the first two grades of secondary school (Bald, 2011), this is unlikely to be the primary driver behind the results. Of course, the ability composition of the class might influence in what detail the teacher treats the material, but such *peer effects* should not be confused with curricular effects.

Fourth, by attending either a low- or a intermediate-track school, students are labelled and explicitly ranked in the Tracked states, whereas in the Comprehensive states they are not (save for being below the academic track). This social comparison might negatively affect their academic self-concept, educational aspirations, motivation to study and, in turn, achievement (Dumont *et al.*, 2017). Contrary to this, Murphy and Weinhardt (2020) show that classroom rank, which is likely to be higher in lower tracks, positively affects student achievement. However, their study concerns the non-tracked English school system where schools are not explicitly ranked. In the German tracking system, between-school sorting is salient and students are well aware of their track's rank, reducing the significance of favourable within-class comparisons in low-track schools. In line with this conjecture, Dumont *et al.* (2017) find that school-leaving certificates, which correspond to (but are not determined by) tracks, are the primary determinant of students' academic self-concept in German non-academic-track schools.

To investigate such socio-emotional channels the remaining columns of Table 2.4 present results for students' educational aspirations, their school-related motivation and feelings of helplessness in school. Only aspirations were measured in grades 5 and 7, allowing for implementation of the DD design. I construct two indicator variables indicating that students aspire higher than the low- or intermediate-track degree, respectively. The results in columns 4 and 5 show

[48]The IQB surveys included questions about private tutoring, too. In Appendix Table B2.8 I show that this result reproduces in the IQB data: if anything, there is more private tutoring in the Tracked states.

Table 2.4. Effects on seventh-grade behavioural and socio-emotional outcomes.

Model specification:	Double Differences					Cross-sectional OLS			
Dependent variable:	Time spent on homework	Help from parents	Private tutoring	Aspirations		Helplessness		Motivation	
				> low cert.	> mid cert.	Maths	German	Maths	German
Variable type:	*(scale)*	*(scale)*	*(dummy)*	*(dummy)*	*(dummy)*	*(scale)*	*(scale)*	*(scale)*	*(scale)*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Comprehensive schooling	0.054	0.048	-0.048	0.120*	0.003	-0.165*	-0.127	0.140	0.178
	(p = 0.49)	(0.63)	(0.61)	(0.07)	(0.97)	(0.08)	(0.17)	(0.41)	(0.33)
	[-0.10, 0.35]	[-0.14, 0.34]	[-0.15, 0.10]	[-0.03, 0.30]	[-0.16, 0.13]	[-0.35, 0.02]	[-0.25, 0.06]	[-0.13, 0.31]	[-0.19, 0.39]
Grade 7	-0.019	-0.203	-0.081	-0.076	0.072				
	(0.91)	(0.31)	(0.34)	(0.60)	(0.44)				
	[-0.37, 0.47]	[-0.60, 0.19]	[-0.30, 0.09]	[-0.31, 0.17]	[-0.19, 0.26]				
Indicator Compr. states	-0.242**	-0.018	-0.023	-0.041	0.095*				
	(0.03)	(0.78)	(0.28)	(0.35)	(0.08)				
	[-0.45, -0.07]	[-0.24, 0.17]	[-0.09, 0.04]	[-0.17, 0.06]	[-0.01, 0.26]				
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
R^2	0.036	0.047	0.036	0.080	0.094	0.031	0.050	0.067	0.061
N state clusters	12	12	12	12	12	12	12	12	12
N Compr. state students	834	819	471	799	799	503	499	486	487
N Tracked state students	3940	3870	2407	3869	3869	2010	2025	1850	1832

Notes: Columns 1–5 present OLS regression results for the unsaturated DD model with student and school covariates applied to different dependent variables in the NEPS DD sample of non-academic-track students, each time retaining all observations with non-missing values for the respective variable. 'Time spent on homework' is the student-reported average time spent on homework per week, standardised to mean zero and standard deviation one for Tracked states students within each grade level. 'Help from parents' is the student-reported frequency of help received with homework from their parents, standardised as before. 'Private tutoring' is a indicator variable equal to one if parents report that their child receives private tutoring. Educational aspirations are measured at two margins: in column 4 (column 5) the dependent variable is an indicator equal to one when students report aspiring higher than the low-track (intermediate-track) certificate. Columns 6–9 are based on the grade 7 cross-section only, as these outcomes were not measured in fifth grade. The dependent variables are regressed on the full set of controls and an indicator for the Comprehensive states. Helplessness is an index, standardised as before, with higher values indicating a higher degree of feeling helpless in the respective school subject. The variable averages 5 survey items, each measured on a 4-point Likert scale. Motivation is an index, standardised as before, with higher values indicating a higher intrinsic motivation for the respective school subject. The variable averages 4 survey items, each measured on a 4-point Likert scale. p-values in parentheses and 95%-confidence sets in brackets stem from 999 wild cluster bootstrap iterations, clustering at the state level and using Webb weights. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

that comprehensive schooling reduces the share of students with low educational aspirations, while there is no effect at the higher margin. Increased aspirations at the bottom seem to mirror the large benefits for low-achieving students found above. The remaining variables were only measured in grade 7, so that I am forced to revert to OLS regressions with large control sets to approximate the effect of comprehensive schooling. With this caveat in mind and despite their limited significance, the results in columns 6–9 suggest that students taught comprehensively are less helpless and more motivated. Appendix Table B2.8 shows very similar patterns in the IQB data; in particular, I find strong evidence for positive effects of comprehensive schooling on motivational outcomes. Although far from conclusive, these results suggest that socio-emotional effects of tracking are relevant.

The fifth and most palpable mechanism for the effects of any tracking policy is certainly peer effects—mind you that the stated goal of tracking is to homogenise classroom peers in terms of ability. According to tracking proponents, this should benefit all students by allowing for more tailored teaching (Duflo *et al.*, 2011). Tracking opponents argue that, instead of homogeneity, peers' ability level is what really matters: more able peers generate direct knowledge spill-overs, increase the quality of classroom interactions (including with teachers) and serve as role models. Numerous papers show positive effects of mean peer achievement on student achievement (e.g., Sacerdote, 2001; Whitmore, 2005; Ammermüller and Pischke, 2009; Carrell *et al.*, 2009; Lavy *et al.*, 2012; Burke and Sass, 2013; Garlick, 2018). A growing literature shows non-linearities in the effects of peers—in particular, very low-achieving peers seem to generate negative spill-overs (e.g., Figlio, 2007; Carrell *et al.*, 2018; Lavy *et al.*, 2012; Bietenbeck, 2020). In a similar vein, Bursztyn and Jensen (2015) present experimental evidence for the presence of peer pressures penalising effort in low-track (non-honours) classes that are absent in high-track (honours) classes. The costs of exposure to low-achieving peer environments might thus well be larger than the benefits of exposure to high-achieving peer environments.

Figure 2.4 showed that peer group composition differs significantly between the two tracking regimes: low-achievers attend lower achieving classrooms on average and, consequently, have a higher probability of being exposed to the lowest-achieving individuals in the Tracked states. Moreover, anecdotes about low-track schools with negative peer dynamics that discourage learning are common. Peer effects are thus a likely candidate to explain the large gains from comprehensive schooling for lower achieving students. To test their role directly I perform a simple mediation analysis. Column 7 of Table 2.3 presents results from the DD model in the IQB data controlling for the mean of classroom peers' test scores.[49] In both reading and listening the effect of comprehensive schooling disappears when controlling for mean peer achievement, indicating the importance of peer effects. The found pattern of effect heterogeneity by previous achievement implies that peer effects are heterogeneous: low-achievers seem to be more sensitive to peer group composition than high-achievers—a result also found by Garlick (2018).

[49] I am prevented from repeating the mediation analysis in the NEPS data because I do not observe students primary school (pre-tracking) classrooms. Note that the IQB data is better suited for the analysis of peer effects anyway, first, because of the large number of observed classes and, second, because participation in the IQB tests is mandatory such that *whole classes* are observed. Accordingly, mean peer achievement is measured much more accurately in the IQB data than in the NEPS.

Altogether, my results confirm the importance of peer effects in explaining the effects of tracking and, inversely, suggest that the homogeneity of classrooms at such an early age might be less important than commonly assumed.

2.5 Discussion and Conclusions

This paper set out to estimate the effect of early between-school tracking in secondary school on student achievement—an issue that, despite its enduring prevalence in educational policy debates, is still not fully understood. Theoretically, the question of tracked *versus* comprehensive schooling seems to involve a trade-off between countervailing forces. On the one hand, homogeneous learning environments are likely to facilitate skill and knowledge acquisition as content and teaching style can be more closely tailored to median classroom ability. On the other hand, the concentration of high ability students in certain schools might impair competence development of students in lower tracks through negative motivational consequences and peer effects. Identifying these effects is notoriously difficult due to the severity of the selection problems involved.

My identification strategy exploits differences in tracking between German federal states: in all states, about 40% of students transition to the academic track after comprehensive primary school. Depending on the state, the remaining student body is either divided between low- and intermediate-track schools or taught comprehensively for another two years. I estimate the effects of these two years of comprehensive instead of tracked schooling on achievement in a triple-differences framework. The estimator compares achievement growth of comprehensively taught non-academic-track students with that of tracked ones, while controlling for tracking-regime-specific trends using unaffected students in the academic track.

I find that student achievement increases when non-academic-track students are not ability tracked between schools but taught comprehensively for another two years: the 95% confidence set for the effect on seventh-grade test scores is estimated to be [0.09, 0.28] SD in maths and [0.04, 0.48] SD in reading. These somewhat imprecisely estimated level differences are composed of large positive effects for low-achievers and null effects for high-achievers. Consequently, comprehensive schooling has an equalising effect on the distribution of test scores without trading off efficiency against equity. There is some fade-out in the effects but comprehensively taught students are still better off towards the end of lower secondary schooling. Auxiliary analyses suggest that students' school-related motivation and educational aspirations are higher in the comprehensive system and that peer effects play an important role in explaining the effects.

With respect to Germany, my results confirm the reform efforts of several West German states to abolish low-track schools and replace their three- with two-tiered school systems. In line with policy-makers intentions, this appears to generate better and more equitable outcomes. Beyond the German context, the effects in this paper are immediately relevant for other countries with multi-track between-school tracking systems, such as the Czech Republic, the Netherlands and Slovakia.

With respect to countries with two-tiered tracking systems, caution must be exercised when extrapolating from my results to the effects of turning those into fully comprehensive school systems. This is because the variation in tracking practices I exploit concerns only the (negatively) selected group of non-academic-track students. Accordingly, my results might not translate to students in the academic track. However, note that the central dimension of effect heterogeneity is previous achievement, which overlaps considerably between tracks. Even for the top quartile of non-academic-track students, who would be medium-high achievers also in the academic track, I find no evidence for negative effects from comprehensive schooling (with *positive* point estimates). This suggests that, if there are negative effects at the young ages considered here, these are confined to the very top of students.

Overall, my results provide a cautionary tale about early and rigid forms of vertical differentiation in schools applicable to all between-school tracking settings. They show that there are limits to efficiency gains from classroom homogeneity as other mechanisms, such as peer effects, motivation and aspirations, start to depress achievement at the bottom once a selective system becomes too differentiated. Accordingly, policy-makers need to carefully balance these forces when determining the degree of vertical differentiation in their school systems and the age at which it starts.

Finally, note that many papers on *within-school* streaming report positive effects for students selected for high-ability classrooms without negative effects for those in regular classrooms (e.g., Card and Giuliano, 2016; Duflo *et al.*, 2011; Figlio and Page, 2002). Rather than contradicting my and previous findings on between-school tracking, this suggests that costs of ability grouping increase with the degree of vertical differentiation between tracks. It makes intuitive sense that mechanisms relating to peer effects, motivational factors and educational aspirations are more pronounced when students are separated between schools. Consequently, forming (subject-specific) *classrooms* based on ability from a certain age onwards, but eschewing vertical differentiation between *schools* to avoid creating detrimental learning environments for low-track students, might allow reaping efficiency gains from homogeneity without incurring large costs in terms of equity.

Appendix A: Additional Information on the Data

The purpose of this appendix is to give a detailed description of the data sets and samples used throughout the paper. This discussion will be guided by Appendix Figure A2.1, which gives a schematic overview of how the different samples correspond to student cohorts and grade levels. The horizontal axis represents school years, divided into first and second term to show whether students were surveyed at the beginning or end of a school year. The vertical axis represents grade levels: the first four grades correspond to primary school, after which students transition to secondary school. Secondary school finishes after grade 9 in the low track, grade 10 in the intermediate track and grade 12 or 13, depending on the state, in the academic track. Cells containing survey names indicate the timing of testing/surveying. The shading shows the progression of sampled cohorts through grade levels within the German school system.

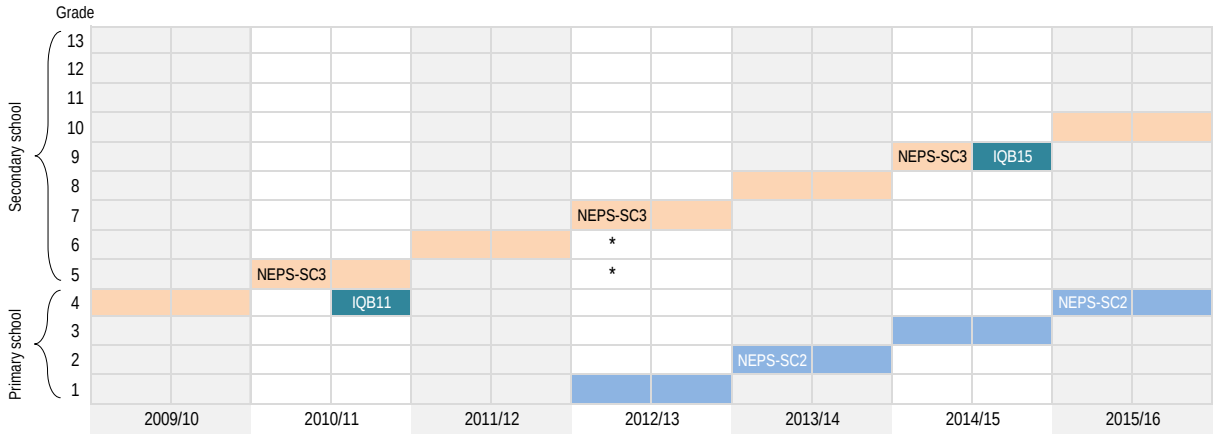


Figure A2.1. Overview of data sets.

2.5.1 National Educational Panel Study

The German National Educational Panel Study (NEPS) is a study carried out by the Leibniz Institute for Educational Trajectories at the University of Bamberg. The NEPS collects longitudinal data on the development of competencies, educational processes, educational decisions, and returns to education for six different 'starting cohorts' (SC): newborns (SC1), kindergarten/primary school students (SC2), lower secondary school students (SC3), upper secondary school students (SC4), university students (SC5) and adults (SC6). For details on the project see Blossfeld *et al.* (2011).

2.5.1.1 NEPS-SC3

The main dataset of this paper is the lower secondary NEPS Starting Cohort 3 (NEPS-SC3), a random sample of newly minted fifth-graders in the school year 2010/11. Students were sampled according to a multi-stage process: (1) Random sampling (with probabilities proportional to

scale) from the population of all German schools at lower secondary level. (2) Random selection of two grade 5 classes within the selected schools. (3) All students of the selected classes were invited to participate in the study. Participating students were surveyed and tested for the first time in the autumn of 2010, at the beginning of their first year in secondary school. Counting only students in regular schools, attending fifth grade for the first time in 2010/11, with non-missing test scores in mathematics and reading, the total size of the 'NEPS grade 5 cross-section' is 4,448, of which 2,303 are non-academic-track students, of whom 330 are from the Comprehensive states.

I use information from student and parent questionnaires to construct student-level control variables: age, a binary indicator for sex, a binary indicator for migration background, a binary indicator for single parent household, a binary indicator for foreign language spoken at home, an index for home possessions (from the student questionnaire), highest level of parental education measured in four categories, monthly household income, a binary indicator for receipt of unemployment benefits (from the parent questionnaire). While the student questionnaire variables are observed for almost everyone (< 1% missing values), parents' answers are missing for about 30% of the sample.

I use information from the school principal questionnaire to construct the following proxies for school quality/schooling inputs: average teacher age, student-teacher ratio, average class size, private school and four standardised indices for a school's facilities[50], extracurricular programmes[51], educational support offers[52] and quality control measures[53]. One additional school-level covariate is retrieved from the teacher questionnaire (but averaged by school to reduce the number of missing values): days of further training received over the past year. School covariates are missing for about 10% of the sample.

The NEPS being a panel, the same students were tested again two years later, in the autumn of 2012, when they had just started seventh grade according to schedule. Students that repeated a grade but remained in the same school are included in the testing, which in Figure A2.1 is indicated by two asterisks at grade levels 5 and 6 (note that to still be in fifth grade students would have to had repeated twice, a case that is not actually observed in the data). Students that switched school are not part of the grade 7 sample, as testing was tied to students remaining in their initial school context. All analyses that use grade-5-to-7 gain-scores as outcomes are based

[50]The index sums the following binary items about the school's facilities: presence of gym; swimming pool; language laboratory; auditorium; common rooms; individual work stations; student library; teacher library.

[51]The index sums the following binary items about the schools afternoon programme: extracurricular homework supervision; remedial teaching for students with non-German background; instruction for students with non-German background; courses in maths; science; German or literature; foreign languages; sports; music or arts; politics or philosophy; handicrafts; offers in technology or media; community activities; social learning; inter-cultural learning; required free-time activities; voluntary free-time activities; project days; project weeks; hot lunches; long-term projects.

[52]The index sums the following binary items about the individual educational support offered by the school: courses in learning techniques; participation in projects or competitions; homework coaching; tutoring; other forms of coaching.

[53]The index sums the following binary items about quality control: complete school mission statement; written school profile; written specification of quality indicators; written specification of performance standards; standardised performance testing; systematic appraisal of data; school brochure; harmonised exams across classrooms.

on the panel sample of students who have non-missing test scores in these first two waves of the NEPS-SC3 survey. This sample is referred to as the 'NEPS 5-to-7 panel sample' in the text.

There is substantial attrition in the NEPS-SC3 panel: of the 4,448 students tested in fifth grade 3,521 are tested again in 2012, of whom 1,646 are non-academic-track students, of whom 269 are from the Comprehensive states. This amounts to an overall panel attrition rate of 21%. In the non-academic tracks the attrition rate is 29%, compared to only 13% in the academic track, indicating that panel drop-out is negatively associated with achievement. Indeed, limiting attention to the non-academic-track sample, drop-outs have 0.06 SD lower maths scores ($p < 0.01$) and 0.04 SD lower reading scores ($p < 0.01$) than their peers (they are also 4 percentage points more likely to be migrants ($p = 0.06$) and 8 percentage points more likely to have low SES ($p < 0.01$)). Further, 49% of panel drop-outs in the Tracked states are from low-track schools, whereas only 35% of all non-academic-track students in the Tracked states belong to the low track.

The reasons for drop-out in the non-academic-track sample are schools withdrawing their participation in the NEPS study (36%), schools or classes being closed (8%) and students switching school (35%). The remaining 21% drop out for an unknown reason, i.e. either because of absence on the day of testing or because students or their parents withdrew their participation in the survey. Overall, attrition is higher in the Tracked states (30% compared to 18% in the Comprehensive states). However, excluding panel drop-out due to schools withdrawing their participation in the survey and schools closing (as these are due to administrative reasons at the school level, unlikely to be related to schooling policy at the state level and clearly not driven by self-selection at the student level), there are no significant differences in any of these shares between Comprehensive and Tracked states.

In addition to students part of the panel sample, the 2012/13 'NEPS grade 7 cross-section' is augmented with a large random refreshment sample of seventh-graders. The refreshment sample was drawn to counteract selective attrition: of the 1,795 additional students, a large majority of 1,125 are from non-academic-track schools (of whom 283 are from the Comprehensive states). Accordingly, the refreshment sample balances the higher rate of attrition in the non-academic tracks and ensures that the NEPS sample remains representative of the student population in both segments of the school system. Together with the 3,521 students from the panel sample, the NEPS grade 7 cross-section has, in total, 5,316 observations, of which 2,771 are in the non-academic tracks, of which 552 are from the Comprehensive states.

The main difference-in-differences (DD) and triple-differences (DDD) regressions presented in Table 2.2 pool the NEPS-SC3 grade 5 and grade 7 cross-sections. The DD model uses non-academic-track students only and, thus, relies on 2,303 fifth-grade and 2,771 seventh-grade student observations (for 5074 student×grade observations in total). This is referred to as the (grade 7) 'NEPS DD sample' in the text. The DDD model adds academic-track students students as an additional control group for an additional 2,145 fifth-grade and 2,545 seventh-grade observations (for 9,764 student×grade observations in total). This is referred to as the (grade 7) 'NEPS DDD sample' in the text. These sample sizes are summarized in Appendix Table A2.1.

Table A2.1. Sample sizes of NEPS cross-sections.

	Non-academic tracks		Academic track	
	Grade 5	Grade 7	Grade 5	Grade 7
Tracked states	1,973	2,219	1,797	2,064
Compr. states	330	552	348	481
	DD sample			
	DDD sample			

After the second wave, students were tested again two years later, in the school year 2014/15, when the cohort attended ninth grade according to schedule. Limiting attention to the non-academic tracks, of the initial panel sample there are 1,286 observations left in mathematics, of which 186 are from the Comprehensive states, and 1,255 in reading, of which 191 are from the Comprehensive states (testing in reading happened later in the year, explaining the difference in the number of observations). All analyses that use grade-5-to-9 gain-scores as outcomes are based on these 'NEPS 5-to-9 panel samples'.

The 'NEPS grade 9 cross-section' of non-academic-track students, which on top of the grade-5-to-9 panel sample includes students from the grade 7 refreshment sample still participating in the survey, comprises 2,149 student observations, of which 433 are from the Comprehensive states. Analogously to the seventh-grade DD regressions, for the ninth-grade DD model of Appendix Table B2.5 the NEPS grade 5 and grade 9 cross-sections are pooled.

2.5.1.2 NEPS-SC2

The NEPS Starting Cohort 2 (NEPS-SC2), a random sample of German primary school students, is used as an additional data source for two reasons: (i) to provide information on primary school inputs and (ii) to investigate achievement trends before tracking starts. The sampling design is very similar to that of NEPS-SC3, with schools as primary sampling units.[54]

A concern for the validity of the DD estimates is that school inputs might have changed differently between primary and secondary school between Tracked and Comprehensive states. Accordingly, it is important to probe their robustness to the inclusion of school input controls. However, only the secondary school environment is observed in the NEPS-SC3, which logically can only affect the post-treatment (grade 7) scores. For the pre-treatment (grade 5) scores, the relevant schooling inputs are those from primary school, which are missing because because I do not observe the primary school students came from. To impute the missing primary school/pre-treatment schooling inputs in the NEPS-SC3 DD sample, I use the SC2 principal questionnaires. In particular, I calculate state-level averages for all the above-mentioned school-level controls, using the earliest available principle questionnaire for each school to minimise the distance between my main cohorts primary school time and the time the primary school information is

[54]Note that the NEPS-SC2 panel commenced two years prior to primary school, when children were still in kindergarten. Still, primary schools served as primary sampling units. As the earlier waves of the panel are irrelevant for the purpose of this paper, they are ignored in this description.

recorded (in vast majority of cases this means 2012), and assign each grade 5 observation in the NEPS DD sample its state-level average.[55]

For the analyses of primary school achievement trends, I use the NEPS-SC2's student-level achievement data. As shown in Appendix Figure A2.1, the first measurement point used here is the beginning of the school year 2013/14, when the surveyed cohort has just entered second grade. Counting only students in regular schools with with non-missing test scores in mathematics, the total size of the 'NEPS grade 2 cross-section' is 5,384, of which 979 are from the Comprehensive states.

Students were tested again two years later, in the autumn of 2015, when they had just started fourth grade according to schedule. Those that repeated a grade but remained in the same school were included in the testing. Students that switched school are not part of the grade 4 sample. Analyses that use grade-2-to-4 gain-scores as outcomes are based on the panel sample of students who have non-missing test scores in these first two waves of the NEPS-SC2. This sample is referred to as the 'NEPS 2-to-4 panel sample' in the text. It comprises 4,676 observations in total, of which 849 are from the Comprehensive states. Hence, the attrition rate between second and fourth grade is 13% overall and in both state groups. Again, panel drop-out is negatively related to performance: drop-outs have 0.18 SD lower maths scores ($p < 0.01$).

In addition to the panel sample the 'NEPS grade 4 cross-section' includes 1,141 newly sampled students[56] for a total sample size of 5,817, of which 1,059 are from the Comprehensive states. The DD model presented in Appendix Table B2.2 pools the NEPS-SC2 grade 2 and grade 4 cross-sections for a total sample size of 11,201 student×grade observations.

2.5.2 IQB National Assessment Studies

For auxiliary analyses the paper draws on two large German cross-sectional educational assessment studies carried out by the Institute for Educational Quality Improvement (IQB) at the behest of the Standing Conference of the Ministers of Education and Cultural Affairs of the States (KMK): the IQB National Assessment Study 2011 (*IQB Ländervergleich in der Primarstufe 2011*; IQB11) and the IQB National Assessment Study 2015 (*IQB-Bildungstrend 2015 in der Sekundarstufe I*; IQB15). The purpose of the IQB studies is to monitor in how far students meet nationally defined educational standards for the primary and lower secondary level. Participation in the IQB tests is mandatory for all sampled students.[57] In contrast to the NEPS' sampling design, the IQB studies do not randomly sample from the population of all German students in a particular grade level but, instead, draw separate random samples within each state. Accordingly, smaller states are heavily overrepresented in the IQB data and the use of student sampling weights is necessary to obtain estimates representative of Germany.

[55] Fortunately, primary school (SC2) principals were asked the same questions as those in secondary school (SC3).

[56] Unlike in the NEPS-SC3, the newly sampled students are not part of a refreshment sample. These are students that were part of the SC2 kindergarten sample but then attended a primary school that did not participate in the NEPS. For financial reasons, after kindergarten these students were only tested again in fourth grade.

[57] However, participation in the accompanying student, parent and teacher questionnaires is not mandatory in some states, such that, control variables have more missing values.

2.5.2.1 IQB11

The IQB11 was the first primary-level National Assessment Study (see Stanat *et al.*, 2012, for details). It tested fourth-graders in mathematics, reading and listening at the end of the 2010/11 school year, when students were at the end of their primary school time. As can be seen in Appendix Figure A2.1, this is one cohort later than the NEPS-SC3 cohort. Within each state, sampling followed a multi-stage process: (1) Random sampling of primary schools. (2) Random selection of one fourth-grade class within selected schools. (3) All students in the selected class were obliged to participate.

Again retaining only students with non-missing test scores on regular schools, the total sample size of the 'IQB11 grade 4 cross-section' is 18,904, of which 11,187 are from the Comprehensive states. Note that the number of sampled schools in each state was chosen depending on earlier estimates of the variance in student performance to achieve similar level of precision in each state. Accordingly, the number of observations is different per state and not proportional to the actual share of a state's schools (or students) of the overall German population.

Most of the analysis restricts attention to non-academic-track students. Hence, classifying students as academic- or non-academic-track is crucial. This presents a challenge for using the IQB11 data as there is is no official assignment of students to tracks in primary school yet. Fortunately, however, the IQB11 survey was conducted at the very end of the school year: data collection ran from the end of May until mid July. Fourth-grade students receive their track recommendation with their mid-term reports in January and then start applying for secondary schools. The application period typically ends in March. Accordingly, at the time of the survey it had been decided which secondary school, and hence track, students would attend in the coming year already. The IQB11 survey asked parents directly which school their child will attend in the coming year. As, unfortunately, this variable has about 20% missing values due to parental non-response I also use information on students' track recommendation, which is reported by the school and has almost no missing values (1%), to classify students as academic or non-academic (see below). The resulting IQB11 grade 4 cross-section of non-academic-track students comprises 11,158 observations, of which 6,573 are from the Comprehensive states.

In order to classify students as accurately as possible based on the two above-mentioned variables, I choose state-specific assignment rules that maximise the fit between the state-specific academic-track shares estimated from my sample and the true shares, obtained from administrative records (Statistisches Bundesamt, 2012). In seven states where the track recommendation is non-binding,[58] I assign all students whose parents report that their child will attend an academic-track school in the coming year to the academic track. Among those students with a missing parent answer, I classify students with an academic track recommendation as academic. The rest is classified as non-academic. In the remaining five states where the track recommendation is binding, two (slightly) different assignment rules emerge as best predictors: In Bavaria and Baden-Württemberg, I only assign students to the academic track if they have both an academic-track recommendation and their parents report that they will attend an academic-track school

[58] These are Bremen, Hamburg, Hesse, Lower Saxony, North Rhine-Westphalia, Saarland and Schleswig-Holstein.

in the coming year. All others are classified as non-academic. In Thuringia, Saxony and Saxony-Anhalt, I assign all students whose parents report that they will attend an academic-track school in the coming year to the academic track, unless they fail to have an academic-track recommendation (however, a missing value on this variable is fine). Of those with a missing parent answer, those with an academic-track recommendation are assigned to the academic track. The rest is classified as non-academic.

2.5.2.2 IQB15

The IQB15 tested ninth-graders in reading and listening at the end of the 2014/15 school year, towards the end of lower secondary schooling (see Stanat *et al.*, 2016, for details).[59] As can be seen in Appendix Figure A2.1, this is the same cohort as the NEPS-SC3 cohort used for the main analysis. Sampling and survey design is largely identical to the IQB11 survey. Within each state, sampling followed a multi-stage process: (1) Random sampling of secondary schools. (2) Random selection of one ninth-grade class within selected schools. (3) All students in the selected class were obliged to participate.

All analysis based on the IQB15 data restrict attention to students on regular *non-academic-track* schools with non-missing test scores. The total sample size of the non-academic-track 'IQB15 grade 9 cross-section' is 13,742 students, of whom 7,009 are from the Comprehensive states.

The DD regressions for reading and listening competencies presented in Table 2.3, as well as the DD regressions for non-cognitive outcomes presented in Table B2.8, pool the non-academic-track IQB11 grade 4 and IQB15 grade 9 cross-sections for the 'IQB DD sample' of 24,900 student×grade observations.

[59] Unfortunately, the IQB15 did not test students in maths.

Appendix B: Additional Tables and Figures

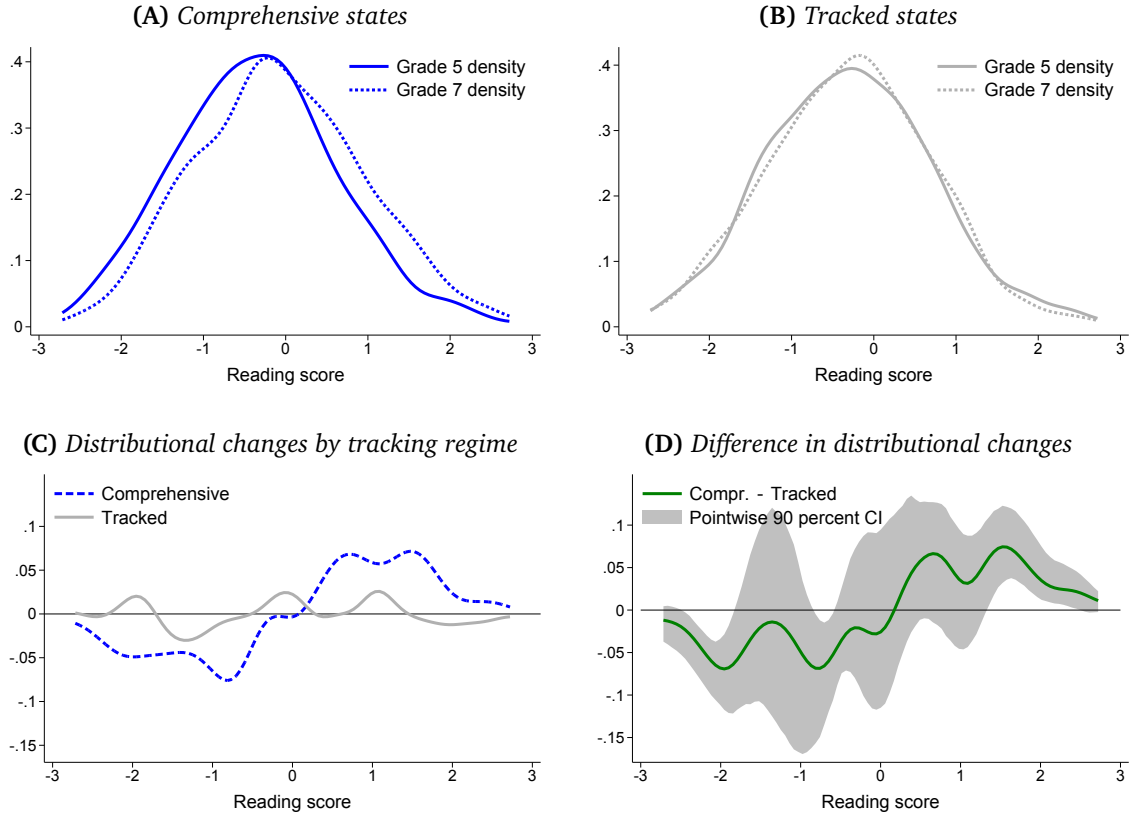


Figure B2.1. Reading score distributions before and after treatment exposure.

Notes: This figure describes how the test score distribution in reading changes differently between grades 5 and 7 depending on the tracking regime. The same notes as in Figure 2.8 apply.

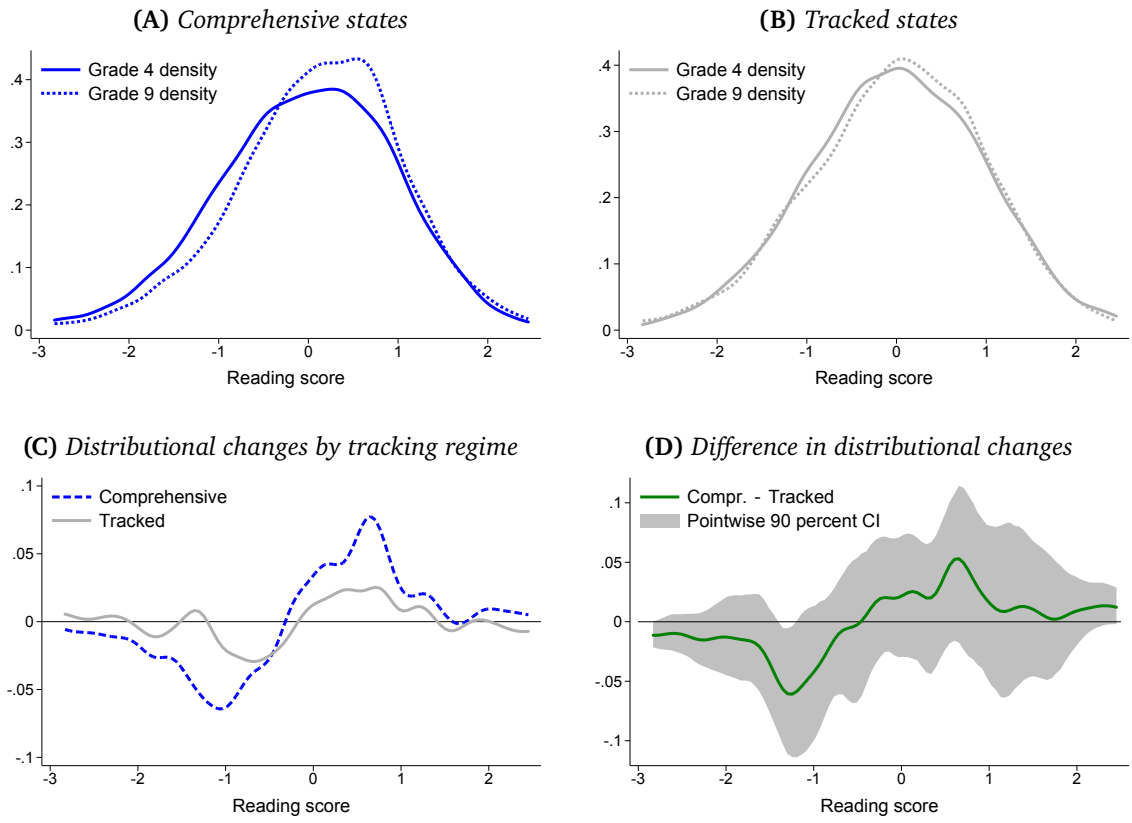


Figure B2.2. Distributional analysis for reading scores in the IQB data.

Notes: This figure repeats the distributional analysis from Figure 2.8 using fourth and ninth grade reading scores in the IQB DD sample ($N = 20,139$). The density estimates are based on the first plausible value and apply student sampling weights. Otherwise the curves are constructed analogously to Figure 2.8.

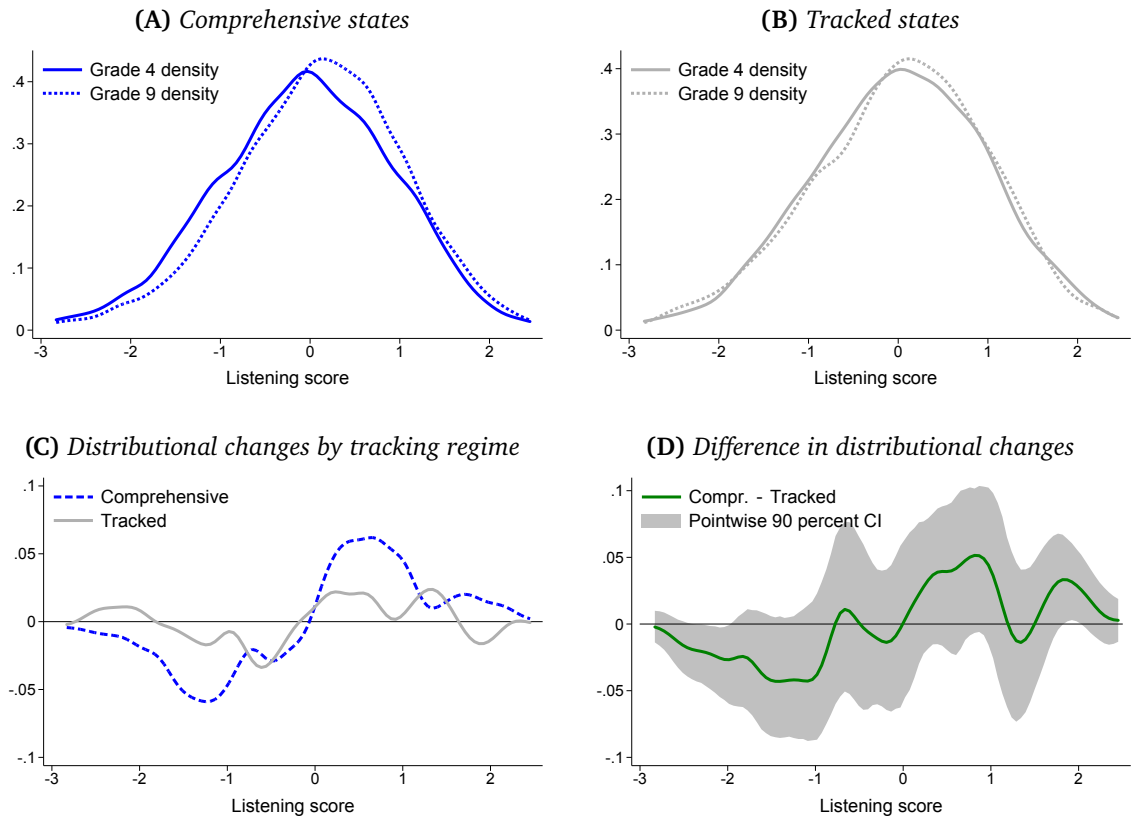


Figure B2.3. Distributional analysis for listening scores in the IQB data.

Notes: This figure repeats the distributional analysis from Figure 2.8 using fourth and ninth grade listening scores in the IQB DD sample ($N = 20,139$). The density estimates are based on the first plausible value and apply student sampling weights. Otherwise the curves are constructed analogously to Figure 2.8.

Table B2.1. Summary statistics and balance test for school characteristics.

	Primary school			Secondary school			Double difference	
	Compr. states	Tracked states	*p*-value (1)=(2)	Compr. states	Tracked states	*p*-value (4)=(5)	$\hat{\beta}^{std}_{DD}$	*p*-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: NEPS data								
Teacher age (years)	47.52	46.64	(0.38)	48.82	47.36	(0.33)	0.14	(0.61)
Further training past year (index)	-0.03	0.02	(0.87)	0.31	0.12	(0.57)	0.14	(0.39)
School size (students)	36.79	47.08	(0.20)	63.92	75.48	(0.52)	-0.04	(0.72)
Student-teacher ratio	14.02	14.66	(0.52)	11.03	12.56	(0.18)	-0.31	(0.46)
School equipment (index)	-0.14	0.07	(0.47)	0.25	0.12	(0.50)	0.31	(0.31)
Educational support (index)	0.14	-0.05	(0.39)	0.92	0.21	(0.02)	0.53	(0.11)
Extracurriculars (index)	0.25	-0.05	(0.29)	-0.10	-0.30	(0.66)	-0.10	(0.89)
Quality control (index)	0.02	0.01	(0.97)	0.05	0.10	(0.91)	-0.06	(0.45)
N schools	62	261		29	133		485	
Panel B: IQB data								
Teacher job experience (years)	23.56	18.27	(0.10)	20.81	13.92	(0.11)	0.14	(0.34)
Further training past two years (hours)	25.60	27.12	(0.58)	25.16	29.50	(0.42)	-0.08	(0.57)
Private school (binary)	0.06	0.02	(0.14)	0.07	0.08	(0.88)	-0.21	(0.20)
Class size	18.89	20.23	(0.06)	23.26	24.10	(0.52)	0.11	(0.66)
All-day schedule/week (days)	2.15	1.94	(0.85)	2.86	2.73	(0.80)	-0.04	(0.95)
Homework support (binary)	0.78	0.72	(0.43)	0.75	0.69	(0.41)	-0.00	(0.99)
Extracurricular learning (binary)	0.32	0.24	(0.13)	0.24	0.28	(0.49)	-0.26	(0.09)
N schools	578	377		340	298		1593	

Notes: This table reports means of school covariates by state group for primary schools (columns 1–2) and non-academic-track secondary schools (columns 4–5). Panel A refers to the NEPS data, where the primary school information comes from the NEPS-SC2 data set and the secondary school information comes from the (main) NEPS-SC3 data set (see Appendix A for details). Panel B refers to the IQB data, where the primary school information comes from the fourth-grade IQB11 data set and the secondary school information from the ninth-grade IQB15 data set. All indices are normalised to mean zero and standard deviation one, separately by grade level. Columns 3 and 6 report *p*-values for tests for zero mean differences in each covariate between the Tracked and Comprehensive states' primary and secondary schools, respectively. Column 7 reports normalised double-differences for each variable, which equal the second difference between Comprehensive and Tracked states between primary and secondary school, divided by the variable's standard deviation. Column 8 reports *p*-values testing for a zero double-difference between state groups and grades. All tests are based school-level regressions and 999 wild cluster bootstraps iterations with Webb weights, clustering at the state level. All figures and tests in this table use school weights provided in the respective data sets.

Table B2.2. DD regression for primary school maths score.

Dependent variable:	Mathematics (1)
Compr. state × Grade 4	-0.015 (0.77) [-0.118, 0.125]
Compr. state	0.043 (0.77) [-0.321, 0.345]
Grade 4	-0.000 (1.00) [-0.161, 0.093]
Adjusted R^2	0.000
N state clusters	12
N Compr. state students	2038
N Tracked state students	9163

Notes: This table reports regression results for the unsaturated DD model applied to primary school maths achievement, i.e. from regressing grade 2 and grade 4 test scores on an intercept, an indicator for the Comprehensive states, an indicator for grade 4 observations and their interaction. This tests whether already in primary school (i.e. before students are exposed to different tracking regimes) mean achievement diverges between Comprehensive and Tracked states. *p*-values in parantheses and 95%-confidence sets in brackets stem from 999 wild cluster bootstrap iterations, clustering at the state level and using Webb weights. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B2.3. Seventh-grade DD robustness checks: alternative model specifications and additional control variables.

	Ind. controls	School controls one-by-one								All school controls				Further school & state controls					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
Panel A: Mathematics																			
Comprehensive	0.165*** (p = 0.01)	0.183*** (0.00)	0.171*** (0.00)	0.135 (0.14)	0.269*** (0.00)	0.173*** (0.00)	0.166*** (0.00)	0.165** (0.03)	0.164*** (0.00)	0.259** (0.05)	0.319* (0.07)	0.246** (0.01)	0.111 (0.24)	0.259** (0.04)	0.307** (0.02)	0.246** (0.03)	0.313*** (0.01)	0.259** (0.04)	0.230** (0.03)
Panel B: Reading																			
Comprehensive	0.243** (0.02)	0.254** (0.04)	0.244** (0.02)	0.240** (0.03)	0.344*** (0.00)	0.251** (0.02)	0.243** (0.01)	0.242** (0.02)	0.238** (0.02)	0.363*** (0.00)	0.375** (0.01)	0.351** (0.01)	0.552* (0.09)	0.362*** (0.00)	0.390*** (0.00)	0.354*** (0.00)	0.365** (0.01)	0.363*** (0.00)	0.333*** (0.00)
Teacher age		✓								✓	✓	✓		✓	✓	✓	✓	✓	✓
Teacher further training			✓							✓	✓	✓		✓	✓	✓	✓	✓	✓
School size				✓						✓	✓	✓		✓	✓	✓	✓	✓	✓
Student-teacher ratio					✓					✓	✓	✓		✓	✓	✓	✓	✓	✓
Facilities						✓				✓	✓	✓		✓	✓	✓	✓	✓	✓
Educational support							✓			✓	✓	✓		✓	✓	✓	✓	✓	✓
Extracurriculars								✓		✓	✓	✓		✓	✓	✓	✓	✓	✓
Quality control									✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Grade × School controls													✓						
Grade × Ind. controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Private school														✓					
Class size															✓				
Timing summer break																✓			
School expenditure																	✓		
Binding teacher rec.																		✓	
Rheinland-Palatinate																			✓
Student sampling weights											✓								
Saturated (w/ state FE)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
Unsaturated (w/o state FE)												✓							
N	5074	5074	5074	5074	5074	5074	5074	5074	5074	5074	4890	5074	5074	5074	5074	5073	5074	5074	5360

Notes: This table reports OLS regression results for variations of the DD model for grade 5 and 7 maths and reading test scores in the NEPS DD sample. Column 1 reposts the saturated DD model with grade-level-interacted student controls from column 4 of Table 2.2 for reference. Columns 2–9 separately add the main eight school controls to this model. Column 10 presents the full model where all school controls are added jointly, corresponding to column 5 in Table 2.2. Column 11 presents results from estimating the full model of column 10 with student sampling weights, provided by the NEPS. Column 12 presents the fully-controlled, *unsaturated* DD model, which instead of state fixed effects includes an indicator for the Comprehensive states. Column 13 returns to the saturated model of column 10 and adds interactions between grade level and all school controls. Columns 14–18 separately add the following variables to the full model of column 10: a school-level indicator for private schools (note: due to the imputation this is a state-level average for grade 5 observations), a school-level measure of average class size (note: state-level average for grade 5 observations), a state-level measure of the time (in months) since the end of the summer break at the day of testing, a state-level measure of per pupil public expenditure for schools, a state-level indicator for binding teacher recommendations. Column 19 presents the full model from column 10, adding observations from the otherwise excluded state Rheinland-Palatinate (RP) to the sample (RP is added to the Tracked states). p-values in parentheses and 95%-confidence sets in brackets stem from 999 wild cluster bootstrap iterations using Webb weights, clustering at the state level. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B2.4. Heterogeneity analysis for seventh-grade achievement in DD and VA models.

Dependent variable:	Level test scores				Grade-5-to-7 gain scores					
Model specification:	DD (1)	DD (2)	DD (3)	DD (4)	FD (5)	VA (6)	VA (7)	VA (8)	VA (9)	VA (10)
Panel A: Mathematics										
Comprehensive schooling	0.207***	0.226**	0.140	0.160**	0.156*	0.153**	0.115	0.144*	0.136**	0.039
	(0.01)	(0.03)	(0.12)	(0.01)	(0.06)	(0.04)	(0.22)	(0.05)	(0.03)	(0.57)
	[0.10, 0.50]	[0.01, 0.67]	[-0.06, 0.30]	[0.06, 0.29]	[-0.01, 0.34]	[0.00, 0.28]	[-0.07, 0.28]	[-0.00, 0.26]	[0.02, 0.29]	[-0.12, 0.23]
× Female		-0.020					0.078			
		(0.88)					(0.42)			
		[-0.28, 0.26]					[-0.16, 0.26]			
× Low SES			0.073					0.019		
			(0.47)					(0.85)		
			[-0.13, 0.49]					[-0.15, 0.35]		
× Migration background				0.026					-0.024	
				(0.91)					(0.87)	
				[-0.65, 0.46]					[-0.33, 0.51]	
× Below median gr. 5 score										0.223
										(0.15)
										[-0.13, 0.48]
Panel B: Reading										
Comprehensive schooling	0.277**	0.212	0.353**	0.209**	0.248**	0.228***	0.178	0.198	0.180*	0.179**
	(0.02)	(0.23)	(0.04)	(0.05)	(0.02)	(0.00)	(0.11)	(0.20)	(0.06)	(0.04)
	[0.05, 0.65]	[-0.10, 0.64]	[0.02, 0.62]	[0.00, 0.37]	[0.06, 0.47]	[0.09, 0.36]	[-0.07, 0.36]	[-0.18, 0.34]	[-0.01, 0.27]	[0.03, 0.38]
× Female		0.145					0.113			
		(0.32)					(0.38)			
		[-0.14, 0.43]					[-0.13, 0.41]			
× Low SES			-0.197					0.064		
			(0.20)					(0.67)		
			[-0.41, 0.10]					[-0.17, 0.58]		
× Migration background				0.178					0.166	
				(0.40)					(0.24)	
				[-0.42, 0.55]					[-0.12, 0.55]	
× Below median gr. 5 score										0.086
										(0.39)
										[-0.24, 0.30]
State FE	✓	✓	✓	✓						
Grade 5 score						✓	✓	✓	✓	✓
N Compr. state students	882	882	882	860	269	269	269	269	269	269
N Tracked state students	4192	4192	4192	4130	1377	1377	1377	1377	1371	1377

Notes: Columns 1–4 are based on the NEPS DD sample of non-academic-track students. Column 1 presents results from the saturated DD model without controls, corresponding to column 2 of Table 2.2. Columns 2–4 fully interact the DD model with indicators for female, below-median socio-economic status (SES) and migration background students, respectively. SES is operationalised as the first principal component of the following variables and their missing dummies: household income, highest parental years of education, home possessions index and the number of books at home. Columns 5–10 are based on the NEPS 5-to-7 panel sample of non-academic-track students. Column 5 presents results for the first-differenced (FD) model, i.e. from regressing grade-5-to-7 gain scores on an indicator variable for the Comprehensive states. The value-added (VA) model in column 6 adds grade 5 test scores as a regressor to the FD model. Columns 7–10 interact the Comprehensive states indicator in the VA model with indicators for female, low-SES, migration background and below-median grade 5 score students, respectively (while each time also adding the indicator as own regressor). *p*-values in parentheses and 95%-confidence sets in brackets stem from 999 wild cluster bootstrap iterations, clustering at the state level and using Webb weights. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B2.5. Effect persistence until ninth-grade.

Dependent variable:	Level test scores		Grade-5-to-9 gain scores	
Model specification:	DD	DD	VA	VA
	(1)	(2)	(3)	(4)
Panel A: Mathematics				
Comprehensive schooling	0.054	0.084	0.088	-0.049
	(p = 0.53)	(0.39)	(0.26)	(0.81)
	[-0.10, 0.33]	[-0.27, 0.43]	[-0.08, 0.25]	[-0.39, 0.25]
× Below median gr. 5 score				0.257
				(0.38)
				[-0.22, 0.78]
N Compr. state students	753	753	186	186
N Tracked state students	3619	3619	1100	1100
Panel B: Reading				
Comprehensive schooling	0.198	0.290	0.163	0.133
	(0.20)	(0.30)	(0.35)	(0.37)
	[-0.23, 0.44]	[-0.30, 0.52]	[-0.14, 0.38]	[-0.29, 0.32]
× Below median gr. 5 score				0.061
				(0.69)
				[-0.25, 0.50]
N Compr. state students	754	754	191	191
N Tracked state students	3522	3522	1064	1064
Controls		✓		
State FE	✓	✓		
Grade 5 score			✓	✓

Notes: Columns 1–2 present OLS regression results for the saturated DD model for grade 5 and 9 test scores in maths (Panel A) and reading (Panel B), estimated on the grade 9 NEPS DD sample of non-academic-track students. Column 1 presents the model without controls and column 2 adds student covariates, interacted with grade level, and school covariates (see notes of Table 2.2). Column 3 presents results for the value-added (VA) model, i.e. from regressing grade-5-to-9 gain scores on an indicator for the Comprehensive states and the grade 5 score. The regressions use the panel sample of non-academic-track students for whom both grade 5 and grade 9 test scores are observed. Column 4 interacts the Comprehensive state indicator with an indicator for students with below-median grade 5 test scores (and adds this indicator as a separate regressor). p-values in parentheses and 95%-confidence sets in brackets stem from 999 wild cluster bootstrap iterations, clustering at the state level and using Webb weights. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B2.6. Summary statistics for the IQB DD sample.

	Non-academic primary school students (IQB11)			Non-academic secondary school students (IQB15)		
	Compr. states	Tracked states	*p*-value (1)=(2)	Compr. states	Tracked states	*p*-value (4)=(5)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Pre-treatment outcomes						
Grade 4 mathematics score	-0.09	-0.00	(0.62)			
Grade 4 reading score	-0.02	-0.00	(0.87)			
Grade 4 listening score	-0.08	0.00	(0.32)			
Self-concept German	0.00	-0.00	(0.94)			
Social integration	-0.13	-0.00	(0.11)			
Reading motivation	-0.09	-0.00	(0.05)			
Attitude towards reading	-0.02	0.00	(0.74)			
Private tutoring	0.23	0.24	(0.81)			
Panel B: Student characteristics						
Female (binary)	0.48	0.46	(0.10)	0.46	0.47	(0.82)
Age	10.63	10.52	(0.09)	15.62	15.66	(0.19)
Migration background (binary)	0.15	0.31	(0.06)	0.18	0.36	(0.06)
Foreign language at home (binary)	0.12	0.21	(0.07)	0.15	0.28	(0.10)
Parental education:						
Low	0.45	0.43	(0.25)	0.50	0.49	(0.60)
Mid	0.41	0.41	(0.88)	0.34	0.34	(0.93)
High	0.14	0.16	(0.13)	0.16	0.17	(0.47)
Parental HISEI score	43.28	44.29	(0.18)	44.82	44.68	(0.91)
N students	6573	4585		7009	6733	

Notes: This table reports means of pre-treatment outcomes and student covariates by state group for primary school students classified as non-academic (columns 1–2) and non-academic-track secondary school students (columns 4–5). The former are based on the grade 4 IQB11 data and the latter on the IQB15 grade 9 data. All figures use student weights. Columns 3 and 6 report *p*-values for tests for zero mean differences between Tracked and Comprehensive states. Test are based on 999 wild cluster bootstraps iterations, clustering at the school level, using Webb weights.

Table B2.7. IQB DD results by plausible value.

Dependent variable:	Reading		Listening	
Plausible value	$\hat{\beta}_{DD}$	*p*-value	$\hat{\beta}_{DD}$	*p*-value
PV1	0.160	(0.12)	0.154	(0.04)
PV2	0.142	(0.21)	0.141	(0.06)
PV3	0.129	(0.25)	0.132	(0.08)
PV4	0.150	(0.14)	0.175	(0.02)
PV5	0.163	(0.18)	0.15	(0.05)
PV6	0.122	(0.24)	0.123	(0.10)
PV7	0.131	(0.18)	0.148	(0.08)
PV8	0.154	(0.12)	0.17	(0.02)
PV9	0.156	(0.14)	0.164	(0.02)
PV10	0.139	(0.13)	0.121	(0.08)
PV11	0.162	(0.10)	0.139	(0.04)
PV12	0.153	(0.11)	0.175	(0.02)
PV13	0.164	(0.11)	0.152	(0.06)
PV14	0.166	(0.10)	0.142	(0.04)
PV15	0.173	(0.07)	0.165	(0.03)
Average	0.151	(.15)	0.150	(.05)

Notes: This table displays coefficient estimates with accordant wild cluster bootstrapped *p*-values for the saturated DD model without controls for reading and listening scores in the IQB sample separately by plausible value. For details about the model see Table 2.3.

Table B2.8. DD regressions for behavioural and socio-emotional outcomes in the IQB data.

Dependent variable:	Self-concept languages (1)	Reading motivation (2)	Attitude towards reading (3)	Social integration (4)	Private tutoring (5)
Comprehensive schooling	-0.012	0.133**	0.152**	0.068	-0.047
	(p = 0.90)	(0.02)	(0.02)	(0.42)	(0.21)
	[-0.20, 0.18]	[0.04, 0.24]	[0.04, 0.26]	[-0.10, 0.24]	[-0.13, 0.02]
Controls	✓	✓	✓	✓	✓
State FE	✓	✓	✓	✓	✓
R^2	0.057	0.077	0.089	0.030	0.030
N state clusters	12	12	12	12	12
N Compr. state students	9465	6743	6907	9594	7872
N Tracked state students	7532	5752	5906	7671	6705

Notes: This table reports results for the fully controlled, saturated DD model applied to different non-cognitive outcomes in the IQB DD sample of non-academic-track students, each time retaining all observations with non-missing values for the respective dependent variable. 'Private tutoring' is an indicator variable equal to one if the student reports receiving private tutoring. All remaining variables are composite scores designed by the IQB, intended to measure the indicated psychological construct, each based on several survey items measured on 4-point Likert scales. I standardise all of them to mean zero and standard deviation one in the group of Tracked states' non-academic-track students, separately by grade. p-values in parentheses and 95%-confidence sets in brackets stem from 999 wild cluster bootstrap iterations using Webb weights, clustering at the state level. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Chapter 3

De-Tracking at the Margin: Local School Supply and Educational Expansion in Germany*

Abstract

Educational expansion is usually ascribed to demand-side factors, such as increasing individual aspirations and labour-market requirements, while supply-side factors, i.e., educational institutions, are often overlooked. We study the effect of local school supply on upper-secondary attainment in Germany's tracked school system. We consider not only the *number*, but also the *type* of available schools by estimating school-type-specific supply effects on attainment of the university-entrance certificate for traditional academic-track schools and newer comprehensive schools and vocational high schools. The latter two offer alternative pathways towards university eligibility, which traditionally could only be obtained on academic-track schools, and thus constitute a partial de-tracking of upper-secondary schooling in Germany. Drawing on yearly administrative records that cover the universe of German students, schools, and graduates, we compile a county-level panel of local school supply and upper-secondary attainment for 13 cohorts between 1995 and 2007. We document that, while attainment has substantially expanded, so has regional dispersion, pointing to growing inequality of educational opportunities. Cross-sectionally, we find that the supplies of all three school types correlate positively with attainment, but for comprehensive and academic-track schools this association is largely spurious, i.e., due to regional differences in educational demand. For vocational high schools, in contrast, we find robust evidence for a positive supply-side effect on attainment, confirmed in two-way fixed-effects, difference-in-differences, and event-study models. The hybrid nature of vocational high schools, combining academic and specialised curricula, might attract students who otherwise would be diverted from academic upper-secondary education towards vocational training.

*This chapter is based on joint work with Camilla Borgna. We benefited from comments and suggestions by Katharina Spieß, Heike Solga, Dalit Contini, seminar participants at WZB, DIW, Collegio Carlo Alberto and conference participants at the CIDER Conference 2019 and the BSE Applied Micro workshop 2021. Funding to purchase the data used in this project was generously provided by the Jacobs Foundation and German Federal Ministry for Education and Finance (BMBF) through the College for Interdisciplinary Educational Research (CIDER).

3.1 Introduction

Over the last century, educational attainment has steadily increased across the world. A large body of research has investigated the distributive consequences of educational expansion, notably in terms of inequality of opportunity (Shavit and Blossfeld, 1993; Breen and Jonsson, 2005; Arum *et al.*, 2007; Breen *et al.*, 2009). We know less about the causes of the expansion, however. Rising level of educational attainment are usually ascribed to an increased demand for higher educational credentials following increased skill demand on labour markets Goldin and Katz (2010) and credential inflation (Berg, 1971; Freeman, 1976; Thurow, 1975). What has received less attention is the supply side—in particular, the capacity of schools to accommodate the increasing demand and to, potentially, even boost it (Walters, 2000; Jackson, 2021).[1]

Expanding school supply can be seen as a particular type of educational reform, which, along with other changes in the organisation of school provision, characterised the second half of the twentieth century in most industrialised countries (Erikson and Jonsson, 1996; Kerckhoff *et al.*, 1996). By definition, the increasing availability of schools expands the opportunity structure in which individuals take their educational decisions, by "allowing more and more students to stay in school for longer and longer periods of time" (Walters, 2000, p. 242). However, in how far students use these opportunities, with the result of higher enrolment and attainment rates, and whether this usage depends on the type of schools supplied is an empirical question. Assessing if reforms affect educational expansion is challenging because of the co-occurrence of many historical processes involving school supply, its allocation rules, and, crucially, educational aspirations (e.g., Jackson, 2021).

This paper takes a step forward in assessing the relationship between school supply on educational expansion by analysing regional heterogeneity and over-time transformations of secondary schooling in Germany. In doing so, we also contribute to the literature on geographical differences as a source of inequality of opportunity. The place where individuals live during their childhood and youth profoundly shapes their life chances, as indicated by a large literature on neighbourhood effects (e.g., Ainsworth, 2002; Brännström, 2008; Owens, 2010; Chetty and Hendren, 2018). An important mechanism through which place of residence affects educational opportunities relates to the local availability of educational institutions. For example, research has shown that living in areas with more college options increases the chances of college attendance (Turley, 2009; Spiess and Wrohlich, 2010; Klasik *et al.*, 2018; Hirschl and Smith, 2020). Our paper contributes by focusing on geographical differences in the availability of secondary schooling options, which might create disparities in the opportunity to even pursue higher education.

Germany makes for an interesting case study because the large expansion of educational credentials has occurred in spite of the apparent stability of its rigid and early between-school tracking system: already at age ten, most students are tracked between three hierarchically

[1] Note that in the literatures on early childhood education and care (see, e.g., Schober and Spiess, 2013; Havnes and Mogstad, 2015) and on tertiary education (see, e.g., Kamhöfer *et al.*, 2019) there has been a stronger focus on supply-side expansions.

ordered educational tracks which correspond to different school types ('tripartite' system). While, in the post-war period, most European countries implemented major school reforms moving from similar selective systems to models of comprehensive schooling (Erikson and Jonsson, 1996; Kerckhoff *et al.*, 1996; Leschinsky and Mayer, 1999), West Germany eschewed such structural de-tracking reforms (Ertl and Phillips, 2000; Neugebauer *et al.*, 2013).[2]

The stability of tracking has been attributed to the strength of vocational education and training (VET) in Germany. In particular, the so-called 'dual' system, which combines firm-based apprenticeships with school-based learning, constitutes an equilibrium that key stakeholders—governments, firms, students, and their families—have little incentive to alter (Thelen, 2004; Busemeyer and Trampusch, 2011). Early tracking is linked to the dual system because, even though lower-secondary schooling is still centred on general skills across all three tracks, the distinct pathways ensure that a majority of students is channelled into upper-secondary VET eventually. The dual system is seen a mixed blessing by the literature on education and work. On the one hand, it is widely praised for facilitating job matches (Shavit and Müller, 1998) and smooth labour-market transitions for non-college-bound youths (Brzinsky-Fay, 2007), while providing firms with workers trained according to their skill requirements (Acemoglu and Pischke, 1998). On the other hand, precisely because of its merits, it constitutes a major institutional barrier to higher education expansion (Powell and Solga, 2011). Indeed, VET is a viable and attractive alternative to university, as it opens up concrete opportunities to enter middle-class jobs. Accordingly, many students are 'diverted' from academic/higher education and forego the benefits of an education centred on general skills, which over the life course might be more profitable than the occupation-specific skills acquired in VET—especially, in times of rapid technological change and globalisation (Hanushek *et al.*, 2017). The 'schism' between academic and vocational education is also criticised for contributing to the intergenerational reproduction of social inequality, because it is especially students from working-class backgrounds who are diverted towards VET, which they perceive as less risky than the university route (Shavit and Müller, 2000; Mayer *et al.*, 2007). Finally, inequality is aggravated by the early tracking system that precedes VET, because school performance at an age when students are strongly influenced by their family upbringing determines their access to highly stratified learning environments in secondary school (e.g., Hanushek and Woessmann, 2006; Pfeffer, 2008; van de Werfhorst, 2019; Matthewes, 2021).

Nevertheless, even in the German case, it would be wrong to talk about institutional stasis. During the last sixty years, gradual and subtle changes have operated at the margins of the tracking system: these transformations have followed the logic of institutional layering, according to which innovations are introduced *on top of* existing organisations and practices instead of re-

[2]This is not to say that de-tracking was never part of the political agenda in Germany: quite the contrary. During the 1960s, the public debate over pedagogical reforms was so ideologically charged that some scholars went as far as characterising it as a 'religious war' (Fend, 1982). The debate resurfaced in the 1980s and 1990s, when the educational system started to be under pressure due to declining birth rates and reunification. Despite these lively debates, efforts for structural reforms have so far failed to garner a broad enough coalition of parties and stakeholders to materialise (Edelstein and Nikolai, 2013).

placing them (Streeck and Thelen, 2005).[3] In this paper, we analyse two reforms of institutional layering, namely the introduction and gradual expansion of two new school types that were added to the traditional tripartite system: comprehensive schools (*integrierte Gesamtschule*) and vocational high schools (*berufliches Gymnasium/Fachgymnasium*). Both constitute alternative pathways toward the university-entrance certificate (*Abitur*), which traditionally could only be attained via the selective academic track (*Gymnasium*). These reforms produced what may be called de-tracking 'at the margin' of the institutional framework: without dismantling the tracking system, they altered the *de facto* structure of opportunities available to students (Leschinsky and Mayer, 1999). However, not all students were equally exposed to these innovations because, even though the new school types were legally introduced by state governments, their implementation depends on local authorities: hence, the educational opportunities actually available to students depend on when and where they go to school (Helbig and Nikolai, 2015).[4]

A first aim of this paper is to carefully document regional differences in both school supply and educational attainment at the upper-secondary level and the extent to which they varied over time. For this, we cannot rely on survey data because it typically is not representative at levels lower than the national or, at best, state level. Instead, we combine regional administrative data sources on school supply, student enrolment, and graduation counts for all German counties in the time frame 1995–2016. This allows us, for any cohort that entered lower-secondary schooling between 1995 and 2007, to reconstruct the school mix locally available at the time of the primary-secondary transition and their attainment levels nine years later. While we find that average attainment is rising steadily, we also find that cross-county differences in upper secondary attainment are striking and increasing over time, indicating growing regional educational inequality in Germany. Regional differences in school supply are equally large, yet more stable over time. However, especially vocational high schools have slowly but steadily expanded over the period considered.

The second aim of this paper is to unravel the role of school supply for educational expansion at the upper-secondary level. Beyond the *quantitative* dimension of school supply, we also focus on its *qualitative* dimension: we analyse if the introduction and expansion comprehensive schools and vocational high schools contributed to attainment growth differently than traditional academic-track schools. Here, regional and temporal differences in the school mix locally available to students represent an advantage because they provide multiple sources of variation we can exploit to disentangle supply- from demand-side factors. Our analytical strategy proceeds in three steps. First, we model regional differences cross-sectionally, controlling for cohort fixed-effects, state fixed-effects, and a large array of county-level socio-economic characteristics, aimed at capturing the most important endogenous dynamics in the demand for education. Second, to circumvent remaining unobserved heterogeneity between counties, we simultaneously use between-county and over-time variation in school supplies in a two-way

[3] Processes of incremental change are not foreign to other sectors of the German educational system, as documented by recent studies on the progressive hybridisation of VET and higher education (Graf, 2013, 2018; Durazzi, 2019).

[4] The tracking system was also subject to other marginal reforms (see below), but these were typically decided and implemented at the state level and did not result in fine-grained regional heterogeneities.

fixed-effects (TWFE) framework. Finally, we exploit the discrete nature of the introduction of new school types in counties where they were not previously available as 'natural experiments' in a difference-in-differences (DD)/event-study framework.

We find that the local supplies of all three school types with upper-secondary provision correlate positively with *Abitur* attainment levels. However, rather than causal, for comprehensive and academic-track schools this association appears to be mainly spurious, namely due to demand-side factors (social composition of resident families and the degree of urbanisation) that correlate with both supply and attainment levels. For vocational high schools, in contrast, we find a positive supply-side effect on attainment that is robust to controlling for various socio-demographic controls and state fixed-effects, as well as weighting for population size and aggregating to higher regional levels. The TWFE and DD regressions (again, regardless of weighting, level of aggregation, sample restrictions and conditioning sets), as well as the event-study analysis confirm a causal positive supply-side effect.

Our estimates suggest that, at the extensive margin, the introduction of a single vocational high school, when it was not previously available in a county, increases attainment rates by about 1.5 percentage points. Across extensive and intensive margins, a supply increase of one school per 1,000 students (corresponding to an increase of about one slot per 100 students) increases attainment rates by roughly the same amount. While the reasons for the success of vocational high schools are probably multifold, we suggest that a crucial factor resides in their hybrid nature: because they combine academic and specialised curricula, they might be especially attractive to risk-averse students who, in the traditional system, would have been diverted towards VET.

The paper proceeds as follows: section 3.2 describes the German school system and the introduction of vocational high schools and comprehensive schools. Section 3.3 describes educational expansion in Germany and develops our hypothesis regarding the role of school supply in that process. Section 3.4 describes the construction of the data. Section 3.5 presents our empirical strategy for the study of school supply effects on attainment. Section 3.6 presents the empirical results. Finally, section 3.7 discusses implications and concludes.

3.2 De-Tracking at the Margin: New Pathways to the *Abitur*

In Germany, students attend four years of comprehensive primary school[5] before they are tracked into distinct school types according to their previous performance (at about age ten). Traditionally, there were three options available at this transition, with a strict correspondence between secondary school type, track, and school-leaving certificate: the selective academic track (*Gymnasium*), lasting nine years and preparing for university; the intermediate track (*Realschule*), lasting six years and preparing for higher-level VET; and the low track (*Hauptschule*), lasting five years and preparing for lower-level VET. Importantly, in the traditional tripartite system, the

[5] In three states (Berlin, Mecklenburg-Vorpommern, and Brandenburg) comprehensive primary school lasts until the end of grade six.

only way to obtain the university-entrance certificate (*Abitur*) was to attend an academic-track school.

Over the past sixty years, the federal states, which are responsible for educational policy-making in Germany, have implemented different changes to the above-sketched tripartite system. Some changed the allocation rules of tracking by abandoning the binding nature of teachers' track recommendations (Neugebauer, 2010; Dollmann, 2016), by opening second-chance options and/or by enhancing track permeability (Buchholz and Schier, 2015; Schindler, 2015, 2017). Some states have even unified the lower two tracks (Neumann *et al.*, 2013; Matthewes, 2021). However, none of them changed the tracking system's most fundamental rationale of dividing the student body between academic-track and non-academic-track schools at an early age.

The introduction of *comprehensive schools* and *vocational high schools* constitute significant disruptions to the tripartite system because, in contrast to its traditional schools, they do not correspond to a single track. Instead, they enable the pursuit of multiple certificates and, most importantly, offer alternative pathways toward university entrance eligibility, i.e., *Abitur* attainment. In doing so, however, they follow two very different logics: comprehensive schools—like traditional academic-track schools—cover the full stage of secondary schooling (grades 5–13), whereas vocational high schools, as the name suggests, only cover the upper-secondary stage (grades 11–13) (see Figure 3.1 for a schematic overview). Both school types were legally introduced at the state level, but their actual provision was left to local authorities. Therefore, at a given point in time, the availability of these alternative pathways to *Abitur* does not only depend on the state a student lives in, but also on her location within that state.

Comprehensive schools (*integrierte Gesamtschule*)[6] were first introduced in the 1960s by some West German states (Berlin, North Rhine-Westphalia, and Lower Saxony) as pilots within a reform process aimed at progressively de-tracking lower-secondary schooling (Köller, 2008). Due to intense opposition from the public, however, the full reform was never brought to completion: comprehensive schools were eventually institutionalised, but without replacing the three traditional tracks as originally intended. Over the decades that followed, most other states followed suit in introducing comprehensive schools as an alternative lower-secondary option *next* to the three school types of the tracked system (as of today, only Bavaria and Saxony do not have comprehensive schools) (Helbig and Nikolai, 2015).[7]

[6]The *integrierte Gesamtschule* should not be confounded with the *cooperative* or *additive Gesamtschule*, which is simply a school building that hosts distinct tracks (partially overcoming the physical segregation of tracked students).

[7]Two periods of reform momentum are worth mentioning. First, in the 1990s, the East-West reunification and related demographic changes provided a window of opportunity for the reorganisation of school supply: many states abolished the lower tracks or reduced their supply, replacing them with hybrid school types (Neumann *et al.*, 2013). Some of these hybrids are comprehensive schools, but most of them are multi-track schools (*Schule mit mehreren Bildungsgängen*) which prepare for low and intermediate certificates but not for the Abitur. A second historical turn happened in the early 2010s, when many states drastically increased the supply of comprehensive schools, replacing low and intermediate tracks: in Berlin, Bremen, Hamburg, Schleswig-Holstein and Rhineland Palatinate the only secondary school options became academic-track or comprehensive schools. Despite the radical character of this reform, in most cases it did not entail the establishment of new schools, as was the case in the past. Rather, these 'new' comprehensive schools were created by merging and relabelling already existing schools, with a significant degree of continuity in terms of resources and reputation (Helbig and Nikolai, 2015, 2017; Neumann *et al.*, 2017). It may

Since comprehensive schools are the product of historically stratified reforms, their names and specificities vary slightly across the country.[8] Nevertheless, they share a number of fundamental elements. Most importantly, comprehensive schools prepare for multiple certificates: during the lower-secondary phase (grades 5–10), they offer curricula that are compatible with continuing general schooling or transitioning to VET in the upper-secondary phase, in which case students leave school with a low (after grade 9) or intermediate (after grade 10) certificate. During the upper-secondary phase (grades 11–13), comprehensive schools provide academic instruction either in-house or (sometimes) via cooperation agreements with other schools, hence allowing the pursuit of the *Abitur*. Because they are not geared towards a specific certificate, comprehensive schools typically have a more heterogeneous student body than traditional track-specific schools. Note that, despite their name, most comprehensive schools implement some form of ability grouping for specific (core) subjects from grade 7 onward (Köller, 2008).[9] Generally, one must keep in mind that comprehensive schools do not imply comprehensive schooling because they coexist next to the tracked system: most high-achieving students attend academic-track schools, so that comprehensive schools are best seen as an option that allows medium-achieving students to postpone their decision of which certificate to pursue.

The introduction of vocational high schools (*Fachgymnasium/berufliches Gymnasium*) has been less controversial than that of comprehensive schools, but similarly gradual and geographically heterogeneous. While some scattered predecessors of vocational high schools were already present at the beginning of the twentieth century Brauckmann and Neumann (2004), strictly they were introduced between the late 1960s and 2000s, with Bavaria as the last state to adopt this innovation (Helbig and Nikolai, 2015).[10] Due to this heterogeneous implementation, their names and specificities differ slightly between states, similar to comprehensive schools. We adopt the conventional definition of vocational high schools as schools that only cover the three academic upper-secondary grades (11–13), award the *Abitur* certificate, but, at the same time, have a vocational focus. Administratively, vocational high schools belong to the VET system but in fact they are much more similar to academic-track schools than to other VET institutions. Vocational high schools follow the standard upper-secondary academic curriculum in core subjects (German, mathematics, foreign languages) merely replacing some secondary subjects with occupation-specific courses in a field of specialisation chosen by the student. The fields students can pick from are: business administration, technical studies, health and social care, agriculture, nutritional sciences, social pedagogics, and biotechnology.[11]

---

therefore be argued that the comprehensive schools established with the 2010s reforms are not fully comparable with the previous ones. In any event, this does not affect out analyses because, our observation window stops with the cohort that entered lower-secondary schooling in 2007.

[8] Comprehensive schools are variously known as *integrierte Gesamtschule*, *integrierte Sekundarschule* and *Gemeinschaftsschule*. We follow the defining approach of national statistics, by grouping together all lower-secondary schools (other than academic-track schools) that enable the pursuit of the *Abitur*.

[9] The degree of streaming (e.g., the number of subjects involved and the number of ability levels) varies between states and sometimes even from school to school.

[10] Note that formally in Bavaria there are no vocational high schools but only *Fachoberschulen*, which we treat as functionally equivalent.

[11] Not all specialisations are available at all schools. In fact, many vocational high schools specialise in one or two fields.

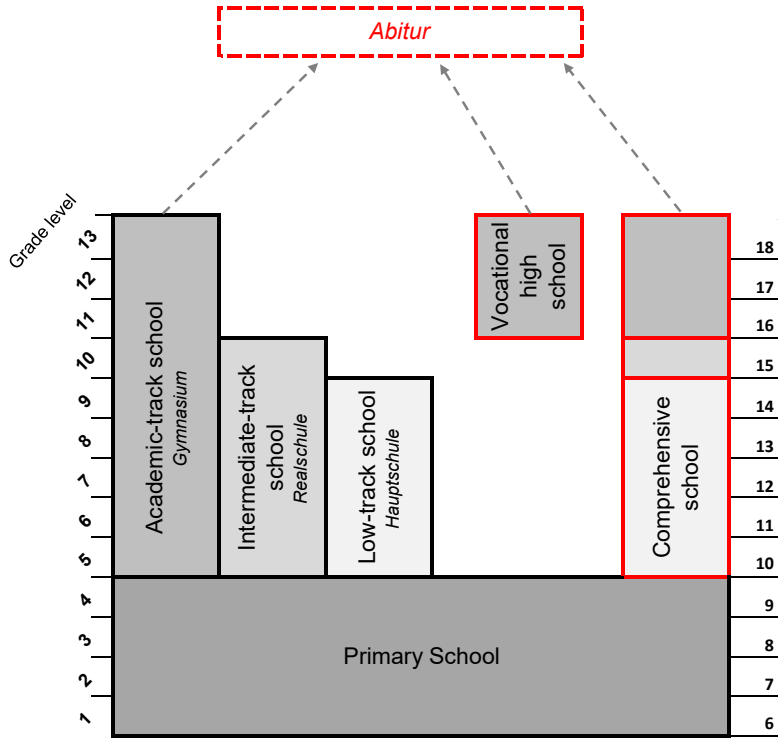


Figure 3.1. Traditional and alternative pathways to the *Abitur*.

Source: Own illustration.

Vocational high schools only cover upper-secondary schooling, so students enrol in these schools after having completed lower-secondary schooling elsewhere. Admission requires the intermediate school-leaving certificate with a sufficiently high grade point average (GPA).[12] The vast majority of enrollees attended lower-secondary schooling in an intermediate-track school (or initially attended the low track and then upgraded to the intermediate one after grade 9) and then directly enrol in vocational high school after their graduation in grade 10 (Winkler, 2017; Zimmermann, 2019). A small share of vocational high school enrollees are students switching from academic-track schools for upper-secondary schooling (e.g., because they seek the vocational specialisation). In some cases, vocational high schools are also formally open to older individuals who chose vocational training after completing lower-secondary school, but empirically this pathway is not common (Winkler, 2017; Zimmermann, 2019).[13]

On vocational high schools, students can obtain three kinds of upper-secondary certificates: (i) the standard *Abitur* (*allgemeine Hochschulreife*) which provides access to all university majors; (ii) a restricted version of the *Abitur* (*fachgebundene Hochschulreife*) which allows students to enrol in university but only in given majors; and (iii) a vocational certificate (*Fachhochschulreife*) which enables the continuation of vocational education at a post-secondary level (in universities of applied science) and can also be attained in other vocational schools.

[12] The minimum GPA requirement vary between 2.5 and 3.0 across states (on a scale from 1.0, the highest GPA, to 6.0, the lowest GPA).

[13] Other institutions (the so-called 'second-chance' sector) are explicitly designed for non-linear pathways, allowing students to enter academic education after VET.

Table 3.1. Summary of school characteristics.

	Academic-track schools	Comprehensive schools	Vocational high schools	Intermediate- & low-track schools
Covers lower secondary phase	YES	YES	NO	YES
Covers upper secondary phase	YES	YES	YES	NO
Awards the *Abitur*	YES	YES	YES	NO
Selection on achievement	Stringent	Absent for lower secondary Moderate for upper secondary	Moderate	Mild/absent
Learning standards	High	Medium-high	Medium-high	Medium/low
Peer composition	Medium-high SES High achievers	Heterogeneous by SES and achievement	Heterogeneous by SES and achievement	Medium-low SES Medium/low achievers
	The same throughout lower and upper secondary	The same in upper secondary except for those who leave after grades 9/10	Completely new peer group in upper secondary	–

Notes: Own illustration, partially based on Brauckmann and Neumann (2004); Dustmann *et al.* (2017); Köller (2008).

Figure 3.1 summarises the structure of the German tracking system by highlighting the three main pathways to *Abitur* attainment: academic-track schools (covering grades 5–13 and preparing exclusively for the *Abitur*), comprehensive schools (also covering grades 5–13, but preparing for multiple certificates, depending on whether students leave after grade 9, 10, or 13), and vocational high schools (covering only grades 11–13 and preparing for various types of upper-secondary certificates). While academic-track schools remain the conventional way to attain the university-entrance certificate (in 2018, 72% of *Abitur*-holders came from this pathway), comprehensive and vocational high schools make up for an important portion of students who reach this qualification (9% and 14%, respectively).[14] It is important to stress that the *Abitur* attained in comprehensive schools and vocational high schools is legally equivalent to the one issued by academic-track schools, and universities are not allowed to discriminate between students based on where they attained the certificate.

Table 3.1 summarises the main characteristics of the different school types along several dimensions relevant for school choice, namely the educational pathways they prepare for, their selectivity and learning standards, and the composition of student body in terms of socio-economic status (SES) and achievement.

3.3 Educational Expansion and School Supply in Germany

Despite the relative stability of its school system, over the last half a century, Germany has experienced a large expansion of secondary credentials. According to data from the Federal

[14] Source: Statistisches Bundesamt (2001b,a)

Ministry of Education and Research, in 1965, most students left secondary school with a low certificate (54.2%) or without any kind of certificate (17.3%). By 2018, these shares have fallen to 17.6% and 6.8%, respectively, making low levels of attainment the exception rather than the norm. Most impressively, over the same period, the share of students who obtain the *Abitur* certificate has skyrocketed from 5.7% to 39.9%.[15] Corresponding to these developments in secondary schooling, the historical dominance of VET has steadily declined, while tertiary education expanded (e.g., university enrolment increased from 15% to 26% between 1980 and 2018[16]). Nowadays there is rough parity in the numbers of new tertiary education enrollees and new apprentices each year (Autorengruppe Bildungsberichterstattung, 2020).

What are the determinants of this striking educational expansion? Undoubtedly, demand factors are decisive. Like elsewhere, qualification requirements on the labour market have grown in Germany. This is partly due to shifts in the occupational structure: the white-collar occupations for which vocational training prepares have been in constant decline, to the advantage of professional and managerial occupations that require skills typically acquired in higher education (Dauth *et al.*, 2021; Oesch, 2013; Oesch and Rodríguez Menés, 2011). At the same time, educational expansion itself can fuel a cycle of credential inflation Berg (1971); Collins (1979), where employers demand increasingly high qualifications and job seekers have to pursue more and more education to secure their position in the labour queue (Thurow, 1975). In Germany, both processes translated into behavioural changes starting with the choice of secondary school: lower tracks have become less and less popular over the years, because students are increasingly interested in pathways that enable the pursuit of the *Abitur*. The latter is not only the necessary requirement to enter university, but also serves as a signal of future trainability in the apprenticeship market. In fact, already more than ten years ago the *Abitur* had become a *de facto* prerequisite for 'higher-ranked vocational training opportunities, such as for bank clerks or information technology clerks' (Powell and Solga, 2011; Solga *et al.*, 2014).[17]

Nonetheless, also supply factors may have played a role in the process of educational expansion. Baethge and Wolter (2015, p. 104) argue that while the "decisive dynamic behind this shift in the educational choices and the allocation processes between alternative school types seems to be the increasing level of educational awareness, aspirations and ambitions in wider parts of the population", all the same "processes of opening up institutional transitions (e.g., between primary and grammar schools or between the lower secondary level and the upper level of grammar schools) have additionally reinforced the expansion". However, to our knowledge, there is a paucity of empirical evidence on such supply-side effects on student attainment. An exception is Schindler and Bittmann (2021) who also study the expansion of alternatives routes toward the *Abitur* and apply decomposition methods to analyse whether this affected

[15] Source: Autorengruppe Bildungsberichterstattung (2020, Table 2.3.16.)

[16] These figures exclude universities of applied sciences. Including those, the numbers are 20% and 45% for 1980 and 2018, respectively. The figures for 2018 exclude international students. Those for 1980 do not because the distinction between domestic and international students is unavailable. However, the number of international students in 1980 is likely negligible. Source: Autorengruppe Bildungsberichterstattung (2020, Table F3-1web).

[17] While, traditionally, apprentices in the dual system came from the lower tracks, a growing share of them now possesses the *Abitur*: in 2018 this share was 23%, equating the percentage of individuals with a low certificate (Autorengruppe Bildungsberichterstattung, 2020, Table E3-2web).

social inequality in attainment, concluding that it likely did not. However, their paper does not investigate the reforms' effect on attainment *levels* and does not differentiate between specific school types. Moreover, they acknowledge that their analysis is unable to identify a key quantity required for a conclusive verdict on the causal reform effects, namely the number of students who are diverted from academic-track schools into non-traditional pathways.

Indeed, it is crucial to establish how non-traditional school types enrollees would have fared *had these schools not been locally available*. A 'diversion' scenario implies that these students would have enrolled in academic-track schools—either by enrolling directly after primary school or by 'upgrading' after grade 10. An 'inclusion' scenario implies that after the lower-secondary stage they would have transitioned to VET. This is where our study aims to contribute by isolating plausibly exogenous variation in the supply of these schools: positive supply-side effects on *Abitur* attainment are evidence for the inclusion mechanism because it implies that students otherwise would have not attained as highly. Our focus on causal identification comes at the cost of using aggregate data that prohibits distinguishing between different social groups.

In light of the school characteristics discussed in Section 2, we hypothesise that the inclusion pattern prevails, i.e., that *Abitur* attainment rates should raise following an increased availability of comprehensive schools (*hypothesis 1*) and vocational high schools (*hypothesis 2*). In particular: (i) the milder selectivity of comprehensive and vocational high schools make them more inclusive compared to academic-track schools; (ii) this should also favour late bloomers, who are not channelled into the academic track as young children; (iii) low-SES students often do not consider academic-track schools because they are perceived as too culturally distant (e.g., Dumont *et al.*, 2019) and non-traditional schools overcome this barrier due to their more heterogeneous student bodies; (iv) students may abstain from upgrading to academic-track schools because they do not want to become outsiders in established friendship networks, an obstacle that is not present in vocational high schools which commence in grade 11; (v) non-traditional schools have, or at least might be perceived to have, lower learning standards, implying a lower risk of failure: this makes them especially appealing for low-SES students, who—even conditional on school performance—often refrain from choosing ambitious pathways due to a higher perceived risk of failure (Breen *et al.*, 2014).

While our main research questions concern the non-traditional pathways, we also consider the effects of traditional academic-track schools as a benchmark for general supply-side effects: in principle, gradual de-tracking and subsequent increases of *Abitur* attainment rates may also be achieved through an expansion of slots for in the traditional academic track (*hypothesis 3*).

3.4 Data

Our analysis is based on regional administrative records for the time frame 1995–2016 provided by the German federal states.[18] The raw data consists of secondary school-type specific counts

[18]These are official administrative education statistics provided by the federal states, the majority of which is freely available from the online database: https://www.regionalstatistik.de/genesis/online. Because some specific

of schools, enrolled students, and *Abitur* graduates in a given year for a given region. The lowest level of regional aggregation available in the data—and our primary regional unit of analysis—is the level of counties (NUTS-3), of which there are 402 across Germany. Deciding on the level of regional aggregation involves a trade-off: lower levels of aggregation allow for a more granular depiction of regional differences and increase statistical power, but come at the cost of increased measurement error due to student mobility across regional borders. Therefore, we repeat all our analyses at the level of planning regions, of which there are 96 across Germany and which thus represent a substantially higher level of regional aggregation. In the appendix we replicate our analyses at even higher levels, namely provinces (NUTS-2), of which there are 41, and states (NUTS-1), of which there are 16.[19]

Regardless of the chosen level of regional aggregation, we have to transform the raw region×year count data into a region×cohort dataset to relate local school supply with attainment rates. We define *secondary school entry* cohorts by the year students enter fifth grade, which corresponds to the first year of secondary school.[20]

Our main independent ('treatment') variables of interest are the local supplies of the three different school types with upper-secondary provision. We operationalise the supply of school type $j \in \{V, C, A\}$ (for vocational high schools, comprehensive schools, and academic-track schools, respectively) for entry cohort c from region i and as the number of schools available in that region (per 1,000 students) in the relevant year of transition, t_{icj}, which in the vast majority of cases equals the year of fifth grade entry, c, for comprehensive and academic-track schools and the year of eleventh grade entry, $(c + 6)$, for vocational high schools:[21]

$$S_{ic}^{j} = 1,000 * \frac{N \text{ schools of type } j \text{ in region } i \text{ in year } t_{icj}}{\text{size of cohort } c \text{ in region } i}$$

Clearly, the number of schools of each type is a rather crude measure of supply compared to the number of available slots which we would ideally use but do not observe. Importantly, however, it ensures that we measure school supply *net of demand*, whereas more granular enrolment counts constitute the equilibrium outcome between supply and demand for specific school types and are thus unsuited for an investigation of supply-side effects. In Appendix Figure A3.1 we

school-type-specific figures are not available there, we bought the remaining data directly from the statistical offices of the federal states.

[19]The German names for counties, planning regions, provinces, and states are *Kreise*, *Raumordnungsregionen*, *Regierungsbezirke*, and *Länder*, respectively. Planning regions do not fall into the European NUTS (Nomenclature of Territorial Units for Statistics) classification, but are defined by the German Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR). As they fall in between NUTS-2 and NUTS-3, we label them NUTS-2.5. Note that, in each step, higher-level units strictly nest lower level ones.

[20]In three states (Berlin, Mecklenburg-Vorpommern, and the excluded Brandenburg) the primary-secondary transition takes place two years later, i.e., after sixth grade. As grade repetitions spike in the year before the transition, we record the size of entry cohorts in seventh grade when all students have transitioned, by summing over the seventh-grade enrolment counts of all secondary school types.

[21]The time point of transition is a function of school type (j) because students enter comprehensive and academic-track schools directly after primary school, but vocational high school only after completing lower-secondary schooling (i.e., after grade 10). It is further a function of county (i) and cohort (c) because in a few states the transition from primary to secondary schools takes place after grade 6 instead of grade 4, as is the norm. The later transitioning states are Berlin and Brandenburg throughout our observation period and Bremen, Lower Saxony and Mecklenburg-Vorpommern for parts of our observation period.

compare the average size of the three school types: the median number of enrolled upper-secondary students per school per cohort is 71 for vocational, 65 for comprehensive, and 85 for academic-track schools, indicating that they are roughly comparable in the number of available slots per school.

Our outcome of interest is the *Abitur* attainment rate, i.e., the share of a given cohort from a given region graduating high school with a university-entrance qualification.[22] As in the data we cannot follow individual students over time, we approximate the attainment rate of cohort c from region i by dividing the region's count of *Abitur* graduates in the year students from that cohort are *expected* to graduate, g_{icj}, by the cohort's size:[23]

$$Y_{ic} = \frac{\sum_{j \in \{V,C,A\}} \left(N \text{ Abitur graduates from school type } j \text{ in region } i \text{ in year } g_{icj}\right)}{\text{size of cohort } c \text{ in region } i}.$$

Accordingly, we introduce measurement error in our outcome variable to the extent that students do not graduate in the expected year (e.g., through grade repetition).[24] This first measurement problem is innocuous for identification as long as grade repetition is not systematically related to our measures of school supply.

A second measurement problem arises if students move to another region *between* lower- and upper-secondary school, because cohort size—the denominator of the outcome variable—is recorded when students are in lower-secondary school (in grade 7), but the number of graduates—the numerator of the outcome variable—is recorded at the end of upper-secondary school (after grade 12/13). There is no problem if students attend a secondary school in another region directly after primary school and by far the most students do not change school again after their first secondary enrolment. Yet, 'upgrading' students, i.e., those who want to pursue *Abitur* but are enrolled in a school without upper-secondary provision, have to enrol in a new

[22] We include in our definition of university-entrance qualification the standard *Abitur* (*allgemeine Hochschulreife*) and its major-bound version (*fachgebundene Hochschulreife*), while we exclude the certificate that only gives access to universities of applied science (*Fachhochschulreife*),

[23] The year of expected graduation is a function of region, cohort, and school type because of the so-called G8 reform, which was introduced by most states at some point during our observation period. The reform reduced the number of school years until *Abitur* graduation from 8 to 9 years *on academic-track schools only*: before introduction of the reform, the expected year of graduation equals $c + 9$ irrespective of school type. After introduction of the reform, the expected year of graduation equals $c + 8$ for students on academic-track schools, but $c + 9$ for students on comprehensive and vocational high schools. We take the staggered introduction of the state-level reform into account when constructing attainment rates (which is possible because our graduation count data is school type-specific). Further, we include dummies for the final G9 cohort and the first G8 cohort in all our regressions to control for any influence the reform might have had on (the timing of) graduating. Information on state-specific reform timing comes from Marcus and Zambre (2019).

[24] One might particularly worry about non-standard school trajectories among graduates from vocational high schools (VocHS). VocHS are supposed to only enrol *Abitur*-bound students *directly after they complete grade 10*, not students who first complete an apprenticeship before returning to school to complete *Abitur*. Such so-called 'second-chance' students are supposed to enrol in *Abendgymnasien* and *Kollegs* instead. Non-compliance with these recommendations means that (in some states) VocHS graduation counts might be inflated by second-chance graduates. Because these students are not directly identifiable, as a workaround, we scale down the count of *Abitur* graduates from VocHS by the state-specific share of VocHS students aged above 20 who are likely to be second-chance graduates, using the state-specific age distribution of VocHS enrollees retrieved from Statistisches Bundesamt (2001b). The mean of this share is 0.13 with a standard deviation of 0.07.

school after tenth grade.[25] If they enrol in a different region for their upper-secondary education, we *underestimate* the graduation rate in the student's original ('sending') region and *overestimate* it in the final ('receiving') region because these students are counted toward the former's cohort size but the latter's graduates. This is innocuous for identification if regional mobility is random. However, it is likely to correlate with our treatments because students from regions with low supply of upper-secondary schools are more likely to enrol in upper-secondary education in a neighbouring region with a higher supply of these schools than *vice-versa*, potentially leading us to overstate the association between upper-secondary school supply and graduation rates. This is a particular concern for vocational high schools, which enrol all their students after tenth grade. As between-region mobility is less likely the higher the level of regional aggregation, it is thus key to probe the robustness of our estimated school supply-attainment associations to different levels of aggregation. However, it should be noted that most counties have self-contained school systems, so that this problem should only be relevant for a small number of bordering rural and urban counties and virtually absent at the level of planning regions, which aggregate over cities and their sub-urban/rural surroundings.[26]

Our final analysis dataset covers the secondary school entry cohorts 1995 to 2007 for 379 counties (91 planning regions, 38 provinces, and 15 states, respectively).[27] The reduced number of regions arises because we are forced to exclude the state of Brandenburg due its incomparable statistical definition of comprehensive schools.[28] In addition to attainment rates and the three measures of school supply, our dataset contains a number of regional socio-demographic indicators, which we use to proxy demand-side factors of educational expansion: unemployment rate, average household income, average population age, share of foreign students, share of college-educated workers, and urbanisation level. We retrieve these variables from the INKAR data base (German Federal Institute for Research on Building, Urban Affairs and Spatial Development)

[25] If students enrol in a secondary school outside of their home region before grade 7 (e.g., with the start of secondary school) there is no measurement problem because cohort size and graduation counts would be recorded in the same region.

[26] The problem arises because some large cities and their surrounding sub-urban/rural areas are classified as separate counties despite potentially having non-negligible overlap in their school systems: for example, students from the suburbs might attend *lower*-secondary school in their rural county but switch to, say, a vocational high school in the city for *upper* secondary education. (Again, note that students who directly enrol in an urban secondary school after primary school—the much more common phenomenon—do not cause a measurement problem because they would count towards the city's cohort size.) Once we aggregate to the level of planning regions, which aggregate over large cities and their surrounding sub-urban/rural areas, the problem disappears. For the county-level analysis, in the case of three Bavarian rural-urban county pairs (*Bamberg* and *Bamberg, Stadt*; *Schweinfurt* and *Schweinfurt, Stadt*; *Würzburg* and *Würzburg, Stadt*) the problem was particularly severe as the three respective rural counties did not have a single upper secondary school in some of the (early) years. For those three cases, we combined each pair into one 'county' (thus reducing the number of total counties by three). We decided against extending this method of consolidation to other cases where supply levels in the rural county are low (but not zero), because it would have involved arbitrary decisions about which counties to merge. Instead we present results at pre-defined higher levels of aggregation, which we deem a more transparent approach.

[27] As expected graduation is (up to) nine years after secondary school entry, the 2007 entry cohort is latest for which we still observe the relevant graduation numbers in our data which runs up to 2016. An advantage of stopping with the 2007 entry cohort is that this is the last cohort not affected by a wave of comprehensive school reforms in 2010 (in Berlin, Bremen and Hamburg), which relabelled many schools as comprehensive with sometimes unclear changes at the school level.

[28] In 2005 Brandenburg adopted the standard definition of comprehensive schools. Accordingly, its data is incomparable to the rest of Germany in the years before 2005, but also internally incomparable over time.

at the county level and aggregate them to higher levels by constructing population-weighted averages over the counties that constitute each higher-level regional unit.[29]

3.5 Empirical Strategy

As described above, the aim of this paper is twofold: first, to describe regional heterogeneities in educational expansion and local school supply in Germany and, second, to assess if the local availability of alternative pathways to Abitur in the form of comprehensive and vocational high schools increases educational attainment, i.e., whether the increase in the supply of these new school types contributed to educational expansion—and regional heterogeneity therein.

With these aims in mind, we first provide a solely descriptive graphical analysis of the regional dispersion and historical evolution of *Abitur* attainment and school supply in Germany by means of heatmaps and scatterplots.

We then begin our investigation of the relationship between school supply and attainment by exploiting the large cross-sectional variation in the types of schools locally available to students by means of OLS regressions for the *Abitur* attainment rate, Y_{ic}, of entry cohort c from region i on our three measures of school supply, S_{ic}^{j} for $j \in \{V, C, A\}$:

$$Y_{ic} = \alpha + \beta_C S_{ic}^{C} + \beta_V S_{ic}^{V} + \beta_A S_{ic}^{A} + \theta_c + \varepsilon_{ic}. \tag{3.1}$$

The presence of cohort fixed-effects, θ_c, ensures that our coefficients of interest, β_C, β_V and β_A, are identified only from contemporaneous *between-region* variation in the supplies of comprehensive, vocational, and academic-track schools, and not from secular trends in attainment and school supply. Here, and throughout the paper, we cluster standard errors at the level of regions.

Pending further investigation, the cross-sectional associations estimated in equation (3.1) hold merely descriptive value because all three school supply measures are potentially endogenous: local demand-side factors connected to educational aspirations, as well as other (state-level) educational policies, might affect both the local availability of different school types and attainment rates. To gauge the extent of confounding in the estimated cross-sectional associations, we probe their sensitivity to conditioning on (i) the extensive list of socio-demographic controls described in the previous section and (ii) state fixed-effects so that our coefficients of interest are identified only from within-state variation in supply, holding state-level factors (e.g., educational policies) constant. Especially the *differential* sensitivity of the three school supply effects to this exercise can be informative of the relative levels of confounding concerning each coefficient. Still, it is clear that conditioning on these observables alone cannot provide convincing estimates of causal supply-side effects.

[29] Among the controls, 'urbanisation level' is the only discrete variable (1=large city, 2=urban county, 3=semi-rural county, 4=rural county). To aggregate it to higher levels, we take the population-weighted average of the *numerical values* (as we do for the other variables), so that at levels of aggregation higher than the county urbanisation level enters as a *continuous* measure.

Consequently, we turn to two-way fixed-effects (TWFE) models in an effort to circumvent the endogeneity of supply, which add region-fixed-effects, δ_i, to equation (3.1):

$$Y_{ic} = \alpha + \beta_C S_{ic}^C + \beta_V S_{ic}^V + \beta_A S_{ic}^A + \delta_i + \theta_c + \tilde{\varepsilon}_{ic}. \tag{3.2}$$

As fixed-effects absorb all permanent differences between regions and cohorts, in equation (3.2) the supply-side effects are identified only from *between-region cross-cohort* variation in school supply, i.e., differences in the size of adjacent cohorts and the regionally differential expansion of these school types over time. Even though the introduction of new schools is not randomised, from the perspective of students the exact timing of when a new school starts operating might in fact be close to random, thus offering much more plausibly exogenous variation in supply than the purely cross-sectional variation exploited above. Acknowledging the possibility that regions with expanding school supply might have had differential attainment growth already prior to the supply expansion, as a robustness check we add region-specific linear cohort trends in equation (3.2). Note that a caveat of the TWFE model is that between-region cross-cohort treatment variation might be much more limited, so that, in statistical terms, we face a bias-variance trade-off when deciding between models (3.1) and (3.2).

Finally, we exploit the discrete nature of the introduction of new comprehensive and vocational high schools as 'natural experiments' in a difference-in-differences (DD) design. That is, we compare how *Abitur* attainment changed in regions that introduced their first comprehensive or vocational high school with the change in regions that continued to have no such alternative pathway to *Abitur*. Under the assumption of parallel counterfactual attainment trends between those regions, DD identifies the causal effect of the *introduction* of these schools (i.e., the extensive margin effect of having at least one such school *vs.* none, ignoring the level of supply). While arguably the cleanest research design in our setting—it imposes no parametric assumptions on treatment effects and uses only sharp jumps in supply for identification—treatment variation is most limited. To estimate the DD, we construct a dichotomous treatment variable indicating the presence vs. absence of the school type in question, $D_{ic}^j \in \{0,1\}$ for $j \in \{V,C\}$, and estimate TWFE models of the form:[30]

$$Y_{ic} = \alpha + \beta_j D_{ic}^j + \delta_i + \theta_c + \ddot{\varepsilon}_{ic}, \text{ for } j \in \{C,V\}. \tag{3.3}$$

To avoid hard-to-interpret comparisons between newly-treated and always-treated units, for this exercise we exclude regions where the school type in question was present throughout the entire observation period. As TWFE-DD models still make undesirable comparisons between early-treated and later-treated units when treatment timing varies between units (Goodman-Bacon, 2021), we also present event-study estimates, which additionally allow us to inspect pre-trends and treatment effect dynamics.[31]

[30] Note that academic-track schools are present everywhere at all times, so this design only works for the two more novel school types.

[31] That is, we estimate $Y_{ic} = \alpha + \delta_i + \theta_c + \sum_{k=-\infty}^{\infty} \beta_k^j \mathbb{I}[K_{ic}^j = k] + \varepsilon_{ic}$ for $j \in \{V,C\}$, where K_{ic}^j indicates 'relative time' (measured in cohorts) with respect to the introduction of school type j in region i (i.e., $K_{ic}^j = 0$ indicates the

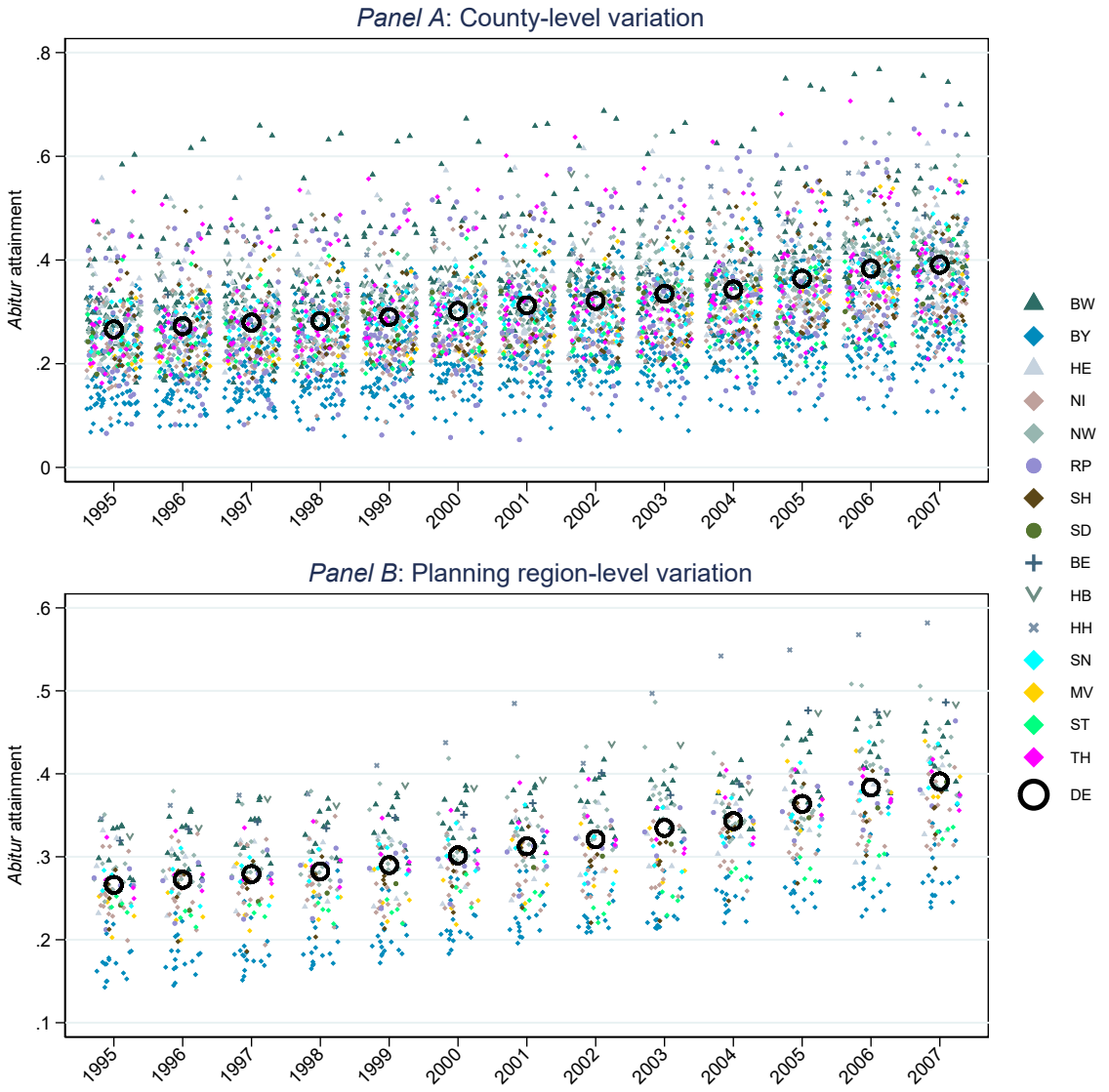


Figure 3.2. Share of *Abitur* attainment by region and entry cohort.

Notes: Each point identifies a region×cohort pair: panel A at the county, panel B at the planning region level. A point's color/symbol identifies the region's state, as indicated in the legend on the right: BW=Baden-Wurttemberg, BY=Bavaria, HE=Hesse, NI=Lower Saxony, NW=North Rhine Westphalia, RP=Rhineland Palatina, SH=Schleswig-Holstein, SD=Saarland, BE=Berlin, HB=Bremen, HH=Hamburg, SN=Saxony, MV=Mecklenburg-Vorpommern, ST=Saxony Anhalt, TH=Thuringia. The hollow circle indicates the national average. *Source:* Own calculations based on administrative education statistics of the federal states (see section 3.4).

3.6 Results

3.6.1 Descriptive Patterns

We begin by depicting the regional heterogeneity and evolution of educational attainment in Germany: Figure 3.2 presents a scatterplot of *Abitur* attainment at the county (panel A) and

cohort of introduction). Given the presence of never-treated regions, the fact that we fully saturate the model in terms of relative time dummies ensures that the event-study estimates, $\{\beta_k^j\}$ for $k \in \{k_{\min}, \ldots, k_{\max}\}$, do not involve comparisons of earlier- vs. later-treated units.

planning region (panel B) level across cohorts.[32] The national average of *Abitur* attainment (depicted by black hollow circles) monotonically increased from less than 3 out of 10 students in the cohort that entered secondary school in 1995 to almost 4 out of 10 in the 2007 entry cohort, showing that educational expansion is a relevant phenomenon during our observation period.

Figure 3.2 reveals large regional heterogeneity in attainment both across and within states. In addition to well-known between-state differences, such as comparatively low attainment rates in Bavaria and comparatively high rates in the city states (Berlin, Bremen, and Hamburg), the figure also shows that a sole focus on the state level (as is common in the German educational literature) misses a lot of variability at lower levels. In particular the differences between counties are very stark, with attainment rates ranging from only slightly above 10% to close to 80% in later cohorts. As explained in section 3.4, differences between counties might be slightly overstated due to measurement issues, so that the most extreme values should be interpreted with some caution. However, even the level of planning regions, which are considerably larger than counties—in fact, clearly too large to depict regional inequality granularly—and thus do not suffer from these measurement issues, the large regional variation between and within states is apparent. Importantly, both panels in Figure 3.2 show that regional dispersion of attainment is increasing over time. These growing regional differences might derive from a divergence in the demands of local labour markets, from processes of residential sorting or from a divergence in the quality and quantity of educational institutions and are suggestive of increasingly unequal educational opportunities in Germany.

Figure 3.3 shows similar scatter plots for the evolution of supply of schools with upper-secondary provision at the county level: vocational high schools in panel A, comprehensive schools in panel B, and academic-track schools in panel C. The figure shows that most of the variation in school supply is cross-sectional: compared to attainment, there is less evolution between cohorts, while between-*state* differences are somewhat more pronounced. Nevertheless, the considerable share of counties that does not have non-traditional school types at all decreases over time, mainly due to the introduction of vocational high schools whose supply is slowly increasing over time. Unfortunately, the supply of comprehensive schools is rather constant throughout our observation period, limiting the potential for identification through TWFE estimation.

One pattern that jumps out in Figure 3.3 particularly for academic-track schools (though it is visible for all three school types) is the sudden and stark increase in supply levels in East German states (depicted in brighter colors) starting with the 2000 entry cohort.[33] This is due to a sudden drop in birth rates in East Germany after reunification in 1990 ('*Geburtenknick*') causing a drastic decrease in (secondary school) entry cohort sizes from 2000 onward (as students generally enter secondary school when they are 10 years old). Hence, these increases in supply are driven by falling student numbers (i.e., decreases in the denominator of our outcome variable), rather

[32] Appendix Figure A3.2 visualises the county-level variation by heatmaps for selected cohorts.

[33] This is even more apparent in Appendix Figure A3.3, which plots the evolution of all three school supply measures at the state level.

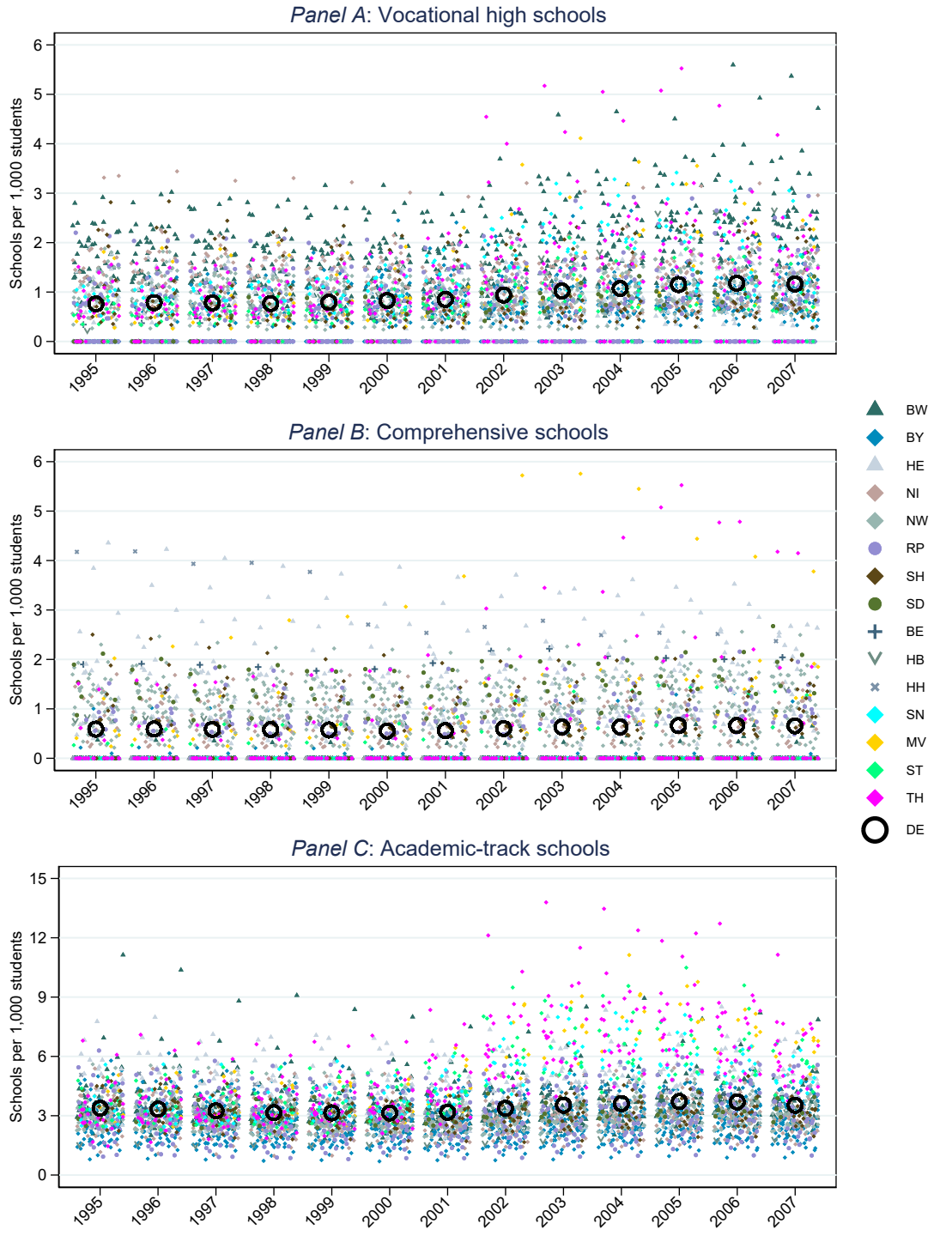


Figure 3.3. School supplies by county and entry cohort.

Notes: Each point identifies a county×cohort pair. Colours/symbols identify states. Hollow circles identify the national average.
Source: Own calculations based on administrative education statistics of the federal states (see section 3.4).

than by the introduction of new schools (i.e., increases in the numerator) as in (the vast majority of) West German states. While part of this reflects actual increases in relative supply (i.e., an increase in the number of available slots per student), part of it is measurement error, as schools adjusted and reduced the number of available slots (and consolidated over the following years

Table 3.2. Cross-sectional OLS regressions at the county level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Vocational high school	0.070***			0.055***	0.053***	0.043***	0.049***	0.051***
	(0.006)			(0.005)	(0.005)	(0.005)	(0.004)	(0.004)
Comprehensive school		0.037***		0.025***	0.034***	0.014**	0.006	0.008
		(0.007)		(0.006)	(0.005)	(0.007)	(0.006)	(0.005)
Academic-track school			0.033***	0.021***	0.026***	0.031***	0.013***	0.027***
			(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.004)
Weighted by cohort size					✓			
State fixed effects						✓		
Demand side controls							✓	✓
Excluding East Germany								✓
Cohort fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
R^2	0.402	0.195	0.342	0.519	0.560	0.606	0.700	0.774
N counties	379	379	379	379	379	379	379	322
N observations	4927	4927	4927	4927	4927	4927	4927	4186

Notes: This table presents results from county-level OLS regressions with standard errors clustered at the county level. The three school supply variables measure the number of schools per 1,000 students. Demand side controls are: unemployment rate, average household income, average age, share of foreign students, college-educated share of workers and level of urbanization. The East Germany excluded in column 8 are: Mecklenburg-Vorpommern, Saxony, Saxony-Anhalt and Thuringia (Brandenburg is excluded throughout and Berlin is retained throughout). All regressions include dummies for the double graduation cohorts and the final pre-reform cohorts resulting from the G8 reform. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Source: Own calculations based on administrative education statistics of the federal states (see section 3.4).

through school closures and mergers, as is visible from the slowly recovering supply levels in the figure and Appendix Figure A3.3). Accordingly, it is somewhat unclear in how far this variation is useful or harmful for our purposes. Consequently, we present all analyses including and excluding the East German states.

3.6.2 Cross-Sectional Regressions

In Table 3.2 we present results for OLS regressions of equation (3.1) at the county level, which show that cross-sectionally all three upper-secondary school types correlate positively with attainment. Both the bivariate associations in columns 1–3, as well as the multivariate model in column 4 suggest that the relationship between local school supply and attainment is strongest for vocational high schools, with a coefficient close to (more than) double the size of that for comprehensive schools (academic-track schools). This pattern is unchanged when we weight counties by their population size to obtain 'effect' estimates representative at the student level (column 5). It is hardly altered when restricting attention to within-state comparisons through the inclusion of state fixed-effects (column 6).

In column 7, we condition on the full set of socio-demographic controls as proxies for local demand-side factors driving *Abitur* attainment: unemployment rate, average household income, average population age, share of foreign students, share of college-educated workers, and urbanisation level. The substantially increased R-squared indicates that these factors indeed explain a lot of between-county differences in attainment. Their inclusion substantially attenuates the coefficients for comprehensive and academic-track schools, suggesting that local supply of these schools is responsive to the social composition of resident families and parents' educational as-

Table 3.3. Cross-sectional OLS regressions at the planning region level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Vocational high school	0.054***			0.056***	0.050***	0.047***	0.049***
	(0.008)			(0.008)	(0.009)	(0.007)	(0.007)
Comprehensive school		0.038***		0.040***	0.040***	0.022***	0.015**
		(0.007)		(0.007)	(0.006)	(0.007)	(0.007)
Academic-track school			0.014***	0.002	0.010**	0.002	0.024**
			(0.005)	(0.003)	(0.005)	(0.003)	(0.010)
Weighted by cohort size					✓		
Demand side controls						✓	✓
Excluding East Germany							✓
Cohort fixed effects	✓	✓	✓	✓	✓	✓	✓
R^2	0.468	0.425	0.363	0.601	0.628	0.761	0.832
N counties	91	91	91	91	91	91	75
N observations	1183	1183	1183	1183	1183	1183	975

Notes: This table presents results from planning region-level OLS regressions with standard errors clustered at the planning region level. The three school supply variables measure the number of schools per 1,000 students. Demand side controls are: unemployment rate, average household income, average age, share of foreign students, college-educated share of workers and level of urbanization. The East Germany excluded in column 8 are: Mecklenburg-Vorpommern, Saxony, Saxony-Anhalt and Thuringia (Brandenburg is excluded throughout and Berlin is retained throughout). All regressions include dummies for the double graduation cohorts and the final pre-reform cohorts resulting from the G8 reform. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Source: Own calculations based on administrative education statistics of the federal states (see section 3.4).

pirations for their children. Consequently, the cross-sectional supply-attainment correlations for these school types appear to be largely spurious. The coefficient for vocational high schools, in contrast, is left almost completely unchanged. This is strong initial evidence that the association between the local supply of vocational high schools and attainment reflects not just confounding but also causal supply-side effects.

As explained above, it is crucial to check whether these results are robust to two challenges in particular. First, it is unclear whether large increases in our measures of school supply in East Germany after 2000 reflect actual expansions in the number of available slots per student, because they are driven mainly by demographic decline instead of the introduction of new schools. Second, students who attend upper-secondary school in another county than where they attended lower-secondary school might cause systematic measurement error that upward-biases our estimated school supply-attainment correlations. We test against these issues by excluding former GDR states from the sample and by repeating the analysis at the level of planning regions, where post-grade 7 student mobility should be largely absent. The final column of Table 3.2 shows that excluding East German states leaves the coefficients for comprehensive and vocational high schools unchanged but increases the one for academic-track schools, confirming that the demographically driven supply increases in East Germany correlate much less with attainment than the remaining variation. Nevertheless, the coefficient for academic-track schools remains only half the size of that for vocational high schools. Table 3.3 repeats the complete analysis at the planning region level (except for state fixed-effects, because most states comprise too few planning regions for this analysis to be meaningful). It confirms that all results carry over to

Table 3.4. Two-way fixed-effects OLS regressions at the county level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Vocational high school	0.013***			0.017***	0.012***	0.015***	0.017***	0.020***	0.020***
	(0.002)			(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)
Comprehensive school		0.001		0.001	-0.006	-0.002	0.011	0.013**	0.005
		(0.009)		(0.008)	(0.016)	(0.005)	(0.008)	(0.005)	(0.010)
Academic-track school			-0.003*	-0.005***	-0.004***	-0.007***	0.005	0.010***	0.005
			(0.001)	(0.001)	(0.002)	(0.001)	(0.003)	(0.003)	(0.005)
Weighted by cohort size					✓				✓
County linear trends						✓		✓	✓
Excl. East Germany							✓	✓	✓
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓
Cohort fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓
R^2	0.950	0.949	0.949	0.951	0.950	0.970	0.957	0.973	0.972
N clusters	379	379	379	379	379	379	322	322	322
N observations	4927	4927	4927	4927	4927	4927	4186	4186	4186

Notes: This table presents results from county-level OLS regressions with standard errors clustered at the county level. The school supply variables are measured as the number of schools per 1,000 students. The East Germany excluded in column 8 are: Mecklenburg-Vorpommern, Saxony, Saxony-Anhalt and Thuringia (Brandenburg is excluded throughout and Berlin retained throughout). All regressions include dummies for the double graduation cohorts and the final pre-reform cohorts resulting from the G8 reform. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Source: Own calculations based on administrative education statistics of the federal states (see section 3.4).

this substantially higher level of regional aggregation, thus alleviating concerns about student mobility.[34]

3.6.3 Two-Way Fixed-Effects

We now turn to estimating the TWFE model of equation (3.2). As explained above, this should bring us closer to identifying causal effects of school supplies on attainment because TWFE only relies on between-region cross-cohort variation for identification. Because this variation might be small compared to cross-sectional differences—with implications for statistical power and the interpretation of effect sizes—we begin by inspecting the remaining treatment variation (Mummolo and Peterson, 2018). Appendix Figure A3.4 plots the distribution of our three treatments before and after conditioning on county fixed-effects (FEs) in addition to cohort FEs, separately for the sample with and without East Germany. For vocational high schools, a substantial share of variation is left after conditioning (the post-FE standard deviation (SD) equals 0.302—about 40% of the original variation), whereas for comprehensive schools much less variation is available (the post-FE SD equals 0.195—about 28% of the original variation). This confirms the pattern suggested by Figure 3.3. Whether East Germany is included or not matters little for the available identifying variation for vocational high schools (the post-FE SD is reduced by 14%), somewhat for comprehensive schools (the, already small, post-FE SD is reduced by 38%) and a lot for academic-track schools, where most of the cross-cohort variation

[34] Appendix Table A3.1 repeats the analysis at even higher levels of regional aggregation—at the province (NUTS-2) and state (NUTS-1) level—with identical conclusions.

Table 3.5. Two-way fixed-effects OLS regressions at the planning region level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Vocational high school	0.008*			0.011***	0.004	0.017***	0.007	0.028***	0.027***
	(0.004)			(0.004)	(0.005)	(0.006)	(0.005)	(0.004)	(0.005)
Comprehensive school		0.001		0.002	-0.014	0.001	-0.028	0.000	-0.003
		(0.016)		(0.019)	(0.022)	(0.006)	(0.019)	(0.012)	(0.015)
Academic-track school			-0.001	-0.003	-0.002	-0.010***	0.020***	0.006	0.002
			(0.002)	(0.003)	(0.003)	(0.003)	(0.006)	(0.007)	(0.010)
Weighted by cohort size					✓				✓
County linear trends						✓		✓	✓
Excluding East Germany							✓	✓	✓
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓
Cohort fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓
R^2	0.949	0.948	0.949	0.949	0.950	0.972	0.960	0.977	0.974
N clusters	91	91	91	91	91	91	75	75	75
N observations	1183	1183	1183	1183	1183	1183	975	975	975

Notes: This table presents results from planning region-level OLS regressions with standard errors clustered at the planning region level. The school supply variables are measured as the number of schools per 1,000 students. The East Germany excluded in column 8 are: Mecklenburg-Vorpommern, Saxony, Saxony-Anhalt and Thuringia (Brandenburg is excluded throughout and Berlin retained throughout). All regressions include dummies for the double graduation cohorts and the final pre-reform cohorts resulting from the G8 reform. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Source: Own calculations based on administrative education statistics of the federal states (see section 3.4).

is due to East German states.[35] Hence, especially for academic-track schools it is important to compare results between the two samples.

Table 3.4 presents the estimation results for the TWFE model at the county level. The first four columns, which are identical to those in Table 3.2 except for added county fixed-effects, show that the coefficients for each school type are substantially reduced compared to the cross-sectional case. However, consistent with the cross-sectional evidence, we continue to find evidence for a positive supply-side effect on attainment for vocational high schools. This effect is highly significant and remarkably robust to a range of model permutations, such as weighting by cohort size, the inclusion of county-specific linear cohort trends, and excluding East Germany from the sample (as well as all combinations of these permutations).

The effect of comprehensive school supply is close to, and statistically indistinguishable from, zero across models when East Germany is included. When the East German states are excluded the coefficient turns positive but remains small and reaches statistical significance only in column 8. As anticipated, the estimated effect of academic-track school supply is most sensitive to the exclusion of East German states: with those states included in the sample, the estimated effected is consistently negative; without these states, the coefficient turns positive but remains small and reaches statistical significant only in column 8, just like comprehensive schools. Given the limited amount of variation in supply over time, especially for comprehensive schools, this is best interpreted as absence of strong evidence for a positive supply-side effect instead of strong evidence for a null effect. In any event, just like the cross-sectional results from above, these results suggest that, if there are supply-side effects for comprehensive and academic-track schools, they are only half as large as those for vocational high schools.

[35] For academic-track schools, the post-FE SD equals 0.629 (about 43% of the original variation) with East Germany but only 0.206 without (i.e., a reduction of 67%).

Table 3.6. OLS regressions for DD model at the county level.

Sample:	All counties	Excluding always-treated counties					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
***Panel A:* Vocational high school**							
Vocational h.s. (binary)	0.013***	0.020***	0.018***	0.012***	0.017***	0.014***	0.013***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)
R^2	0.950	0.933	0.945	0.962	0.943	0.967	0.971
N clusters	379	147	147	147	134	134	134
N observations	4927	1911	1911	1911	1742	1742	1742
***Panel B:* Comprehensive school**							
Comprehensive s. (binary)	-0.001	0.002	0.005	-0.007	0.006	-0.001	-0.001
	(0.006)	(0.006)	(0.004)	(0.008)	(0.006)	(0.007)	(0.005)
R^2	0.949	0.950	0.956	0.970	0.959	0.975	0.977
N clusters	379	246	246	246	200	200	200
N observations	4927	3198	3198	3198	2600	2600	2600
Specification:							
Weighted by cohort size			✓				✓
County-specific linear trends				✓		✓	✓
Excluding East Germany					✓	✓	✓
County fixed effects	✓	✓	✓	✓	✓	✓	✓
Cohort fixed effects	✓	✓	✓	✓	✓	✓	✓

Notes: This table presents results from county-level OLS regressions with standard errors clustered at the county level. The East Germany excluded in column 8 are: Mecklenburg-Vorpommern, Saxony, Saxony-Anhalt and Thuringia (Brandenburg is excluded throughout and Berlin retained throughout). All regressions include dummies for the double graduation cohorts and the final pre-reform cohorts resulting from the G8 reform. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Source: Own calculations based on administrative education statistics of the federal states (see section 3.4).

As before, we repeat all regressions at the level of planning regions to confirm that our estimates are not driven by student mobility-induced measurement error. Reassuringly, the estimates for the effect of vocational high schools from this exercise, presented in Table 3.5, closely reproduce those at the county level (with slightly smaller effect sizes).[36] The estimates for comprehensive and academic-track schools become even less suggestive of positive effects. Altogether, the TWFE estimates suggest that supply increases of vocational high schools increase *Abitur* attainment rates, whereas for comprehensive and academic-track schools we cannot reject the null hypothesis of no supply-side effects.

3.6.4 Difference-in-Differences

In the final step of our analyses we exploit the sharp nature of the introduction of new schools as 'natural experiments' in a DD design to estimate the extensive margin effects of the local availability of these school types. Again, it is important to note the limited treatment variation in the case of comprehensive schools: during our observation period, 90 counties introduced vocational high schools for the first time but only 16 counties introduced comprehensive schools for the first time.

[36] Appendix Table A3.2 repeats the analysis at even higher levels of regional aggregation—at the province (NUTS-2) and state (NUTS-1) level—with identical conclusions.

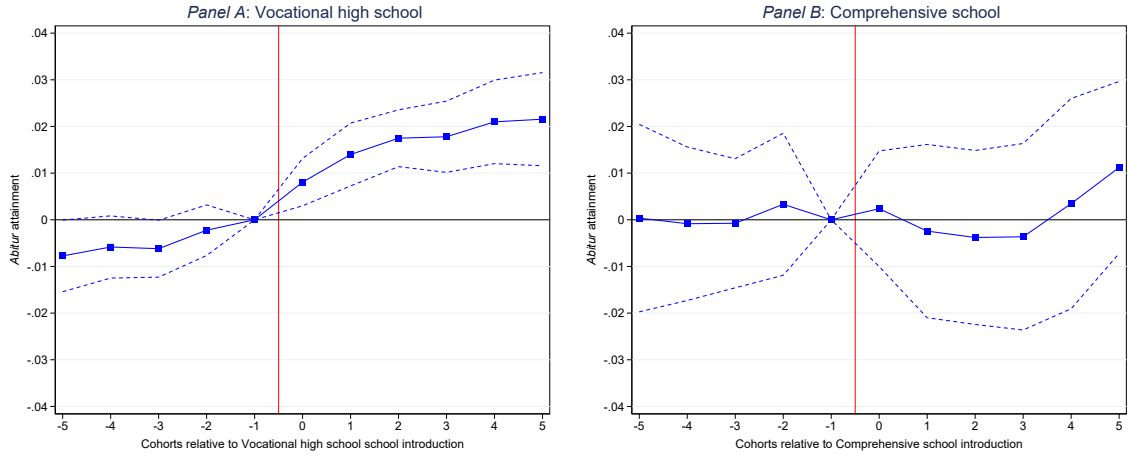


Figure 3.4. County-level event studies for introduction of vocational high schools and comprehensive schools.

Notes: The figure plots coefficients and 95%-confidence intervals (from standard errors clustered at the county level) for relative/event time indicators from -5 to +5 for event-study regressions fully saturated in event time indicators, as described in footnote 31. Differences between treatment and control units in event time -1 are normalised to zero. For panel A, we exclude 232 always-treated counties and 7 counties that switched from treated to untreated during our study period from the full sample, for an estimation sample of 50 control and 90 treatment counties. Treatments are staggered across all cohorts between 1996 and 2007, so that the first treatment lead (i.e., event time = -1) and the treatment period (i.e., event time = 0) are observed for all treated units. For panel B, we exclude 189 always-treated counties and 1 county that switched from treated to untreated during our study period from the full sample, for an estimation sample of 164 control and 25 treatment counties. Treatments are staggered across all cohorts between 1996 and 2007, except for 2004, so that the first treatment lead (i.e., event time = -1) and the treatment period (i.e., event time = 0) are observed for all treated units. *Source:* Own calculations based on administrative education statistics of the federal states (see section 3.4).

Table 3.6 presents the DD results for vocational (panel A) and comprehensive schools (panel B) estimated by the TWFE model of equation (3.3) at the county level. Column 1 uses the full sample to allow for direct comparison to the previous results. Columns 2 excludes regions that had at least one vocational/comprehensive high school throughout our study period for a cleaner comparison. The remaining columns present the same robustness checks as above. The estimates in panel A indicate that the extensive margin effect for vocational high schools is substantial: depending on model and sample definitions, the introduction of such a school when it was not previously present increases the *Abitur* attainment rate by between 1.2 and 2.0 percentage points.[37] The estimates for comprehensive schools in panel B, in contrast, show no evidence for extensive margin effects.

3.6.5 Event study

Finally, we present event-study estimates for the introduction of both school types to inspect treatment effect dynamics. Additionally, they allow us to assess the plausibility of the 'parallel trends' assumption underlying DD by inspecting if pre-trends, i.e., attainment trends before the introduction of new schools, were indeed parallel between treatment and control regions.

[37] Appendix Table A3.3 shows that these results are corroborated at the planning region level. The effect estimates are substantially smaller but note that, at higher levels of aggregation, extensive margin effects (i.e., the effect of introducing a single school) are smaller *mechanically* because the reference population is larger. In contrast to the previously used relative measures of supply, the binary treatment variable does not consider cohort sizes.

Further, it does not include earlier-treated counties in the construction of the control group for later-treated ones.

The estimation results displayed in Figure 3.4 confirm the positive effect of vocational high school introduction (panel A) and continue to show no evidence for an effect of comprehensive schools (panel B). The treatment effect of the vocational high school introduction seems to grow over time until it stabilises at around 2 percentage points. Note that panel A also shows slightly non-parallel pre-trends between treatment and control counties. Closer inspection reveals that the trends are almost perfectly parallel until two cohorts prior to the introduction when an upward trend starts to emerge. Remember that we measure cohort-specific attainment rates imperfectly as we use the *expected* rather than the actual year of graduation to count graduates. If some students actually take slightly longer to enter upper-secondary schooling (e.g., through repeating a grade or studying abroad for a year after grade 10, as is common), they would in fact profit from a vocational high school that *according to schedule* is only available to one or two cohorts after theirs. Given this imprecision in measurement, one might even expect attainment to rise at least one cohort early. As the observed deviations from parallel pre-trends are well explicable by the measurement imprecisions and small compared the deviations post-treatment, we do not believe them to seriously jeopardise the validity of the DD design in this context.

3.7 Conclusions

This paper investigated the relationship between upper-secondary attainment and local school supply in Germany's tracked school system, where the university-entrance certificate (*Abitur*) was traditionally only awarded by academic-track schools. However, the introduction and gradual expansion of vocational high schools and comprehensive schools opened up new pathways towards the *Abitur*, so that upper-secondary schooling has been partially de-tracked wherever these schools became locally available. This means that in Germany school supply has not only a quantitative but also a qualitative dimension, i.e., it depends not only on the number but also on the types of schools that are available to students. We therefore differentiate between the supply of academic-track schools, comprehensive school and vocational high school in our analysis, to shed light on the effects of this decentralised and incremental 'de-tracking at the margin'.

To observe attainment and supply levels at the local level, we marshal administrative data that covers the universe of German students, schools and graduates to construct a county-level panel of local school supply and upper-secondary attainment for 13 cohorts between 1995 and 2007. Using this novel data set, we reveal that the expansion of secondary credentials on net masks important geographical heterogeneity: *Abitur* attainment rates vary strongly between counties and these differences have been increasing over time. Similarly, the number and types of schools locally available to students differs markedly between regions, though these differences are more stable over time.

Our cross-sectional results indicate that supply levels of all three school types that award the *Abitur* strongly correlate with its attainment. However, for academic-track and comprehensive

schools, these cross-county correlations can largely be accounted for by regional differences in educational demand, as proxied by the degree of urbanisation and the social composition of resident families. Consistent with the absence of substantial supply-side effects, for these school types we fail to find significant coefficients in two-way fixed-effects (TWFE) and difference-in-differences (DD) models that rely only between-region cross-cohort variation in school supplies to circumvent endogeneity. With the caveat that the over-time variation for comprehensive schools is limited, we thus find no empirical support in favour of hypotheses 1 and 3.

Accordingly, academic-track and comprehensive schools seem to have primarily catered to existing educational demands instead of causally facilitating educational expansion above and beyond that. This might be less surprising for the traditional academic-track schools, which expand their slots only slowly in response to persistent demand-side pressures by parents. For comprehensive schools, however, this result stands in contrast to the weight these schools have received in the public debate. In actual fact, their local availability seems to have mainly facilitated already ongoing processes of educational expansion, but had these schools not been locally available, current comprehensive school graduates would have found other ways to reach the *Abitur* (i.e., they would have enrolled in an academic-track school straight after primary school or transferred to it after lower-secondary school). The reason for the absence of a substantial general attainment effect of comprehensive schools might be their heterogeneity: in many places, these schools effectively function as substitutes for intermediate- and low-track schools, rather than as institutions catering to students with higher aspirations.

For vocational high schools, in contrast, we find clear evidence of a positive supply-side effect on attainment. The cross-sectional supply-attainment association is robust to controlling for various socio-demographic controls and state fixed-effects, as well as weighting for population size and aggregating to higher regional levels. The TWFE and DD regressions, across various model specifications and sample permutations, as well as the event-study analysis confirm this result and, thus, hypothesis 2. Our estimates suggest that, at the extensive margin, the introduction of a single vocational high school, when it was not previously available in a county, increases attainment rates by about 1.5 percentage points. Across extensive and intensive margins, a supply increase of one school per 1,000 students (corresponding to an increase of about one slot per 100 students) increases attainment rates by roughly the same amount.

This implies that, in contrast to traditional academic-track schools and comprehensive schools, vocational high schools induce a group of students to take up (and complete) academic upper-secondary education that otherwise would not have done so. This result is especially noteworthy because, unlike previous research (e.g., Schindler, 2017), we adopt a restrictive definition of upper-secondary attainment that focuses only on the university-entrance certificate and excludes certificates that grant access to the vocational, second-tier sector of higher education. Despite having received relatively little attention in the public (and academic) debate, vocational high schools have thus contributed to educational expansion through a process of 'inclusion' (cf., Arum *et al.*, 2007; Schindler and Bittmann, 2021).

While the reasons for the success of vocational high schools are probably multifold, we suggest that a crucial factor resides in their hybrid nature: because they combine academic and specialised curricula, they might be especially attractive to risk-averse students who, in the traditional system, would have been diverted towards VET. From the perspective of students, vocational high schools might partly solve the trade-off between short-term employability and long-term human capital investments. From a societal perspective, vocational high schools might offer 'a safety net without diversion' from higher education eligibility, especially to more risk-averse low-SES students (cf., Shavit and Müller, 2000). Whether increased university eligibility translates into increased university enrolment, what the effects are on students' labour market outcomes and in how far vocational high schools really contribute to reducing social inequality are questions that need to be addressed by future research based on individual-level data.

Appendix A: Additional Tables and Figures

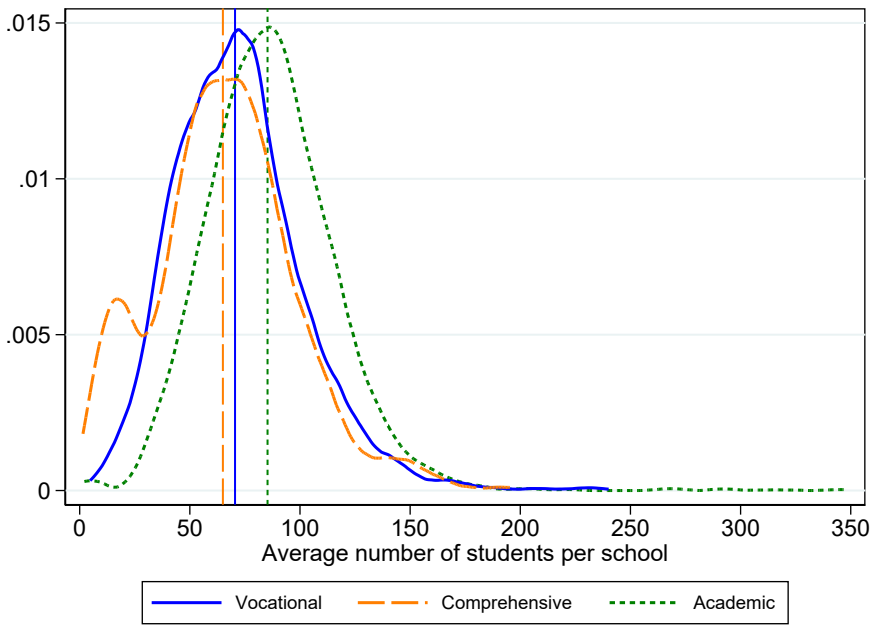
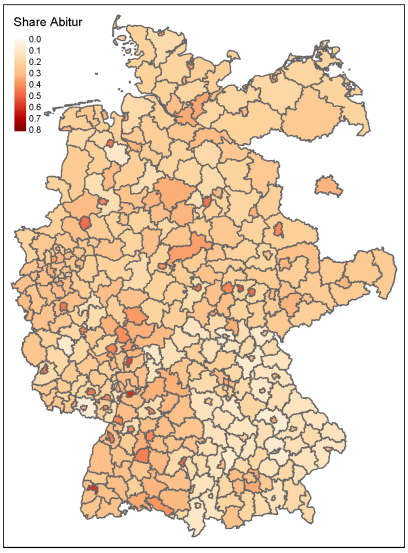


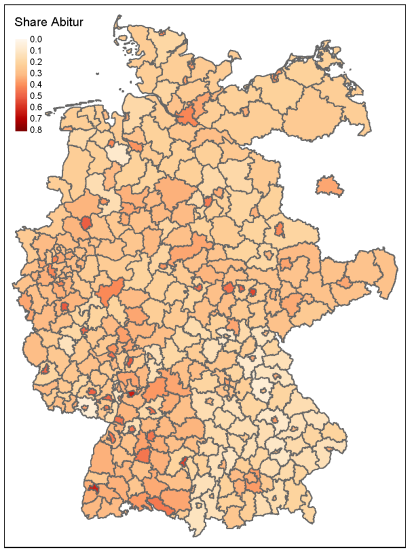
Figure A3.1. Density plot of average school size (at the county level).

Notes: The plot shows the distribution of county-level average school size at the upper-secondary level (i.e., the number of enrolled students in the first year of upper-secondary education) for the three school types of interest. Note that we plot the distribution of county-level averages instead of the school-level distribution because we observe (school type-specific) enrolment only at the county but not at the school level. The three vertical lines indicate the median of each distribution. *Source:* Own calculations based on administrative education statistics of the federal states (see section 3.4).

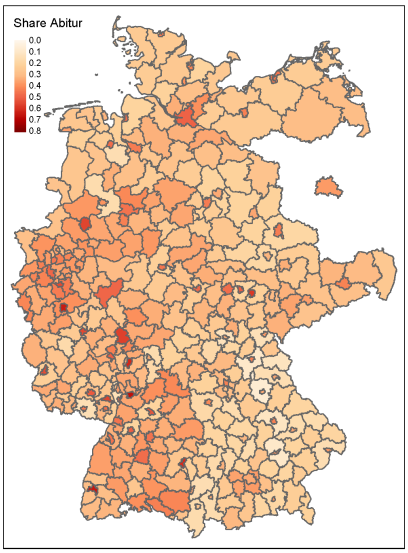
(A) *Entry cohort 1995*



(B) *Entry cohort 1999*



(C) *Entry cohort 2003*



(D) *Entry cohort 2007*

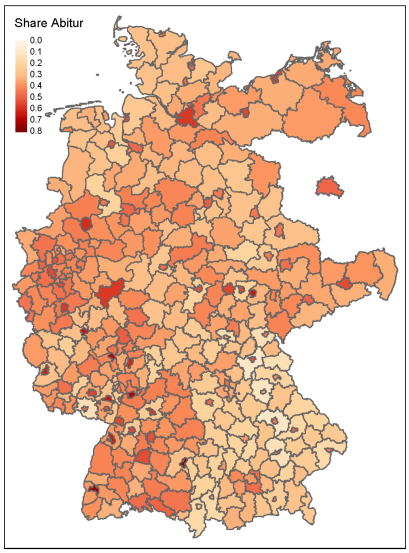


Figure A3.2. Share of *Abitur* attainment by county for four selected cohorts.

Notes: The four secondary school entry cohorts are 1995 (top-left), 1999 (top-right), 2003 (bottom-self) and 2007 (bottom-right). *Source:* Own calculations based on administrative education statistics of the federal states (see section 3.4).

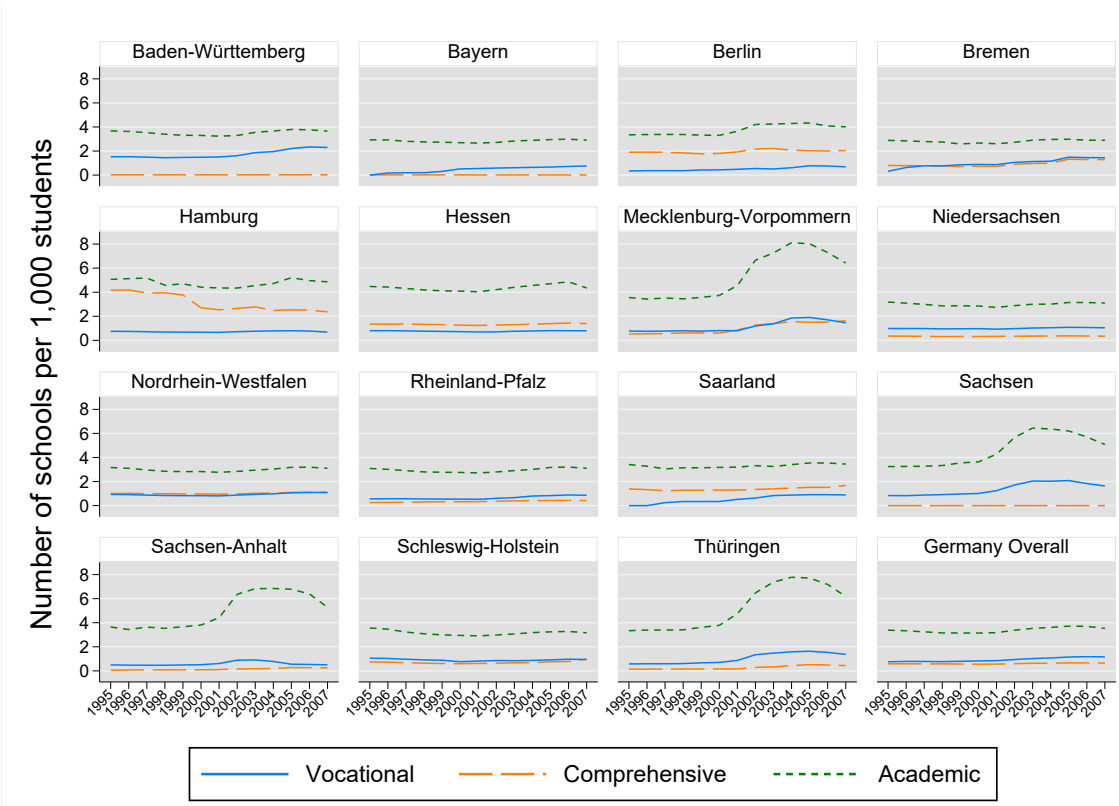


Figure A3.3. Development of state-level school supplies over entry cohorts.

Source: Own calculations based on administrative education statistics of the federal states (see section 3.4).

Table A3.1. Cross-sectional OLS regressions at different levels of aggregation.

Level of aggregation:	County (NUTS-3)		Planning region		Province (NUTS-2)		State (NUTS-1)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Vocational high school	0.055***	0.049***	0.056***	0.047***	0.068***	0.051***	0.064***	0.047**
	(0.005)	(0.004)	(0.008)	(0.007)	(0.010)	(0.008)	(0.016)	(0.020)
Comprehensive school	0.025***	0.007	0.040***	0.022***	0.049***	0.025***	0.051***	0.021*
	(0.006)	(0.005)	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.010)
Academic-track school	0.021***	0.012***	0.002	0.002	-0.006	-0.007	-0.013*	-0.009
	(0.003)	(0.002)	(0.003)	(0.003)	(0.005)	(0.005)	(0.007)	(0.007)
Demand side controls		✓		✓		✓		✓
Cohort fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
R^2	0.519	0.696	0.601	0.761	0.693	0.819	0.714	0.828
N clusters	379	379	91	91	38	38	15	15
N observations	4927	4927	1183	1183	494	494	195	195

Notes: This table presents results from OLS regressions at different levels of aggregation with standard errors errors clustered accordingly. The school supply variables are measured as the number of schools per 1,000 students. Demand side controls are: unemployment rate, average household income, average age, share of foreign students, college-educated share of workers and level of urbanization. All regressions include dummies for the double graduation cohorts and the final pre-reform cohorts resulting from the G8 reform. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Source: Own calculations based on administrative education statistics of the federal states (see section 3.4).

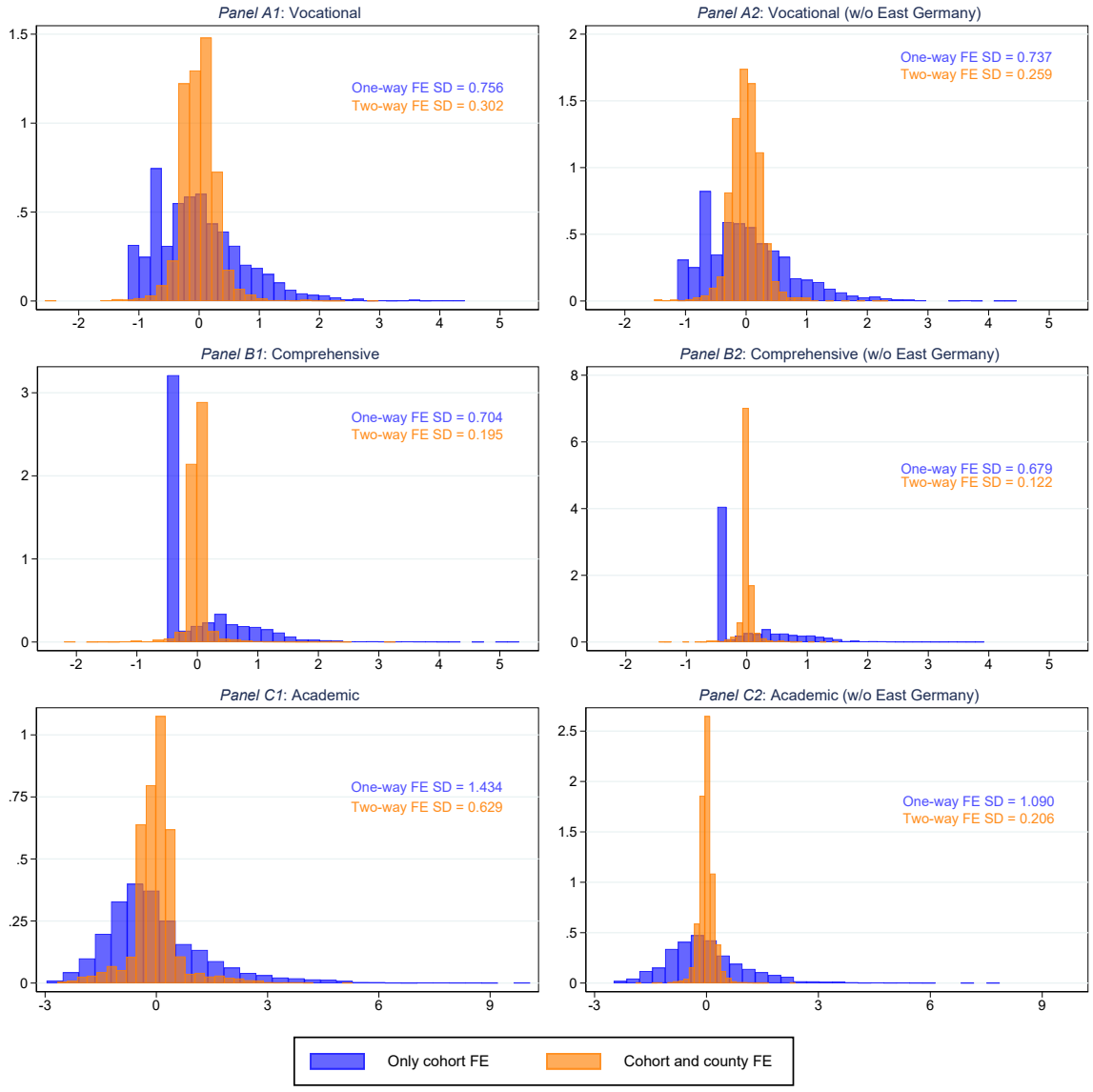


Figure A3.4. Treatment variation before and after the inclusion of county fixed effects.

Notes: The left panels refer to the full sample. The right panels to the sample excluding East German states. *Source:* Own calculations based on administrative education statistics of the federal states (see section 3.4).

Table A3.2. Two-way fixed-effects OLS regressions at different levels of aggregation.

Level of aggregation:	County (NUTS-3)	Planning region	Province (NUTS-2)	State (NUTS-1)
	(1)	(2)	(3)	(4)
Vocational high school	0.017***	0.011***	0.015*	0.021
	(0.002)	(0.004)	(0.008)	(0.018)
Comprehensive school	0.001	0.002	-0.015	-0.027
	(0.008)	(0.019)	(0.029)	(0.021)
Academic-track school	-0.005***	-0.003	-0.007	-0.004
	(0.001)	(0.003)	(0.005)	(0.006)
Spatial unit fixed effects	✓	✓	✓	✓
Cohort fixed effects	✓	✓	✓	✓
R^2	0.951	0.949	0.953	0.952
N clusters	379	91	38	15
N observations	4927	1183	494	195

Notes: This table presents results from OLS regressions at different levels of aggregation with standard errors errors clustered accordingly. The school supply variables are measured as the number of schools per 1,000 students. All regressions include dummies for the double graduation cohorts and the final pre-reform cohorts resulting from the G8 reform. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Source: Own calculations based on administrative education statistics of the federal states (see section 3.4).

Table A3.3. Difference-in-differences for Vocational introduction at the planning region level.

Sample:	All counties	Excluding always-treated counties					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Vocational (binary)	0.005*	0.006**	0.007**	0.005*	0.006**	0.005*	0.004*
	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.002)	(0.002)
Weighted by cohort size			✓				✓
County-specific linear trends				✓		✓	✓
Excluding East Germany					✓	✓	✓
County fixed effects	✓	✓	✓	✓	✓	✓	✓
Cohort fixed effects	✓	✓	✓	✓	✓	✓	✓
R^2	0.949	0.974	0.980	0.985	0.974	0.985	0.989
N clusters	91	19	19	19	19	19	19
N observations	1183	247	247	247	247	247	247

Notes: This table presents results from planning region-level OLS regressions with standard errors clustered at the planning region level. The East Germany excluded in column 8 are: Mecklenburg-Vorpommern, Saxony, Saxony-Anhalt and Thuringia (Brandenburg is excluded throughout and Berlin retained throughout). All regressions include dummies for the double graduation cohorts and the final pre-reform cohorts resulting from the G8 reform. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Source: Own calculations based on administrative education statistics of the federal states (see section 3.4).

Chapter 4

Labour Market Returns to Vocational Education in the Presence of Multiple Alternatives*

Abstract

Many countries consider expanding vocational secondary education reduce skill shortages and unemployment among non-college-bound youths. However, critics fear this could divert students from academic routes, whose focus on general, instead of occupation-specific, skills might be better suited for today's rapidly changing labour markets. This paper studies labour market returns to vocational education in England, where students choose between a vocational track, an academic track and quitting education at age 16. Identification is challenging, first, because self-selection is strong and, second, because students' next-best alternatives are unknown. Our research design leverages multiple instrumental variables to estimate margin-specific treatment effects, i.e., causal returns to vocational education for students at the margin between vocational and academic education and, separately, for students at the margin between vocational and no post-16 education. We use linked administrative education and earnings data covering the universe of students in public schools. Identification comes from plausibly exogenous variation in distance to the nearest vocational provider conditional on distance to the nearest academic provider (and *vice-versa*), while controlling for granular student-, school- and neighbourhood-level characteristics. We find that the vast majority of marginal vocational students are at the margin with academic education (instead of no further education), so that the first-order effect of expanding vocational-track access is diversion from the academic track. Diversion leads to losses in earnings at age 29 of about 9% for males and 7% for females. These effects are not driven by employment but due to wages (or working hours). A substantial part of the effect is explained by reduced university degree completion. For the few marginal students at the margin with no further education, we find tentative evidence of positive employment and earnings effects but results are imprecise and insignificant. Our findings caution against an expansion of vocational upper secondary education in England in its current form.

*This chapter is based on joint work with Guglielmo Ventura. We benefited from comments and suggestions by Sandra McNally, Andrew Eyles, Jack Mountjoy, Jan Marcus, Katharina Spieß, Aureo de Paula, Pedro Carneiro, seminar participants at CEP-LSE, CVER-LSE, University of Potsdam, BSE-Berlin, WZB Berlin, DIW Berlin and conference participants at ifo-EffEE 2021 and IWAEE Catanzaro 2021.

4.1 Introduction

Over the last decades, rising demand for high-skilled work has made it increasingly difficult for young workers without tertiary education to secure stable and well-paying jobs (Autor, 2019), widening the social chasm between university graduates and what has been referred to as 'the forgotten half' (Neumark and Rothstein, 2007). At the same time, firms in many post-industrial economies lament a lack of skilled workers in fast-growing technical and professional occupations (OECD, 2017). In terms of education policy, these developments have led to a heightened interest in improving and expanding vocational programmes in secondary education—particularly so in countries with weaker traditions of providing high-quality vocational education and training (VET) to its students, like the US and UK.[1] Often in reference to countries with strong apprenticeship systems, like Germany and Switzerland, VET is heralded as a means to relieve skill shortages, while improving the employment and earnings prospects of non-college-bound students (Fersterer *et al.*, 2008). However, considerable disagreement remains on whether vocational routes really benefit young people economically—especially in settings where firm involvement in VET through apprenticeships is rare.

This paper aims to contribute to this debate by delivering causal evidence on labour market returns to vocational upper secondary education in England, where at the age of 16, after completing compulsory schooling, students choose between a vocational track, an academic track and no upper secondary education. Our research design leverages multiple instrumental variables (IVs) to estimate margin-specific treatment effects, i.e., causal returns to vocational education for students at the margin between vocational and academic education and, separately, for students at the margin between vocational and no post-16 education. Using the fact that vocational and academic education are offered by distinct institutions, identification comes from plausibly exogenous variation in distance to the nearest vocational provider conditional on distance to the nearest academic provider (and *vice-versa*), while controlling for granular student-, school- and neighbourhood-level characteristics. Our analysis is based on linked administrative education and earnings data for the whole of England, which also allow us to closely inspect effect heterogeneity and potential mechanisms.

The debate on expanding vocational content in secondary school typically revolves around the types of skills vocational (also known as technical) and academic (also known as general) educations provide and their respective value on the labour market. The main benefit of an academic curriculum lies in equipping students with general knowledge and analytical skills that are well transferable across jobs (Goldin, 2001). Compared to the more occupation-specific curricula taught in vocational tracks, this might help students on the labour market through increased flexibility, especially in the long run (Hanushek *et al.*, 2017). This argument is particularly cogent in the face of rapidly changing labour demand due to technological change and decarbonisation. Advocates of vocational education counter that the general skills taught in academic tracks might, in fact, be too generic to be readily deployable on labour markets

[1] See, e.g., US Department of Education (2012); Jacoby and Dougherty (2016) for the US and Independent Panel on Technical Education (2016) for the UK.

unless complemented with tertiary education, which far from all students pursue (Bertrand *et al.*, 2021). In addition, the abstract nature of learning in academic tracks might disengage less academically inclined students, leaving them at risk of dropping out of secondary education altogether (Hall, 2016).

Despite the long-standing debate on these issues, there is a paucity of compelling empirical evidence on the labour market returns to vocational education. Descriptively, vocational education involves a trade-off between short-run benefits and long-run costs: studies that compare age-earnings profiles of vocationally and academically educated students generally find that initial advantages for the former have reversed by students' early thirties (Brunello and Rocco, 2017; Hanushek *et al.*, 2017; Hampf and Woessmann, 2017). While there is some consensus that vocational education indeed facilitates the school-to-work transition (Shavit and Müller, 2000), if differences in earnings at later ages are to be interpreted causally remains contested. The selection problems when comparing labour market outcomes between those with and without vocational degrees are severe because, in most settings, vocational students have much lower previous achievement and come from more disadvantaged backgrounds (Ryan, 2001).

A second challenge to estimating meaningful returns to vocational education is that they are likely to be heterogeneous, with students sorting into educational tracks with at least partial knowledge of their idiosyncratic returns (Dahl *et al.*, 2020). However, in the presence of selection on gains, the average treatment effect (ATE), which average age-earnings profiles implicitly aim to estimate, is a poor heuristic for judging economic efficiency or guiding policy, because it does not correspond to the effect for marginal students (who are more likely to respond to policy changes). Identifying returns for marginal students is particularly complicated when students face more than one alternative to vocational education because returns will depend on students' (typically unobserved) next-best alternative. For example, a student whose alternative to vocational education is quitting education and entering the labour market directly might benefit from enrolling in a vocational track to acquire additional work-related skills. At the same time, for a student whose alternative to vocational education is enrolling in the academic track, returns might be more ambiguous, especially in the longer run. Accordingly, it is crucial to identify the most relevant margins of vocational enrolment in a given setting and separately estimate margin-specific returns against the respective alternatives.

In light of this discussion, the English setting is interesting not only because it allows to investigate the effect of vocational *vs.* academic in a country that, much like the US, has a vocational education system with a historically rather poor reputation and a labour market with high wage dispersion. Moreover, because compulsory education in England ends at age 16, students choose not only between vocational and academic education but also consider leaving the education system altogether. The two corresponding margins of choice—vocational *vs.* academic and vocational *vs.* no further education—map directly into the theoretical debate about who benefits from vocational education and who does not.

However, multiple unordered education choices pose a challenge from a methodological standpoint. Standard IV based methods, such as two-stage least squares (2SLS) generally fail to

recover alternative-specific treatment effects in settings with multiple unordered treatments and effect heterogeneity (Heckman and Urzúa, 2010), even with as many instruments as treatments available (Kirkeboen *et al.*, 2016). Some studies have managed to identify alternative-specific effects by combining data from centralised admission systems where students' preference orderings can be directly observed with regression discontinuity designs exploiting admission cut-offs generated by over-subscription (Dahl *et al.*, 2020; Kirkeboen *et al.*, 2016; Silliman and Virtanen, 2022). However, this strategy is not feasible in the English setting where there is neither a centralised admissions system nor systematically oversubscribed education providers in the post-16 sector.

To overcome these challenges and identify the two alternative-specific effects of interest, net of self-selection into educational tracks, we apply an identification approach based on multiple instrumental variables (IVs) proposed by **?**. We exploit the fact that upper secondary educational tracks in England are linked to specific institutions: the vocational track is offered by vocational colleges ('Further Education Colleges') and the academic track is offered by designated academic colleges ('Sixth Form Colleges'), as well as by secondary schools. By focusing on students from schools that do not offer the academic track, i.e., on students who need to switch to a new institution at age 16 regardless of which track they choose, we can construct two alternative-specific IVs based on students' geographical proximity to the nearest vocational and the nearest academic college. Thus equipped, identification of margin-specific local average treatment effects (LATEs) is secured under intuitive assumptions by 'cross-instrumenting' students' education choices with their distance to the specific alternative of interest. Our estimated margin-specific complier treatment effects speak directly to the effects of expanding access to vocational colleges, but offer a more nuanced view than would be possible with conventional IV methods that are only able to identify the combined (net) effect of such a policy.

For estimation we leverage unique education administrative data linked to tax records that allow us to follow three full cohorts of state-school educated pupils in England through their school careers, post-compulsory education and into the labour market. We record our two main outcomes of interest, employment and annual earnings, thirteen years after the education choice, i.e., at age 29. To construct the two required distance instruments we combine geospatial information on students' home addresses with the locations of all post-16 education providers in England. To account for the non-random location of post-16 education providers, next to detailed student- and school-level controls, we directly control for distance to local economic centre and fine-grained measures of neighbourhood quality, as well as region fixed effects to ensure that we compare similar students from similar neighbourhoods who face similar labour market conditions when they make their education choices. Identification stems from conditional variation in the distance to vocational college, holding constant distance to academic college, and *vice-versa*, which is much more plausibly exogenous than distance to any kind of education provider (**?**). In balance tests we show that our instruments are empirically balanced across a range of student characteristics, including nationally administered achievement tests at age 11.

We find that the vast majority (about 85%) of marginal students, i.e., of those whose choice to enrol in vocational education is responsive to incentives like distance, is choosing between vocational and academic education, not considering the option of no post-16 education. Accordingly, any policy that seeks to increase vocational enrolment by increasing the attractiveness of the vocational track will do so mainly by diverting students from academic education. For these vocational-academic compliers, we find that vocational education substantially reduces earnings. For males, our estimate for the margin-specific LATE on annual earnings at age 29 equals -£2,500, corresponding to a earnings reduction of 9%. For females the effect is slightly smaller at -£1,500, corresponding to a 7% reduction. The negative return is not driven by extensive margin effects: we find null effects for the probability of being in sustained employment at age 29. Rather, vocational education seems to channel these students into lower-paying jobs with worse wage progression: when inspecting returns by age, we find that the earnings effect worsens close to linearly over students twenties. Together with the arguments that occupation-specific skills may depreciate faster, this trend suggests that the earnings penalty from vocational *vs.* academic education might even continue to grow as students age.

To investigate potential mechanisms behind this finding we inspect effects on various education outcomes. For males at the vocational-academic margin, the vocational track substantially harms upper secondary attainment and progression to higher education. We show that about 60% of their negative earnings effect can be explained by reduced university degree completion. In line with the smaller earnings effect, for females the effects on education outcomes are somewhat less pronounced, which seems to stem from the fact that their curriculum choice during upper secondary education is less affected by the track they attend. Only about one-third of their earnings penalty can be explained by lower degree completion.

To investigate effect heterogeneity, we use our two continuous instruments as local instrumental variables (Heckman and Vytlacil, 2005) and stratify the local IV estimates across different distances to vocational and academic colleges, thus comparing marginal students with different unobserved preferences for vocational *vs.* academic education. We confirm that students select into tracks based on their comparative advantage: returns to vocational education increase (i.e., become less negative) with students' underlying preferences for the vocational track (i.e., with distance from vocational college) and decrease with their underlying preferences for the academic track (i.e., with distance to academic college). While returns are negative for most vocational-academic compliers, those with the highest preferences for vocational education experience positive returns. This suggests that vocational education might well be beneficial for a large share of non-marginal students, but that—at least in its current form—too many students in England choose the vocational track from an efficiency standpoint, especially among males.

Results at margin between vocational and no post-16 education look very different. Point estimates indicate large returns from pursuing vocational instead of no further education both in terms employment and earnings (e.g., for males a 6 percentage point increase in sustained employment and a £3,000 increase of annual earnings). However, given the low share of compliers at this margin (about 15%), the estimates are imprecise and generally not statistically significant

(only the local IV estimates for those with the highest preference for vocational education reach marginal significance). Nevertheless, the divergent point estimates across the two margins of treatment highlight the importance of margin-specific identification in this context. Large but imprecise point estimates for the small group of compliers at the no post-16 education margin contaminate the net return for vocational education that conventional IV methods would estimate. We show that a 2SLS regression, which instruments vocational enrolment with distance to vocational college, yields a small and insignificant estimate of the net LATE, thus shrouding the large negative effects for the majority of marginal students and nurturing an ambiguous and more positive impression of vocational education in England than warranted.

This paper contributes new causal evidence to the limited literature on the labour market returns to vocational education. Most of the existing evidence comes from Continental Europe, where there is a well-established tradition of channelling secondary students into distinct academic and vocational tracks. Starting in the 1970s, a number of countries reformed their vocational tracks by giving more weight to the general curriculum. Evaluations of such reforms in the Netherlands (Oosterbeek and Webbink, 2007), Romania (Malamud and Pop-Eleches, 2011), Sweden (Hall, 2016) and Croatia (Zilic, 2018) all find no effect on students' labour market outcomes, suggesting that replacing vocational with academic content does not generate much benefits on net. Stronger even, Bertrand *et al.* (2021) consider a similar reform in Norway and find that it increased vocational track enrolment at the expense of both drop-out and academic enrolment, leading to higher earnings on net, especially for disadvantaged men. However, these studies estimate general equilibrium effects and therefore offer only limited guidance for students' education choices or policymakers' allocation problems, because they are unable to zoom in on marginal students and alternative-specific effects.

Two recent studies tackle this question more directly. First, Silliman and Virtanen (2022) exploit admission cut-offs in Finland, where students' preferences over secondary tracks are recorded under a centralised admission system. This allows the authors to estimate margin-specific effects of vocational *vs.* academic education by focusing on students marginally admitted to vocational programmes whose next choice would have been an academic programme (and *vice versa*). Interestingly, they find a positive earnings effect from vocational education that persists until at least age 33 and shows no sign of fading out. Consistent with the notion of comparative advantage, returns are larger for students who initially indicated stronger preferences for the vocational track by ranking it first. Second, Birkelund and van de Werfhorst (2022) study returns to vocational education in Denmark using a conventional IV strategy. Like in our setting, students in Denmark choose whether to enrol in an academic track, a vocational track or to quit their education. The authors instrument students education choice with peers' choices under a strong exogeneity assumption and the implicit assumption that students' unobserved preferences are weakly ordered, thus effectively ruling out the existence students at the margin between academic and leaving education. With this caveat in mind, the authors find null effects of vocational *vs.* academic education and positive effects of vocational *vs.* leaving education.

While these studies from Nordic countries paint an encouraging picture of vocational education, it is questionable whether their findings translate to countries with radically different education and labour market institutions, such as the UK or the US. Until recently, the provision of high-quality vocational education has been neglected in these countries, fuelled by (and fuelling) an ingrained distrust of vocational programmes (cf. Dewey, 1916; Oakes, 1985; Raffe *et al.*, 2001). In the UK, this is reflected in stark differences in teaching resources, peers and reputation compared to academic programmes. Moreover, the UK labour market is characterised by much more dispersed wage distributions with high returns to university degrees. Birkelund and van de Werfhorst (2022) find that the students diverted from academic to vocational education in Denmark are more likely to be employed in occupations with a lower socio-economic status (according to the ISEI classification). In the Danish setting this does not translate into lower earnings, which the authors attribute to wage compression from widespread collective bargaining and strong trade unions, none of which is present in the UK.

Direct causal evidence from the British or American context is scarce. In the US, where vocational courses (known as Career and Technical Education) are typically part of comprehensive high-school students' curriculum choice, Kreisman and Stange (2020) find a small positive association between vocational course take-up and wages. Quasi-experimental evidence using admission cut-offs to Technical High Schools in Connecticut points to positive earnings effects for men (Brunner *et al.*, 2021). In the UK, earlier studies are mostly descriptive in nature and focus on earnings differentials accruing from different types of qualifications. To the best of our knowledge, this paper is the first to consider students' enrolment in different types of post-16 institutions in England as a means to establish causal returns to vocational education.

Our findings caution against an expansion of vocational education in England in its current form as this would reduce earnings for the majority of affected students who are diverted from the academic track and associated higher earnings. Instead, policy should aim to sway students who are qualified for academic upper secondary education to enrol in the academic track. Still, our results offer at least suggestive evidence that swaying students at risk of dropping out, to instead enrol in the vocational track, might generate substantial returns. Finally, our findings suggest that the existing vocational institutions could be improved by establishing clearer pathways from vocational to higher education.

The paper is structured as follows: section 4.2 describes the institutional background and the data and presents descriptive findings. Section 4.3 explains our identification framework. Section 4.4 present our main results for labour market outcomes, including robustness checks and heterogeneity analyses. Section 4.5 seeks a better understanding of the margin-specific LATE estimates by characterising marginal students, exploring mechanisms and probing external validity. Finally, section 4.6 concludes.

4.2 Background, Data and Descriptives

4.2.1 Post-Compulsory Education in England

In England the compulsory phase of schooling lasts from age 5 to 16, during which students study a common nationally defined curriculum. To conclude their compulsory education, all students take a set of standardised exams, the General Certificate of Secondary Education (or GCSEs), in typically eight to ten subjects. Afterwards, they choose if and what type of upper secondary education to pursue. In the period considered by our analysis, students at age 16 faced three principle alternatives: (i) to pursue upper secondary education in the academic track, (ii) to pursue upper secondary education in the vocational track or (iii) to conclude their education and directly enter the labour market.

The academic track comprises two years of study towards academic qualifications known as A-Levels, which are the traditional prerequisite for university entrance. These two years of academic upper secondary education are referred to as 'sixth form'. They are offered by secondary schools that have their own sixth form (which thus provide lower and upper secondary schooling) and by designated, publicly funded Sixth Form (SF) Colleges, of which there are 94 across England. For students from schools *with* a sixth form, in the vast majority of cases choosing the academic track thus means continuing on one's secondary school. Academic-track students from schools *without* a sixth form generally enrol in SF Colleges. Admission to the academic track is decentralised. Typically, entry requires five GCSEs at grade C or higher (often those have to include Maths and English), though the exact requirements vary by institution and the specific A-Level subjects students pick (typically students pick three).

In contrast to the academic track and to other European vocational education systems, which usually comprise a limited and well-defined number of programmes, the vocational track in England is much less structured. Students can choose '*a la carte*' from a plethora of vocational qualifications that differ in level (from 1 to 3), subject and duration (from 1 to 2 years). Level 3 qualifications (*Diplomas*) are notionally equivalent to A-Levels. They have similar entry requirements in terms of GCSEs, are taught full-time, mostly as two-year courses, and count towards university admission.[2] For students that do not meet the entry requirements for Level 3, there are less demanding vocational qualifications at Level 2 (which count as equivalent to GCSEs) and Level 1 (which count as equivalent to primary school education). At each level the number of courses to choose from is very large: at Level 3 alone, there are more than 3,700 different qualifications (Hupkau *et al.*, 2017). The vast majority of vocational courses for 16–18-year-old learners are classroom-based; apprenticeships offering workplace training typically only start after completion of a classroom-based qualification.

About 80% of vocational-track students attend publicly funded Further Education (FE) Colleges, of which there are 247 across England (Hupkau and Ventura, 2017). The rest is trained by other publicly funded providers, like local authorities, or private providers, both of which

[2]Though students with vocational qualifications at Level 3 face more restrictions in terms of the degrees and universities they can apply for compared to students with A-Levels.

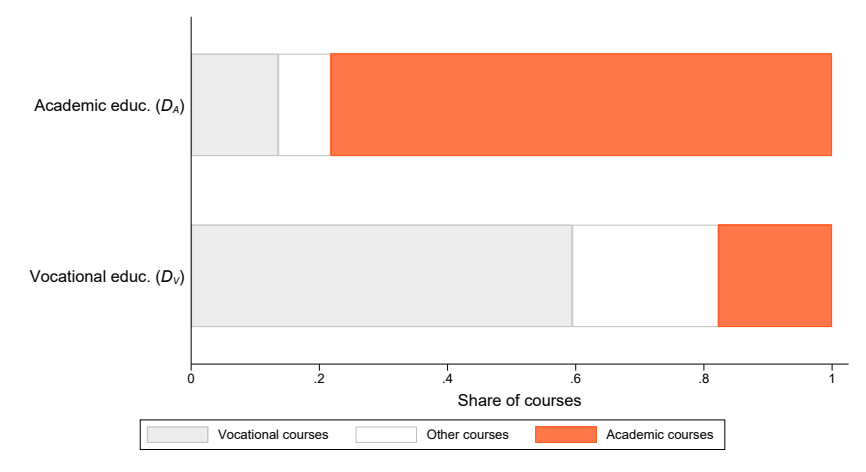


Figure 4.1. Course contents by educational track.

Notes: This figure plots average shares of academic, vocational or 'other' courses students take by initial enrolment. Shares are constructed considering all courses and modules (of more than one month of length) studied within 24 months of the relevant enrolment (which starts with the first observed study spell after GCSEs). We weight courses by total time required to complete them ('guided learning hours'). Courses are classified as follows: A(AS)-Levels and GCSE qualifications are classified as academic; non-A-Levels qualifications in vocational subjects are classified as vocational; qualifications whose recorded subject is 'Preparation for Work and Life' (such as qualifications known as Key Skills or Functional Skills) are classified as other.

are numerous but small. Similar to community colleges in the US, FE Colleges were historically established to offer adult education. While this function remains important, over the last decades they have increasingly shifted their focus to the education of 16–18-year-old students pursuing alternatives to purely academic education. Next to vocational courses, FE Colleges offer courses in basic and soft skills (such as employability or communication skills) and some academic courses (remedial courses in English and maths but also A-Levels). Therefore, not dissimilarly to other countries, vocational-track students typically study a more mixed curriculum than those in the academic track. Note that in England (much like in the US) this is the result of students 'mixing and matching' courses instead of a predefined curriculum.

For the empirical analysis, we conceptualise the education choice students face after completing compulsory schooling as a choice between one of three treatments: the academic track, defined as enrolling in an academic institution (i.e. SF College or a school's sixth form); the vocational track, defined as enrolling in a vocational institution (i.e. FE College or other vocational education provider); and no post-16 education, defined as not enrolling in any upper secondary education. Importantly, in our empirical analysis we focus on students from secondary schools *without* sixth form, who have to switch institutions at age 16 regardless of whether they want to attend the vocational or the academic track. This is crucial for our identification strategy which relies on students' proximity to the closest academic and the closest vocational provider as two alternative-specific instrumental variables influencing students' choices through shifting the respective alternative's costs.

Accordingly, our analysis primarily compares vocational-track students on FE Colleges with academic-track students on SF Colleges, largely excluding those on schools. FE and SF Colleges have similar governance and funding structures, are both relatively large institutions and only cater to students above 16, all of which contrasts with secondary schools. Accordingly, this

sample restriction allows us to minimise the confounding influence of institutional features in our comparisons despite adopting an institution-based treatment definition. To explore how this maps into the type of qualifications studied, Figure 4.1 shows the teaching hours-weighted distribution of courses students enrol in by treatment status. On average, students at vocational institutions spend the majority of their time studying vocational subjects, though they also study a substantial amount of basic and soft skills courses and some academic subjects. This sharply contrasts with academic-track students who predominantly study academic subjects.

4.2.2 Data Sources, Sample Construction and Variables

We use a unique ensemble of administrative datasets from England, known as Longitudinal Education Outcomes (LEO), to follow three full cohorts of state-school educated pupils through their school and post-compulsory education into the labour market until age 29. The three cohorts we study took their GCSEs in the academic years 2001/02 through 2003/04—these are the earliest cohorts for which we can observe all the required information for our analysis.

To define our base sample we use the pupil census of the National Pupil Database (NPD), which reports information on the universe of students enrolled in state-funded schools in England.[3] For the three above-mentioned academic years, we retain all students in the final year before their GCSEs (year group 11) to define our cohorts of interest.[4] The pupil census includes information on students' gender, ethnicity, special educational need, language spoken at home and free school meal (FSM) eligibility, which we use as controls. Further, we record students' test scores in standardised national end-of-primary-school exams in English, maths and science. These so-called Key Stage 2 (KS2) exams serve as our main ability controls.[5] We standardise scores to mean zero and standard deviation one within cohorts.

The three exhaustive and mutually exclusive treatments of interest are enrolling in the academic track (i.e. SF College or a school's sixth form), enrolling in the vocational track (i.e. FE College or other vocational education provider) and not pursuing any upper secondary education after completing compulsory education. To observe all post-16 education choices we link the NPD data with the Individualised Learner Records (ILR), a dataset which covers the universe of publicly-funded education and training activities.[6] Equipped with this information, we define treatment by the institution type of students' first observed enrolment, if any, within a two-year window after completing their GCSEs.[7]

[3] State-funded schools comprise 93% of the total English student population.

[4] We exclude 3% of students from special educational needs (SEN) schools and 0.5% of students for whom we do not observe a GCSE exam (who mostly are of SEN status).

[5] We also observe students' GCSE results, but do not use them as controls because incentives to perform in the GCSE exams are stronger for students who intend to pursue the academic track (see previous section), making performance in these exams potentially endogenous to post-16 education choices.

[6] We thus observe any enrolment at a school's sixth form in the NPD and any enrolment at a SF College, FE College or other private or public vocational education provider in the ILR.

[7] In order to avoid misclassification from short courses or initial enrolments that are subsequently not actually taken up, we ignore learning spells shorter than one month in the treatment assignment.

Labour market outcomes come from Her Majesty's Revenue and Customs (HMRC) tax records, which we can link to our student data for the tax years 2004 to 2017. The data covers earnings and employment spells of all employed individuals in England. From 2014 onwards, we also observe earnings from self-employment. We sum the earnings accruing from all employment spells and self-employment in a given year, deflate by the annual UK consumer price index (base year 2017) and winsorise at the 99^{th} percentile to construct our primary outcome of interest: real annual earnings. This includes observations with zero earnings, whether unemployed or inactive, because labour market attachment is likely endogenous to education choices.[8] Still, in some analyses we focus exclusively on students with positive earnings who are unequivocally part of the labour market. To disentangle extensive margin effects, we construct an indicator for sustained employment, which takes value one if a student was employed more than six months in a given year.[9] Our main models focus on outcomes at age 29, the oldest age at which we observe outcomes for all three cohorts. To study dynamics, we also construct a student-level panel of employment and earnings, though for this we have to exclude (earnings from) self-employment to ensure comparability across the full age range.

Moreover, we construct a number of educational outcomes. First, using the NPD and ILR, we calculate the share of vocational *vs.* academic courses studied during their first two years of post-compulsory education (see Figure 4.1). Second, we construct three indicators for students' upper secondary attainment: whether students complete any Level 3 qualification (i.e. A-Levels or an equivalent vocational qualification), whether they obtain a 'full' Level 3 (i.e. two Level 3 qualifications), which is the minimum university entry requirement, and whether they do so by completing two A-Levels. Third, using the ILR, we construct an indicator for whether students ever start an apprenticeship, which is seen as a desirable outcome for vocational courses. Fourth, we link our sample to data from the Higher Education Statistics Agency (HESA) containing the universe of university enrolments to construct indicators for starting a 3-year university degree; doing so at a more selective pre-1992 university; and completing a university degree.

Our two instrumental variables measure students' proximity to the nearest FE College—as the main provider of upper secondary vocational education for all students regardless of secondary school—and the nearest SF College—as the main provider of upper secondary academic education for students from schools without sixth form. In the pupil census, we observe the Lower Super Output Area (LSOA) of students' home address in their final year of compulsory education. LSOAs are small geospatial areas that divide the surface of England into about 33,000 units of 1,000–1,500 inhabitants each. We proxy students' residential location with the population-weighted centroid of their LSOA, calculate ellipsoidal distances in kilometres to all FE and SF Colleges and take the minimum within each set. Because both distance measures are heavily

[8] 6% of students cannot be matched to the tax records, meaning they do not show a single earnings spell even 13 years after leaving school. In principle, we retain these observations to avoid selecting our sample on outcomes. At the same time, this group is likely to include many individuals that under no circumstances would have entered the labour market (e.g., severely disabled students) and whose education choices are unlikely to respond to incentives (such as distance). Hence, among the set of unmatched students, we at least exclude those with SEN status or missing KS2 scores (which is often associated with special education needs).

[9] As we do not observe employment spells for the self-employed, we set the sustained employment indicator to one if an individual earned more than £10,000 in a given year.

skewed and the effect of distance on choices is likely to vary with distance, we transform both to natural logarithms for the analysis.[10] Henceforth, we will refer to the two (logged) instruments as distance to vocational college, Z_V, and distance to academic college, Z_A, respectively.[11]

Our use of distance instruments makes it paramount to control for residential sorting (e.g. Spiess and Wrohlich, 2010). Accordingly, we construct an elaborate control set consisting of student-, school- and neighbourhood-level covariates. At the student level, our control set contains all above-mentioned student demographics, including all their two-way interactions, and cubic polynomials in all three KS2 test scores.[12] At the secondary school level, we include indicators for school type, averages of the three KS2 test scores and the shares of FSM eligible, White British and English as a second language students. To measure neighbourhood quality we include cubic polynomials in seven domain-specific Indices of Deprivation (IoD), which are constructed by British Ministry of Housing, Communities and Local Government and vary at the fine-grained LSOA level: IoD income; IoD employment; IoD education, skills and training; IoD health and disability; IoD crime; IoD housing and service; and IoD living environment. On top of these covariates, we include fixed effects for the nine regions of England and for student cohorts to compare students that face similar local labour market conditions when they complete compulsory education. Finally, to alleviate concerns about educational providers concentrating in local centres and families sorting along similar dimensions, we add students' distance to the nearest local economic centre as region-specific cubic polynomials.

4.2.3 Summary Statistics

As mentioned above, our estimations focus on students from secondary schools without sixth form because, unlike their peers from schools with sixth form, at age 16 these students need to enrol at a new institution for either upper secondary track. The first two columns of Table 4.1 compare the full student population to the 40% of students who attend schools without sixth form. Panel A shows that the latter are 12 percentage points (pp) more likely to enrol in vocational, 2 pp more likely to pursue no post-16 education and, conversely, 13 pp less likely to enrol in the academic track than the average (panel A). This might be due to higher barriers to academic track entry when it is not offered by one's secondary school, school quality differences and/or different student body compositions, apparent from the remainder of the table: students from schools without sixth form are more likely to have a special educational need or to be economically disadvantaged (panel B), have lower primary school tests scores (panel C), live in more deprived neighbourhoods (panel D) and earn less by age 29 (panel F). They do not appear to live in more rural areas as evidenced by parity in distance to the closest local economic centre (panel D). However, they do live closer to academic and vocational colleges, indicating some sorting of these institutions towards their constituencies or *vice-versa* (panel E).

[10] Furthermore, we apply one final sample restriction and drop the 3% most remote students (who live farther than 63km from any college) as the data becomes sparse and the first stages break down at distances that large.

[11] Appendix Figure A4.1 shows their distribution in levels and in logs.

[12] As about 10% of students have missing values for at least one KS2 score, we include missing dummies for all three scores.

Table 4.1. Summary statistics.

	All students (1)	Schools w/o sixth form (2)	Academic education (3)	Vocational education (4)	No post-16 education (5)
A. Treatment choices					
No post-16 education (D_N)	0.14	0.17	0.00	0.00	1.00
Vocational education (D_V)	0.43	0.54	0.00	1.00	0.00
Academic education (D_A)	0.43	0.30	1.00	0.00	0.00
B. Demographic characteristics					
Female	0.49	0.49	0.54	0.49	0.43
White British	0.81	0.81	0.74	0.82	0.86
English as second language	0.09	0.10	0.16	0.09	0.06
Special educational need	0.14	0.16	0.06	0.18	0.28
Free school meal (FSM)	0.13	0.17	0.11	0.18	0.26
C. Previous achievement					
KS2 score English	0.00	-0.13	0.42	-0.29	-0.64
KS2 score Maths	0.00	-0.11	0.41	-0.27	-0.57
KS2 score Science	0.00	-0.11	0.38	-0.25	-0.58
GCSE points	0.00	-0.15	0.59	-0.30	-1.00
D. Neighbourhood characteristics					
IoD income	0.00	-0.22	0.06	-0.27	-0.54
IoD employment	0.00	-0.25	0.05	-0.33	-0.54
IoD education	-0.00	-0.23	0.16	-0.31	-0.65
IoD environment	-0.00	-0.14	-0.05	-0.15	-0.28
Iod crime	0.00	-0.18	0.01	-0.21	-0.41
IoD housing	-0.00	0.05	-0.03	0.08	0.07
IoD health	0.00	-0.28	0.00	-0.35	-0.53
Distance to local centre in km	8.0	8.1	7.4	8.6	7.6
E. Distance instruments					
Distance vocational college in km	6.1	5.0	5.3	4.9	5.0
Distance academic college in km	15.3	10.7	7.2	12.5	11.2
F. Labour market outcomes at age 29					
Sustained employment	0.74	0.72	0.80	0.73	0.56
Annual earnings in £ (incl. 0s)	17,434	15,988	20,519	15,158	10,599
Annual earnings in £ (excl. 0s)	21,691	20,268	24,163	18,991	16,281
Observations	1,570,992	618,825	183,269	332,699	102,857

Notes: This table reports means of key variables for different samples. Panel A shows the treatment indicators, i.e., students' initial post-16 education choice, panels B–D show various control variables (except for GCSE points, which are not part of the control set), panel E shows the two distance instruments and panel F shows the three main labour market outcomes. The test scores in panel C and the (inverted) indices of deprivation (IoD) in panel D have been standardised to zero mean and unit standard deviation in the full sample for each cohort. See section 4.2.2 for detailed variable definitions. Column 1 describes the full sample, i.e., all students enrolled in English state-funded schools who were in their final year of compulsory education in the academic years 2001/02–2003/04, save for minor sample restrictions described in 4.2.2 (e.g., excluding students from SEN schools). Column 2 describes our estimation sample, i.e., students from secondary schools without a 'sixth form' (i.e., without upper secondary provision). Columns 3–5 split the estimation sample by treatment group.

Our estimation sample thus represents a moderately negatively selected group in terms of achievement and socio-economic background for whom vocational education plays a particularly important role. Accordingly, they are the group likely most affected by changes in the provision of vocational education and are likely to have a higher share of students at the margin between vocational and academic education. These facts make our estimation sample relevant from a

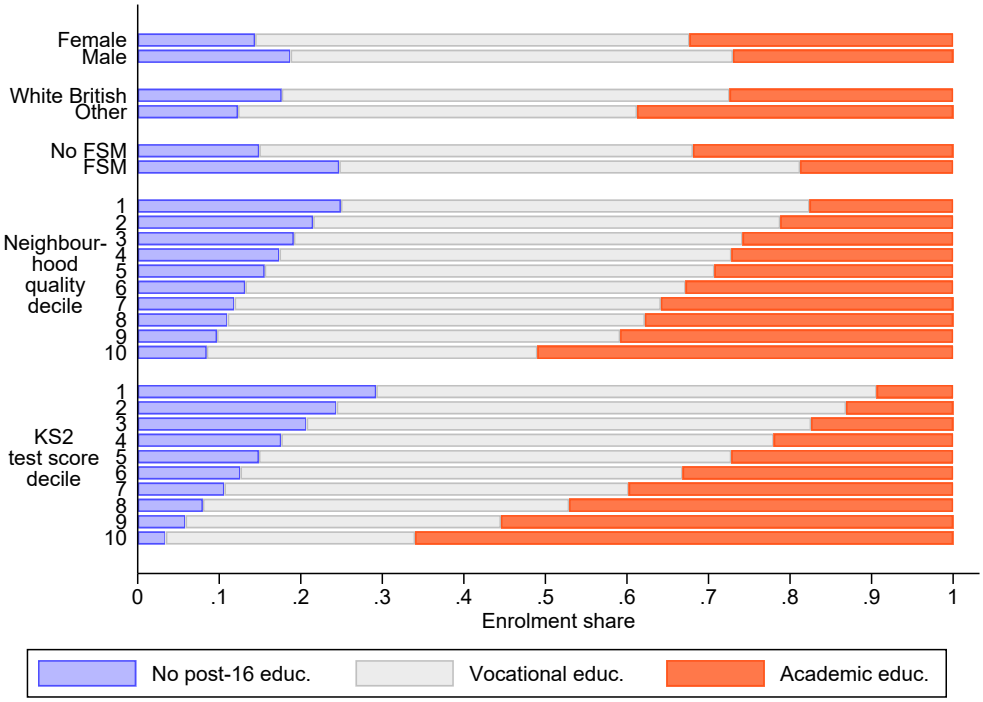


Figure 4.2. Track choices by observable characteristics.

Notes: This figure shows the distribution of students over treatments by observable characteristics. The figure is based on the estimation sample of students from schools without sixth form. FSM stands for free school meal eligibility. Neighbourhood quality deciles are deciles of the first principal component (PC) of all seven (inverted) IoDs. KS2 test score deciles are deciles of the first PC of all three end-of-primary-school (KS2) test scores. PCs are extracted (and their deciles calculated) in the full sample, so that the deciles refer to the same categories in the estimation sample as for students from schools with sixth form (see Appendix Figure A4.2.)

policy perspective, even though it does not cover the whole population. We revisit the question of external validity in section 4.5.4.

4.2.4 Selection into Educational Tracks

Columns 3–5 of Table 4.1 show stark differences in the composition of the three treatment groups. To better understand the selection patterns, Figure 4.2 plots the distribution of educational choices conditional on different observable student characteristics in our estimation sample.[13]

The treatments of academic and no post-16 education correlate more strongly with observables than that of vocational education. For example, vocational enrolment is very similar between genders, but females are substantially more likely to enrol in the academic track and substantially less likely to pursue no upper secondary education. Economically disadvantaged (i.e. FSM) students are far less likely to enrol in the academic track, only somewhat more likely

[13] Appendix Figure A4.2 repeats this exercise for students from schools with sixth form. It shows that, while the share of students choosing vocational and no post-16 education is lower and the share choosing academic is higher than in our estimation sample within each covariate cell, next to these level differences the selection patterns are very similar. This suggests that treatment selection is governed by the same process and our complier treatment effect may well extrapolate to students at the margin between vocational and academic education coming from schools with sixth form.

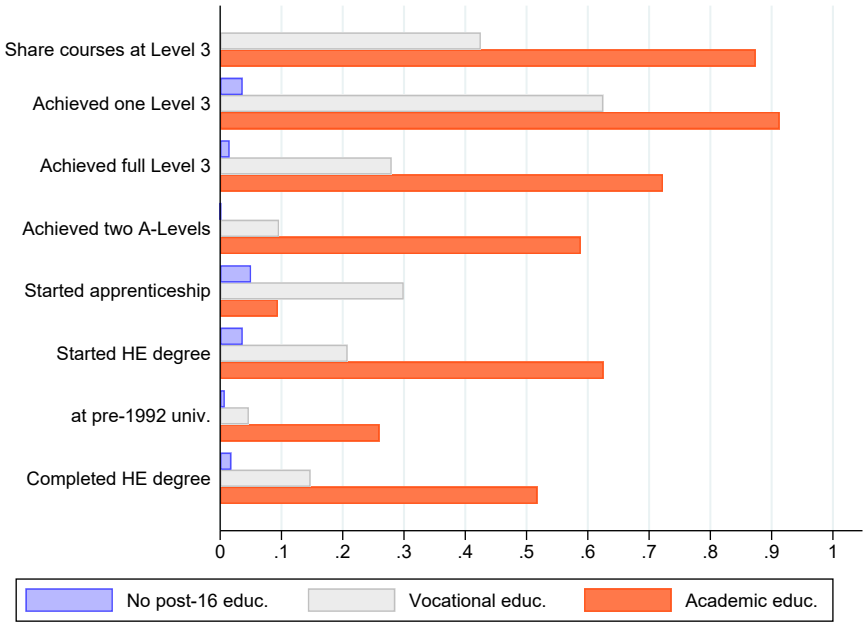


Figure 4.3. Educational outcomes by initial enrolment.

Notes: This figure shows means for several indicators of educational attainment by treatment group. The figure is based on the estimation sample of students from schools without sixth form. The share of courses at Level 3 is computed over all courses (longer than one month) started within 24 months of the relevant enrolment. The other indicators are measured over the whole period covered by the data (i.e., up to age 29–31 depending on the cohort). 'Full Level 3' stands for successfully passing two Level 3 qualifications (A-Levels or vocational equivalent). Pre-1992 universities are more selective universities.

to enrol in the vocational track and much more likely to not enrol in either than those without disadvantage. With respect to neighbourhood quality, the enrolment gradient is steepest for academic education, with students from the highest decile more than twice as likely to enrol in the academic track than those from the lowest decile, whereas the no post-16 education and vocational education shares decrease roughly equally over the same range. By far the best predictor of education choices, however, is previous achievement. Academic enrolment monotonically increases with test scores from under 10% in the lowest decile to almost 70% in the highest decile. Conversely, the no post-16 education share monotonically decreases from about 23% to 3% over the same range. Also the vocational share is decreasing in test scores, albeit less steeply.

Altogether, Figure 4.2 reveals that the group of students in vocational education is more heterogeneous than the other two, who are more clearly either negatively or positively selected. For higher-achieving students, the vocational-academic margin seems most relevant; for lower-achieving students the vocational-no post-16 education margin seems most relevant.

4.2.5 Differences in Education Outcomes

Figure 4.3 compares the three treatment groups in our estimation sample in terms of their educational progression and attainment. Differences in upper secondary attainment between groups are stark: only two-third of vocational-track students successfully complete an upper secondary (Level 3) qualification and not even one-third manage to reach a full Level 3 (i.e. two

Level 3 qualifications). This sharply contrasts with the academic track, where these attainment rates lie above 90% and 70%, respectively. As expected, the vast majority of academic-track students who achieve a full Level 3 do so by completing two A-Levels. Most vocational-track students who complete a full Level 3 do so through vocational qualifications, though about one-third also complete two A-Levels in the process, indicating that curricular mixing is more common in vocational colleges. Unsurprisingly, upper secondary attainment of the no post-16 education group is negligible.

Differences in post-secondary educational progression are equally strong. About one-third of vocational-track students start an apprenticeship, compared to only 9% of academic-track students and 5% of those who initially chose no post-16 education. While most academic-track students progress to higher education and go on to complete a three-year degree, only 20% of vocational-track students go to university and only 14% complete a degree. While these difference are large, vocational track attendance clearly does not rule out progression to higher education *per se*. Still, vocational-track students are much less likely to graduate conditional on starting a degree and generally enrol in less selective universities.

4.2.6 Differences in Labour Market Outcomes

Figure 4.4 plots raw age-employment (panel A) and age-earnings profiles including and excluding individuals with zero earnings (panel B and C) by initial enrolment. All three treatment groups start off with similar employment probabilities at age 18, but in the early years those of vocational- and academic-track students grow at a much faster rate than those of students without post-16 education. A small initial advantage of vocational- over academic-track students reverses by age 22 for females and 23 for males, roughly corresponding to the age many academic-track students leave university. For both genders, the raw differences have stabilised by age 28 and are much larger between vocational and no-post 16 education than between academic and vocational education.

The earnings trajectories show the expected pattern: those without post-16 education earn most initially, are quickly overtaken by those with vocational education, who in turn are overtaken by those with academic education. In both cases, women tend to experience this overtaking a year earlier than men, as was the case for employment. While the earnings differences between the vocational and the no post-16 education group stabilise rather quickly, those between the academically educated and the rest continue to grow throughout students' twenties—but at a decreasing rate. For women, differences seem to have roughly stabilised by age 29; for men, the trends suggest that they continue to grow beyond age 30 but at a slower pace. In contrast to employment, for earnings the differences between academically and vocationally educated students are more pronounced than those between vocational and no post-16 education students. Note that the raw education premiums in terms of employment and earnings are larger for women than for men, particularly so for the vocational-academic contrast we are primarily interested in. Accordingly, we will perform all of our analyses separately by gender.

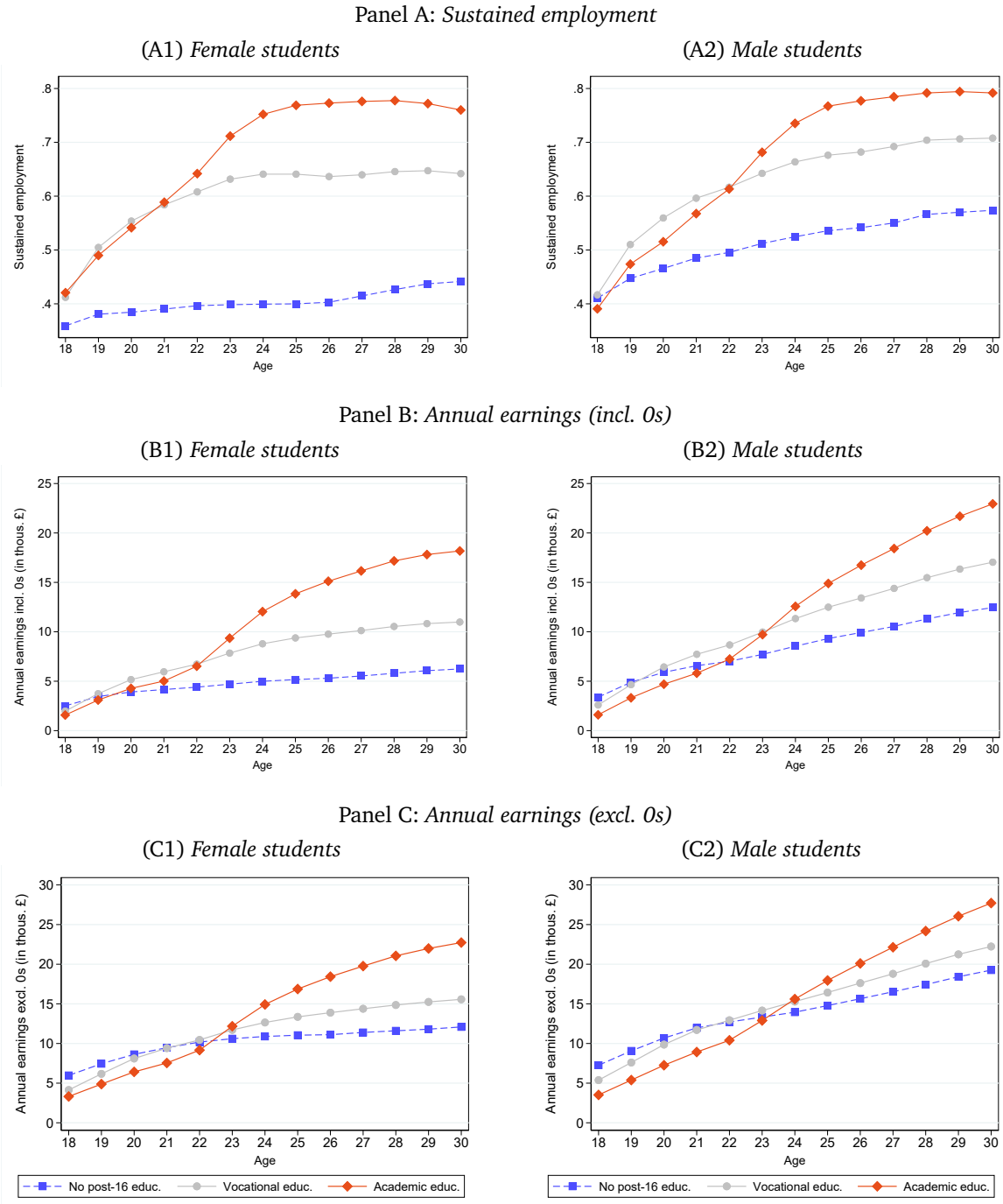


Figure 4.4. Labour market trajectories by initial enrolment.

Notes: The figure plots average labour market outcomes from age 18 to 30 by treatment group. The figure is based on the estimation sample of students from schools without sixth form. Annual earnings are measured in real 2017 British pounds. Sustained employment indicates being employed more than 6 months in a given year. For comparability across the whole age range only earnings (and employment) from employed, but not from self-employed, work are included. Outcomes at age 30 are not available for the 2004 cohort.

4.2.7 OLS Results

The raw labour market outcome differences between education groups represent a mixture of causal effects and selection. As a first step towards approximating causal returns to upper

Table 4.2. OLS regressions for labour market outcomes at age 29.

Dependent variable:	Sustained employment		Annual earnings (incl. 0s)		Positive earnings (excl. 0s)	
	Raw (1)	Controlled (2)	Raw (3)	Controlled (4)	Raw (5)	Controlled (6)
A. Female students						
Academic education (D_A)	0.125***	0.069***	7,150***	3,878***	6,900***	3,668***
	(0.002)	(0.002)	(56)	(56)	(57)	(56)
No post-16 education (D_N)	-0.216***	-0.181***	-4,977***	-3,575***	-3,422***	-2,023***
	(0.003)	(0.003)	(59)	(57)	(79)	(75)
Dependent var. mean	0.68	0.68	13,030	13,030	17,522	17,522
Observations	303,607	303,607	303,607	303,607	225,771	225,771
B. Male students						
Academic education (D_A)	0.060***	0.034***	4,745***	1,963***	4,632***	1,714***
	(0.002)	(0.002)	(68)	(70)	(66)	(68)
No post-16 education (D_N)	-0.121***	-0.106***	-4,030***	-3,000***	-2,529***	-1,417***
	(0.002)	(0.002)	(71)	(70)	(73)	(72)
Dependent var. mean	0.76	0.76	18,838	18,838	22,632	22,632
Observations	315,217	315,217	315,217	315,217	262,373	262,373

Notes: The table results from OLS regressions of labour market outcomes on indicators for academic education and no post-16 education (making vocational education the reference category), separately by gender. Results are based on the estimation sample of students from secondary schools without a sixth form. The raw specification controls for cohort fixed effects. The controlled specification controls for all two-way interactions between indicators for White British ethnicity, English as a second language, free school meal (FSM) eligibility and special educational need (SEN); for cubic polynomials of KS2 scores in English, maths and science; for cubic polynomials of all seven neighbourhood-level indices of deprivation; for secondary school type indicators and school-level averages of KS2 test scores and demographic characteristics; for cohort and government-region fixed effects; and for region-specific cubic polynomials of students' distance to the closest economic centre. Standard errors, reported in parentheses, are clustered at the LSOA×cohort level. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

secondary education, we use OLS regressions and our rich background data to estimate *controlled* education premiums. In particular, we estimate models of the following form:

$$Y = \alpha + \beta_A D_A + \beta_N D_N + \gamma \boldsymbol{X} + \varepsilon, \tag{4.1}$$

where the dependent variable Y is either employment or annual earnings (including and excluding zeros) at age 29, both of which now include self-employment in contrast to Figure 4.4. D_A and D_N are indicator variables for academic and no post-16 education, respectively, making vocational education the reference category. $\boldsymbol{X}$ is the flexibly interacted control set containing student demographics, previous performance, school- and neighbourhood characteristics, described in section 4.2.2.

The OLS results in Table 4.2 show that gaps between education groups are large, even after conditioning on observables. Taken at face value the coefficients in column 4, they suggest that choosing vocational instead of academic education at age 16 reduces earnings at age 29 by roughly £2,000, or 10% of average annual earnings, for men and £4,000, or 30% of annual earnings, for women. Focusing on individuals with positive earnings (column 6) yields slightly smaller but still substantial gaps. Extensive margin effects play a role in explaining the larger

effect for female students, as academically educated women are 7 pp more likely to be employed than vocationally educated women, while for men this coefficient is only half as large. The contrast between vocationally educated students and those without post-16 education is even larger, especially in terms of employment.

4.3 Research Design

4.3.1 Limitations of Traditional Approaches

Obtaining meaningful estimates of the returns to vocational education is challenging in the English setting. The first and foremost problem is that student self-selection into educational tracks is strong. With upper secondary admissions and enrolment fully decentralised, students' education choices likely correlate with unobserved traits and preferences that also determine later labour market outcomes. Hence, despite our comparatively elaborate control set, selection on observables is unlikely to hold and the remaining omitted variable bias in the controlled OLS estimates might well be substantial.

A second problem is that treatment effect heterogeneity is likely to be important. Already returns to different *levels* of education generally display a lot of heterogeneity (see e.g. Carneiro *et al.*, 2011). Returns to different *types* of education can be expected to vary even stronger and likely correlate with students' education choices (Dahl *et al.*, 2020). Accordingly, for many questions the average treatment effects (ATEs) estimated by OLS might not be the most relevant target parameters. For example, in the presence of selection on gains, instead of ATEs we would require effect estimates for students *at the two margins of choice* to judge whether too many or too few students are choosing vocational over academic or no post-16 education from an efficiency standpoint. From a policy perspective, ATEs might well be unrepresentative of effects for students whose education choices are responsive to policy changes, like an expansion or contraction of the vocational sector.

Third, even absent such 'essential' heterogeneity, OLS gives no indication of the relative responsiveness of the two margins of choice, i.e., whether increasing the attractiveness of the vocational track primarily draws in students from no post-16 or from academic education. This limits the guidance OLS can give to policymakers. For example, a large ATE of vocational *vs.* no post-16 education is misleading if students at the corresponding margin cannot be reasonably be induced to change their education choice, but only students at the vocational *vs.* academic margin react to incentives.

Instrumental variables (IV) are a canonical solution to the problems of OLS. Given a valid instrument, IV eliminates selection bias, allowing for estimation of *causal* returns to vocational education. Further, IV identifies local average treatment effects (LATEs) for marginal students (i.e., compliers with respect to the instrument), which can therefore be more policy-relevant than ATEs (Imbens and Angrist, 1994). However, in our setting there are multiple margins of treatment because students choose between three, instead of only two, unordered alternatives. Among

students that are 'marginal' with respect to vocational education, returns will likely systematically differ depending on students' next-best alternatives, i.e., whether they would counterfactually choose academic or no post-16 education. Therefore, a comprehensive description of the returns to vocational education requires the estimation of two separate *margin-specific* LATEs: the effect of vocational *vs.* academic education for students at the corresponding margin; and the effect of vocational *vs.* no post-16 education for students at that margin. However, conventional IV does not identify alternative-specific treatment effects in multi-valued treatment settings, even with as many instruments as treatments available (Kirkeboen *et al.*, 2016).[14] These shortcomings of conventional multivariate IV motivate our use of **?**'s (**?**) alternative IV-based identification approach.

Before outlining its details, it is worthwhile to highlight that a simple univariate IV—instrumenting the binary indicator for vocational education, D_V, with a single instrument (such as distance to vocational college)—identifies a generic *net* LATE of vocational education against compliers' unobserved next-best alternative. As shown by Kline and Walters (2016), this net LATE decomposes into the two unidentified margin-specific LATEs of interest with weights equal to the share of compliers at each margin:[15]

$$\beta_V^{IV} = \underbrace{\text{LATE}_V}_{\text{Net effect of } V} = \overbrace{\lambda}^{\substack{\text{Complier share} \\ \text{at } V\text{-}A \text{ margin}}} \underbrace{\text{LATE}_{V\text{-}A}}_{\text{Effect of } V \text{ vs. } A} + \overbrace{(1-\lambda)}^{\substack{\text{Complier share} \\ \text{at } V\text{-}N \text{ margin}}} \underbrace{\text{LATE}_{V\text{-}N}}_{\text{Effect of } V \text{ vs. } N} . \tag{4.2}$$

Using distance to vocational college as an instrument, we could thus estimate the net effect of vocational education for all 'marginal' vocational education enrollees. While arguably an important parameter, the net effect offers only a limited understanding of returns to vocational education in England. First, equation (4.2) shows that the same net effect can be composed of many different combinations of margin-specific effects, with potentially very different policy implications. Second, as alluded to above, normative judgment about inefficiencies in the allocation of students to tracks also requires margin-specific effects.

4.3.2 Empirical Strategy

To identify margin-specific treatment effects we follow an identification approach proposed by **?**. It extends existing theory for identifying complier potential outcomes (POs) in IV settings developed by Imbens and Rubin (1997) and Abadie (2002) to multi-valued treatments by leveraging multiple (alternative-specific) instruments. The key ingredient of **?**'s (**?**) approach is

[14] For example, Two Stage Least Squares (2SLS) applied to equation (4.1), instrumenting the two treatments D_A and D_N with the two distance instruments Z_V and Z_A, does not identify the two margin-specific effects of interest but yields fundamentally uninterpretable quantities that amalgamate all three possible effect margins (**?**). In particular, it can be shown that: $-\beta_{V\text{-}A}^{2SLS} = \theta_A \text{ LATE}_{V\text{-}A} + (1-\theta_A)(\text{LATE}_{V\text{-}N} - \text{LATE}_{A\text{-}N})$ and that $-\beta_{V\text{-}N}^{2SLS} = \theta_N \text{ LATE}_{V\text{-}N} + (1-\theta_N)(\text{LATE}_{V\text{-}A} + \text{LATE}_{A\text{-}N})$ where $\text{LATE}_{A\text{-}N}$ is the effect at the margin of academic *vs.* no post-16 education and θ_A and θ_N depend on the multivariate 2SLS first-stage equations.

[15] The weight λ equals the share of Z_V-compliers who are at the vocational-academic margin and is identified by the reduction in $Pr(D = A)$ induced by Z_V as a share of the increase in $Pr(D = V)$. The two margin-specific complier treatment effects, however, are not identified.

'cross-instrumentation': using instrumental variation in the attractiveness of *alternative* choices, instead of the choice in question, to identify margin-specific POs, which can then be subtracted from each other to form alternative-specific complier treatment effects. We can implement his method thanks to our two alternative-specific distance instruments: for example, conditional on distance to academic college, variation in distance to vocational college only changes the attractiveness of vocational education, but not that of no post-16 or academic education.

To explain our empirical strategy we require some basic notation. Define the three discrete and mutually exclusive treatment conditions as $D = V$ (vocational education), $D = A$ (academic education) and $D = N$ (no post-16 education), with corresponding binary treatment indicators D_V, D_A and D_N. Denote the associated POs as Y_V, Y_A and Y_N, so that observed outcomes are given by $Y = D_V\, Y_V + D_A\, Y_A + D_N\, Y_N$. Further, denote potential treatment choice as $D(z_V, z_A, \mathrm{x}) \in \{N, V, A\}$, representing the education choice a student of type $\boldsymbol{X} = \mathrm{x}$ would make if exogenously assigned to instrument values $(Z_V, Z_A) = (z_V, z_A)$. Corresponding binary indicators are defined analogously.

4.3.2.1 The Effect of Vocational *vs.* Academic Education

It is well known that IV not only identifies complier treatment effects but also POs (Abadie, 2002). For example, in the simple univariate IV that instruments D_V with Z_V we merely have to replace the outcome variable Y with the treatment-outcome interaction YD_V to identify compliers' mean vocational education PO instead of LATE_V.[16] The intuition behind this is simple: $YD_V = Y_V$ if $D_V = 1$ and $YD_V = 0$ otherwise, so that Z_V-induced changes in YD_V contain information about Y_V for students switching into or out of vocational education in response to changes in distance to vocational college. Yet, as explained above, in our setting these compliers are students who are switching from/to academic education *and* students who are switching from/to no post-16 education. We thus only identify a *net* PO analogous to the net LATE_V in equation (4.2), i.e., a weighted average of these two complier types' vocational education POs with weights equal to their share. Instead of averaging over these two types of compliers, we would like to focus on one margin at a time.

? shows that, with comparable alternative-specific instruments available, this can be achieved through cross-instrumentation: if in the same IV, instead of using distance to vocational college, we use distance to academic college, Z_A, to instrument the treatment of vocational education, while holding fixed distance to vocational college, we restrict complier flows to the vocational *vs.* academic education margin. This is because, conditional on distance to vocational college, variation in distance to academic college only changes the attractiveness of the academic track, leaving the attractiveness of the other two alternatives unaffected. Hence, any Z_A-induced changes in D_V must be due to students switching between vocational and academic education; movements between vocational and no post-16 education are ruled out. Therefore, the univariate IV for the effect of D_V on YD_V, (cross-)instrumenting D_V with Z_A instead of Z_V, but conditioning on

[16] Analogously, with YD_A as the outcome, IV identifies compliers' mean *academic* education PO.

Z_V, identifies the mean vocational education PO for compliers *at the vocational vs. academic education margin only*.[17]

By symmetry, the univariate IV for the effect of D_A on YD_A, (cross-)instrumenting D_A with Z_V, while conditioning on Z_A, identifies the mean *academic education* PO for compliers at the vocational *vs.* academic education margin.[18] Under the assumption that Z_V- and Z_A-induced compliers do not systematically differ (to be discussed below), the first margin-specific effect of interest—that of vocational *vs.* academic education for compliers at the vocational-academic margin—is then identified by the difference between these two 'cross-instrumented' univariate IV estimands.

In principle, the identification results in **?** allow for non-parametric estimation of point-specific marginal treatment effects (MTEs) by using the two continuous instruments as local instrumental variables (LIV) (Heckman and Vytlacil, 2005).[19] Nevertheless, we impose some parametric assumptions to increase statistical power and to obtain readily interpretable LATE estimates for our main results. In particular, we implement both univariate IVs using global linear (in logs) regression models for the reduced form and first stage equations, which control flexibly—but parametrically—for covariates. In section 4.4.4, we relax these assumptions and estimate more local MTEs across different points of the instrument support to test the robustness of our main results and to inspect effect heterogeneity by unobservables.

Thus, for the (margin-specific) mean vocational education PO, we instrument vocational enrolment with conditional distance to academic college by estimating the following pair of reduced form and first stage equations:

$$YD_V = \alpha_0 + \alpha_1 Z_A + \alpha_2 \boldsymbol{X} + \alpha_3 Z_V + \varepsilon \tag{4.3}$$

$$D_V = \pi_0 + \pi_1 Z_A + \pi_2 \boldsymbol{X} + \pi_3 Z_V + \nu, \tag{4.4}$$

[17] Formally, under assumptions A1, A2 and A3, to be discussed below, the average vocational education PO for vocational-academic compliers at point $(Z_V, Z_A, \boldsymbol{X}) = (z_V, z_A, \mathrm{x})$ is identified as a ratio of partial derivatives as follows:

$$\lim_{z_V' \uparrow z_V} \mathbb{E}\left[Y_V \mid D(z_V', z_A, \mathrm{x}) = V,\, D(z_V, z_A, \mathrm{x}) = A\right] = \frac{\partial \mathbb{E}\left[YD_V \mid Z_V = z_V, Z_A = z_A, \boldsymbol{X} = \mathrm{x}\right]}{\partial Z_A} \Big/ \frac{\partial \mathbb{E}\left[D_V \mid Z_V = z_V, Z_A = z_A, \boldsymbol{X} = \mathrm{x}\right]}{\partial Z_A}.$$

The right-hand side is the local instrumental variables (LIV) estimand (Heckman and Vytlacil, 2005) for the effect of D_V on YD_V, instrumenting D_V with Z_A and conditioning on Z_V and $\boldsymbol{X}$.

[18] Analogously, under assumptions A1, A2 and A3, the average academic education PO for vocational-academic compliers at point $(Z_V, Z_A, \boldsymbol{X}) = (z_V, z_A, \mathrm{x})$ is identified as a ratio of partial derivatives as follows:

$$\lim_{z_V' \uparrow z_V} \mathbb{E}\left[Y_A \mid D(z_V', z_A, \mathrm{x}) = V,\, D(z_V, z_A, \mathrm{x}) = A\right] = \frac{\partial \mathbb{E}\left[YD_A \mid Z_V = z_V, Z_A = z_A, \boldsymbol{X} = \mathrm{x}\right]}{\partial Z_V} \Big/ \frac{\partial \mathbb{E}\left[D_A \mid Z_V = z_V, Z_A = z_A, \boldsymbol{X} = \mathrm{x}\right]}{\partial Z_V}$$

[19] The MTE is the limit version of LATE as the instrument shifts tend towards zero. MTE is the continuous instrument analague to LATE because, just like LATE in the binary instrument case, it is defined without parametric assumptions or restrictions on effect heterogeneity (Kennedy *et al.*, 2019). In the binary instrument case (without covariates), 2SLS equals the Wald ratio and thus non-parametrically identifies LATE. This is no longer true in the continuous instrument case, where 2SLS imposes parametric assumptions on the first stage relationship. LIV, in contrast, non-parametrically identifies the point-specific MTE.

to construct the PO's IV estimate as the usual 'Wald' ratio between reduced form and first stage coefficients: $\hat{\alpha}_1/\hat{\pi}_1$ (remember that Z_V is merely a control variable in equations (4.3) and (4.4)). Note that this is numerically equivalent to a 2SLS regression of YD_V on D_V instrumented with Z_A and controlling linearly for Z_V and $\boldsymbol{X}$. Further note that the coefficient ratio is simply the global regression analogue to the ratio of local partial derivatives from footnote 17. Accordingly, the implicit parametric assumption we make is that the relationship between log-distance and choices is constant and linear.

Analogously, for the (margin-specific) mean academic education PO, we instrument academic enrolment with conditional distance to vocational college by estimating the following pair of reduced form and first stage equations:

$$YD_A = \beta_0 + \beta_1 Z_V + \beta_2 \boldsymbol{X} + \beta_3 Z_A + \tilde{\varepsilon} \tag{4.5}$$

$$D_A = \rho_0 + \rho_1 Z_V + \rho_2 \boldsymbol{X} + \rho_3 Z_A + \tilde{\nu}. \tag{4.6}$$

to construct the PO's IV estimate as $\hat{\beta}_1/\hat{\rho}_1$.

The estimate for $\text{LATE}_{V\text{-}A}$ is then formed by differencing the two margin-specific PO estimates:

$$\widehat{\text{LATE}}_{V\text{-}A} = \frac{\hat{\alpha}_1}{\hat{\pi}_1} - \frac{\hat{\beta}_1}{\hat{\rho}_1}.$$

The share of compliers at the vocational-academic margin, λ, is estimated by the ratio of first stage coefficients $\hat{\rho}_1/-\hat{\pi}_3$, which, intuitively, equals the share of the total reduction in vocational enrolment upon an increase in distance to vocational college ($\hat{\pi}_3$) that is due increased academic enrolment ($\hat{\rho}_1$). To obtain standard errors we block bootstrap at the LSOA×cohort-level (the level at which the instruments vary) with 999 repetitions.

4.3.2.2 The Effect of Vocational *vs.* no Post-16 Education

Identification at the second margin, between vocational and no post-16 education, proceeds similarly—with one complication. Analogously to the above, the average no education PO for compliers at the vocational *vs.* no post-16 education margin is identified by a univariate IV for the effect of D_N on YD_N, (cross-)instrumenting D_N with Z_V.[20] We thus estimate the following pair of reduced form and first stage equations:

$$YD_N = \gamma_0 + \gamma_1 Z_V + \gamma_2 \boldsymbol{X} + \gamma_3 Z_A + \check{\varepsilon} \tag{4.7}$$

$$D_N = \tau_0 + \tau_1 Z_V + \tau_2 \boldsymbol{X} + \tau_3 Z_A + \check{\nu}, \tag{4.8}$$

[20] Analogously, under assumptions A1, A2 and A3, the average no education PO for vocational-no post-16 education compliers at point $(Z_V, Z_A, \boldsymbol{X}) = (z_V, z_A, \mathrm{x})$ is identified as a ratio of partial derivatives as follows:

$$\lim_{z_V' \uparrow z_V} \mathbb{E}\left[Y_N \mid D(z_V', z_A, \mathrm{x}) = V,\ D(z_V, z_A, \mathrm{x}) = N\right]$$

$$= \frac{\partial \mathbb{E}\left[YD_N \mid Z_V = z_V, Z_A = z_A, \boldsymbol{X} = \mathrm{x}\right]}{\partial Z_V} \Big/ \frac{\partial \mathbb{E}\left[D_N \mid Z_V = z_V, Z_A = z_A, \boldsymbol{X} = \mathrm{x}\right]}{\partial Z_V}$$

to construct the IV estimate for the no education PO as $\hat{\gamma}_1/\hat{\tau}_1$.

Unfortunately, we cannot apply the same logic to identify the vocational education PO for compliers at this margin. This is because we lack a third instrument shifting only the attractiveness of no post-16 education that could be used to (cross-)instrument vocational enrolment. Yet, as discussed by **?**, a workaround is available. Remember that instrumenting vocational enrolment with (conditional) distance to vocational college in the univariate IV for the effect of D_V on YD_V identifies the *net* vocational education PO for all compliers, which decomposes into the two margins with weights equal to the respective complier shares. As we know these shares and the vocational education PO for vocational-academic compliers to be identified, we can back out the missing vocational education PO for vocational-no education compliers arithmetically from this decomposition.[21] Only using coefficient estimates from the previous reduced form and first stage equations, the PO estimate can be constructed as $(-\hat{\alpha}_3)/\hat{\tau}_1 - (\hat{\alpha}_1/\hat{\pi}_1)\,(\hat{\rho}_1/\hat{\tau}_1)$. Note how the fact that this PO needs to be backed out propagates uncertainty, reducing statistical power at this margin.

As before, the estimate for $\text{LATE}_{V\text{-}N}$ is then formed by differencing the two margin-specific PO estimates:

$$\hat{\text{LATE}}_{V\text{-}N} = \left(\frac{-\hat{\alpha}_3}{\hat{\tau}_1} - \frac{\hat{\alpha}_1}{\hat{\pi}_1}\frac{\hat{\rho}_1}{\hat{\tau}_1}\right) - \frac{\hat{\gamma}_1}{\hat{\tau}_1}.$$

We have thus identified all parts of equation (4.2)'s decomposition of the net effect, LATE_V, into its two margin-specific components, $\text{LATE}_{V\text{-}A}$ and $\text{LATE}_{V\text{-}N}$.

4.3.3 Assessing the Identification Assumptions

We now discuss the identification assumptions required for **?**'s (**?**) procedure. We, in turn, state them formally, discuss their interpretation in our setting and empirically assess their plausibility. To simplify notation, we implicitly condition on the control set $\boldsymbol{X}$ in everything that follows.

4.3.3.1 Independence and Exclusion

The first assumption is the canonical IV assumption of independence and exclusion, adapted to the multiple treatments and two instruments setting:

A1 Independence and Exclusion:

[21] Formally, under assumptions A1, A2 and A3, the average vocational education PO for vocational-no post-16 education compliers at point $(Z_V, Z_A, \boldsymbol{X}) = (z_V, z_A, \mathrm{x})$ is identified as follows:

$$\begin{aligned}
&\lim_{z_V' \uparrow z_V} \mathbb{E}\left[Y_V \mid D(z_V', z_A, \mathrm{x}) = V,\, D(z_V, z_A, \mathrm{x}) = N\right] \\
&\quad = \frac{\partial \mathbb{E}\left[YD_V \mid Z_V = z_V, Z_A = z_A, \boldsymbol{X} = \mathrm{x}\right]}{\partial Z_V} \Big/ \frac{\partial \mathbb{E}\left[D_N \mid Z_V = z_V, Z_A = z_A, \boldsymbol{X} = \mathrm{x}\right]}{\partial Z_V} \\
&\qquad - \frac{\partial \mathbb{E}\left[YD_V \mid Z_V = z_V, Z_A = z_A, \boldsymbol{X} = \mathrm{x}\right]}{\partial Z_A} \Big/ \frac{\partial \mathbb{E}\left[D_V \mid Z_V = z_V, Z_A = z_A, \boldsymbol{X} = \mathrm{x}\right]}{\partial Z_A} \\
&\qquad\quad * \frac{\partial \mathbb{E}\left[D_A \mid Z_V = z_V, Z_A = z_A, \boldsymbol{X} = \mathrm{x}\right]}{\partial Z_V} \Big/ \frac{\partial \mathbb{E}\left[D_N \mid Z_V = z_V, Z_A = z_A, \boldsymbol{X} = \mathrm{x}\right]}{\partial Z_V}.
\end{aligned}$$

Table 4.3. Instrument balance tests.

Dependent variable:	White (1)	FSM (2)	KS2 English (3)	KS2 Maths (4)	KS2 Science (5)
Distance vocational college (Z_V)	0.0024***	0.0000	0.0003	-0.0006	-0.0012
	(0.0008)	(0.0007)	(0.0015)	(0.0014)	(0.0015)
Distance academic college (Z_A)	-0.0004	0.0020***	0.0001	-0.0020*	0.0040***
	(0.0007)	(0.0006)	(0.0012)	(0.0011)	(0.0013)
Controls:					
White British		✓	✓	✓	✓
Free school meal (FSM)	✓		✓	✓	✓
KS2 score English	✓	✓		✓	✓
KS2 score Maths	✓	✓	✓		✓
KS2 score Science	✓	✓	✓	✓	
Remaining controls	✓	✓	✓	✓	✓
N clusters	62,560	62,560	61,100	61,192	61,239
N students	618,824	618,824	563,157	568,867	570,600

Notes: The table reports results from OLS regressions of selected student characteristics on the two distance instruments and the full control set, excluding the covariate in question (see notes of Table 4.2 and section 4.2.2 for a description of the control set). Results are based on the estimation sample of students from schools without sixth form. Standard errors reported in parentheses are clustered at the LSOA×cohort level. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

$$(Z_V, Z_A) \perp\!\!\!\perp \left(Y_N, Y_V, Y_A, \{D(z_V, z_A)\}_{\forall (z_V, z_A)}\right)$$

A1 requires the two distance instruments to be as good as randomly assigned with respect to students' potential outcomes and treatment choices, conditional on the implicit control set $\boldsymbol{X}$. Threats to A1 are residential sorting of students and the non-random location of colleges. If students with a stronger motivation for enrolling in a particular institution type live (or move) closer to it, the exogeneity of distance with respect to education choices would be violated. In this context, we expect residential sorting to be minimal because academic and vocational colleges do not allocate slots based on geographical proximity (nor are they oversubscribed) and generally cater to large regions, so that families' have no incentives to live in their proximity save for shorter travelling times during two years of upper secondary education. Note that we construct the distance instruments using students' residence a full year before post-16 enrolment and potential relocation decisions are made.

A more serious concern is the non-random location of colleges, which may induce a backdoor association between the distance instruments and students' potential labour market outcomes via local economic conditions (thus violating the exclusion restriction).[22] For this reason the control set $\boldsymbol{X}$ contains detailed geographical controls, including distance to the closest economic centre, next to student- and school-level variables. Accordingly, we only leverage distance variation among similar students from similar schools living in similarly remote neighbourhoods with similar economic and social characteristics within the same region. Further note that for identification we only exploit distance to one type of college while holding distance to the other type fixed, which should reduce the scope for backdoor associations.

[22] For example, vocational colleges historically catered for adult workers suggesting they might have been originally located in more disadvantaged areas. Another channel might be any economic (or education) spillover of colleges' presence on surrounding areas, although these are unlikely to be particularly large.

Table 4.3 assesses the plausibility of A1 empirically by means of balance tests on observed pre-determined covariates. One at a time, we exclude the White British indicator, free school meal (FSM) eligibility indicator and the three test scores from the control set to regress the excluded variable on the two distance instruments and the remaining controls. This way we assess covariate balance with respect to changes in distance.

While we do observe some statistically significant associations, economically these are extremely small. Distance to vocational college only significantly correlates with the White British indicator but in negligible magnitude: a one percent increase in Z_V corresponds to a 0.2 pp increase in the probability of being of White British ethnic background. Distance to academic college significantly correlates with FSM and test scores in maths and science. Yet again, these associations are negligible in size: a one percent increase in Z_A is associated with a 0.2 pp increase in the probability of being FSM-eligible and a 0.4% of a standard deviation increase in the KS2 science score. Note that the former is a sign of disadvantage and the latter a sign of advantage so this does not point at a clear directional pattern of sorting. Their small size, inconsistent pattern and the fact that this exercise gives an upper bound of the remaining selectivity (because each time one control variable and its interactions are excluded from the control set) suggest that any potential remaining selectivity is small enough to be ignored.

4.3.3.2 Instrument Relevance (First Stages)

Of course, our research design is only feasible if the two distance instruments are strong predictors of education choices. In particular, identification at the vocational-academic margin requires two non-zero first-stage relationships: the conditional effect of distance to academic college on vocational education (corresponding to π_1 in equation (4.4)) and the conditional effect of distance to vocational college on academic education (corresponding to ρ_1 in equation (4.6)). Identification at the vocational-no education margin additionally requires the conditional effect of distance to vocational college on no post-16 education to be non-zero (corresponding to τ_1 in equation (4.8)).[23]

Figure 4.5 visualises these three first stage relationship relationships by means of binned scatter plots, which non-parametrically control for the other distance instrument and the full control set (Cattaneo *et al.*, 2021). Conditional on distance to vocational college (and all other controls), vocational enrolment increases monotonically and approximately linearly in (log) distance to academic college (panel A). A similar picture arises for the relationship between academic enrolment and distance to vocational college (panel B). This visual inspection thus supports the instrument strength conditions for identification at the vocational-academic margin, as well as the chosen linear approximation.

The conditions for identification at the vocational-no education margin are somewhat less favourable. Panel C shows that the choice of no post-16 education is much less responsive to distance to vocational college. Only at distances larger than 3km a positive slope becomes

[23]The instrument relevance condition is formally embedded in A2 below.

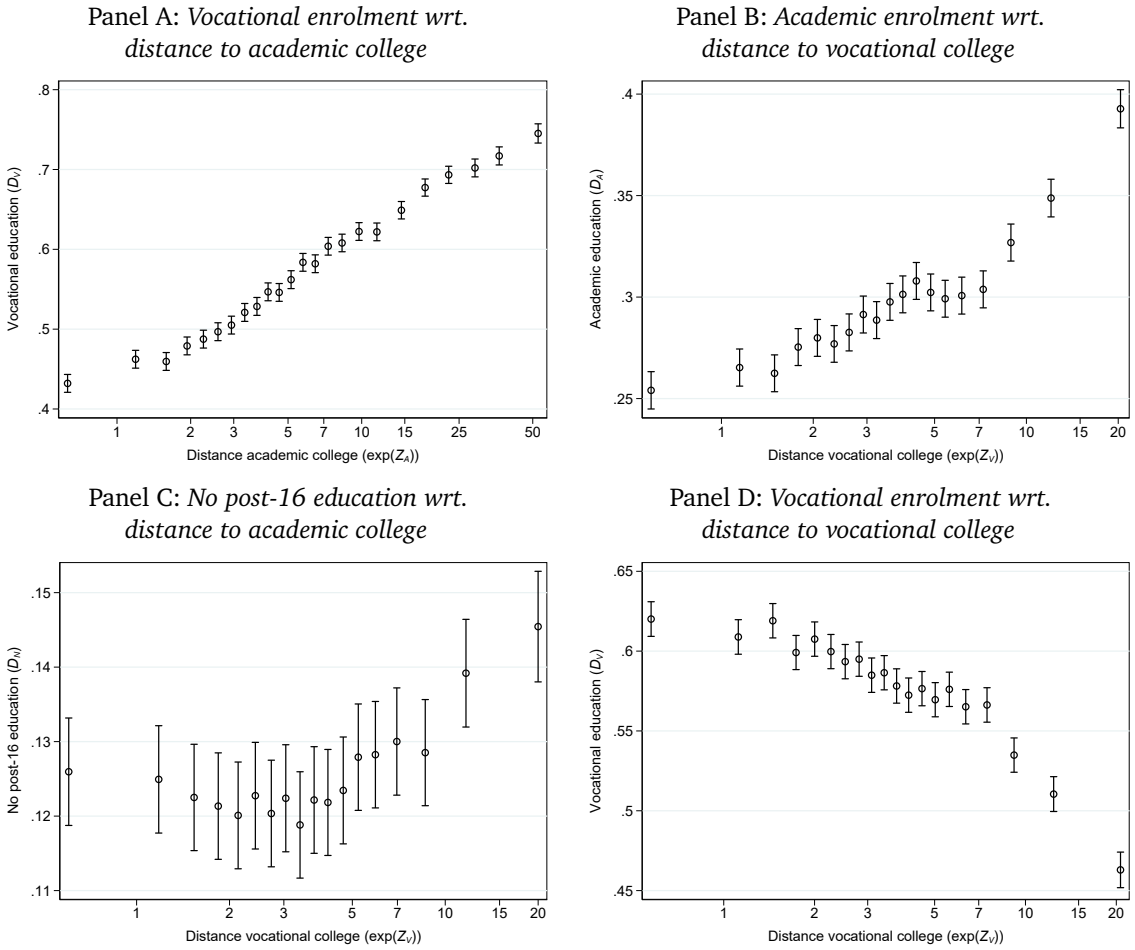


Figure 4.5. Binned scatter plots of first stage relationships.

Notes: This set of figure plots the relevant first-stage relationships between students' treatment choices and distance instruments, built using the 'binscatter' regression routine (Cattaneo *et al.*, 2021). Each panel plots the estimated probability of students choosing a given option (alongside 95% CIs) within optimally-spaced bins of the relevant distance instrument (in natural logs), while controlling non-parametrically for the other distance instrument and the full control set. The figure is based on the estimation sample of students from schools without sixth form.

visible. Panel D shows the relationship between vocational enrolment and distance to vocational college—the traditional first stage for the net effect of vocational education. Note that because $D_V = 1 - D_A - D_N$, the decreases in vocational enrolment with distance to vocational college from panel D mirror the increases in academic and no post-16 enrolment from panels B and C. This tells us that at closer distances to vocational colleges the decline in vocational enrolment is entirely at the advantage of academic education; as distance grows, both margins contribute. Still, even at further distances flows across the academic-vocational margin are much more important: most marginal students, whose decision to enrol in vocational college is responsive to distance, appear to choose between vocational and academic education.

Table 4.4 reports regression results for the three first stage equations (4.4), (4.6) and (4.8), separately by gender. For each first stage, we present two specifications: one excluding and one including test score controls. All coefficients show their expected signs, consistent with the plots of Figure 4.5: distance to vocational college decreases the probability of vocational but increases

Table 4.4. First stages with and without test score controls.

Dependent variable:	Vocational educ. (D_V)		Academic educ. (D_A)		No post-16 educ. (D_N)	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Female students						
Distance voc. college (Z_V)	-0.0388***	-0.0393***	0.0345***	0.0353***	0.0043***	0.0041***
	(0.0016)	(0.0015)	(0.0014)	(0.0014)	(0.0010)	(0.0010)
		[645.5]		[640.9]		[17.6]
Distance acad. college (Z_A)	0.0826***	0.0822***	-0.0881***	-0.0876***	0.0055***	0.0054***
	(0.0013)	(0.0013)	(0.0012)	(0.0012)	(0.0008)	(0.0008)
		[4104.2]				
R^2	0.08	0.11	0.18	0.26	0.06	0.08
B. Male students						
Distance voc. college (Z_V)	-0.0335***	-0.0341***	0.0260***	0.0274***	0.0075***	0.0067***
	(0.0016)	(0.0015)	(0.0013)	(0.0013)	(0.0011)	(0.0011)
		[491.8]		[455.9]		[39.5]
Distance acad. college (Z_A)	0.0702***	0.0700***	-0.0770***	-0.0770***	0.0068***	0.0070***
	(0.0013)	(0.0013)	(0.0011)	(0.0011)	(0.0009)	(0.0009)
		[3042.2]				
R^2	0.06	0.08	0.18	0.26	0.07	0.10
KS2 test scores		✓		✓		✓
Remaining controls	✓	✓	✓	✓	✓	✓

Notes: The table reports results from OLS regressions of treatment indicators on the two distance instruments and the full control set, excluding (odd columns) and including (even columns) cubic polynomials of the three KS2 test scores. Results are based on the estimation sample of respectively female students (panel A) and male students (panel B) from schools without sixth form. Standard errors reported in parentheses are clustered at the LSOA×cohort level. The number of observations is 303,608 (55,079 clusters) in the female sample and 315,217 (55,888 clusters) in the male sample. In square brackets, we present 'effective' F-statistics testing first-stage strength in the presence of heteroskedasticity and clustering, whose critical values for a single-instrument 2SLS lie between 5.53 and 16.38 (Olea and Pflueger, 2013). Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

that of no post-16 and academic enrolment, while distance to academic college decreases the probability of academic but increases that of vocational enrolment. To test first-stage strength, we report robust F-statistics for the three coefficients of interest (Olea and Pflueger, 2013).[24] For both females and males, the F-statistics associated to D_A and D_V far exceed relevant critical values for first stage strength. The F-statistic for D_N is much lower, confirming that compliers at the vocational-no education margin are few (particularly among female students). The weakness of the first-stage relationship suggests caution when interpreting results at this margin.

Note that the first stages are remarkably unaffected by the inclusion of test score controls. Given that test scores are highly predictive not only of students' education choices but also of later labour market outcomes, this suggests that our elaborate demographic, school- and neighbourhood-level controls alone suffice for purging the distance-choice relationships of potential confounders, thus lending further credence the instruments' independence and exclusion (A1).[25]

[24] We report 'Kleibergen-Paap' F-statistics which are robust to heteroskedasticity and clustering and equivalent to the 'effective' F-statistic of Olea and Pflueger (2013) in this case of a single instrument (per first stage).

[25] Appendix Figure A4.3 demonstrates a tight relationship between KS2 test scores and labour market performance: for example, average annual earnings at age 29 more than double across the test score distribution.

4.3.3.3 Partial Unordered Monotonicity

The next assumption extends the intuition of 'no defiers' from the binary to the multi-valued treatment case, considering only conditional instrumental variation:[26]

A2 Partial Unordered Monotonicity:

For all triples (z_V, z'_V, z_A) with $z'_V < z_V$ we have: $D_V\left(z'_V, z_A\right) \geq D_V\left(z_V, z_A\right)$
but $D_A\left(z'_V, z_A\right) \leq D_A\left(z_V, z_A\right)$ and $D_N\left(z'_V, z_A\right) \leq D_N\left(z_V, z_A\right)$ for all individuals,
with each inequality holding strictly for at least some individuals.

For all triples (z_A, z'_A, z_V) with $z'_A < z_A$ we have: $D_A\left(z_V, z'_A\right) \geq D_A\left(z_V, z_A\right)$
but $D_V\left(z_V, z'_A\right) \leq D_V\left(z_V, z_A\right)$ and $D_N\left(z_V, z'_A\right) \leq D_N\left(z_V, z_A\right)$ for all individuals,
with each inequality holding strictly for at least some individuals.

A2 requires that a decrease (increase) in the distance to either type of college, holding constant distance to the other, renders the associated education choice weakly more (less) attractive for *all* students.[27] It does not restrict the complier flows to a certain margin, however. For example, as the distance to vocational college decreases ($z'_V < z_V$), but distance to academic college is held fixed, some people may switch into but no one out of vocational education ($D_V\left(z'_V, z_A\right) \geq D_V\left(z_V, z_A\right)$); whether these compliers come from academic education ($D_A\left(z'_V, z_A\right) \leq D_A\left(z_V, z_A\right)$) or no post-16 education ($D_N\left(z'_V, z_A\right) \leq D_N\left(z_V, z_A\right)$) is left unrestricted. However, nobody can switch between no post-16 and academic education.

Given their exogeneity, partial unordered monotonicity is a natural assumption for our distance instruments. The only plausible violation would stem from complementarities between the two college types, so that the one's attractiveness is tied to that of the other. However, academic and vocational colleges are substitutes and enrolling in one is not a preparatory step for enrolling in the other at a later stage.

To empirically assess the plausibility of A2 we follow the literature and test whether the first stages are consistent across different subsamples of the data (e.g. Dobbie *et al.*, 2018; Bhuller *et al.*, 2020; Agan *et al.*, 2021). Appendix Table A4.1 presents estimates for all three first stage of interest across a large variety of covariate-defined data cells. The first stages of D_A and D_V are positive throughout. The weaker first stage of D_N is zero in some subsamples but never negative. Accordingly, we find no evidence for the presence of 'defiers' in our sample.

4.3.3.4 Complier Comparability

The final assumption is specific to **?**'s (**?**) framework and draws a connection between the two sets of vocational-academic compliers induced by Z_V and Z_A:

[26] Heckman and Pinto (2018) develop the general 'unordered monotonicity' condition for the unordered multi-valued treatment case. It requires that treatment responses are uniform across *all* possible shifts in the instruments. **?**'s (**?**) 'partial unordered monotonicity' relaxes this assumption by looking only at *conditional* variation in the instruments, i.e., focusing on the subset of shifts where one of the two instruments stays constant. This means that we make no assumptions about the behaviour of students in cases where distance to both colleges decreases simultaneously.

[27] And strictly for some students, thus formally embedding an instrument relevance condition.

A3 Complier Comparability:

For all pairs (z_V, z_A):

$$\lim_{z'_V \uparrow z_V} \mathbb{E}\left[Y_V \mid D(z'_V, z_A) = V, D(z_V, z_A) = A\right] = \lim_{z'_A \downarrow z_A} \mathbb{E}\left[Y_V \mid D(z_V, z'_A) = V, D(z_V, z_A) = A\right]$$

A3 states that compliers shifted from academic to vocational college by a marginal *decrease* in distance to vocational college (left-hand side) must be comparable, in terms of POs, with those shifted by a marginal *increase* in distance to academic college (right-hand side). It is required because we can only identify the right-hand side from the data (using distance to academic college to instrument vocational enrolment, as explained above) but not the left-hand side (because distance to vocational college induces compliers also from the other margin). Given that both are sets of students at a margin of indifference between vocational and academic education and both instruments represent simply the distance to the closest respective provider, it is hard to picture how these two complier types could systematically differ.[28]

The unidentified PO on left-hand side of course prohibits a direct test of A3. Still, we can assess its plausibility by comparing the two complier types in terms of their pre-determined characteristics, which are separately identified under A1 and A2 alone.[29] Because A3 is about potential *outcomes*, we use covariate-predicted earnings instead of any single covariate for this comparison. We predict earnings using coefficients from an OLS regression of earnings at age 29 (including zeros) on the control set $\boldsymbol{X}$.

Table 4.5 presents the results from this exercise, separately by gender. Column 1 reports the average predicted earnings for vocational-academic compliers induced by Z_V (left-hand side of A3) and column 2 reports average predicted earnings for those induced by Z_A (right-hand side of A3). Column 3 reports the difference between the two. For female students, predicted earnings differ by -£371 (2.5% taking column 2 as the baseline). For male students, they differ by £324 (which is only 1.5%). Economically, these differences are small enough to suggest that the two complier types are well comparable. They are, however, marginally statistically significant, which is why in section 4.4.2 we probe the robustness of our results to correcting the PO (and effect) estimates for these "expected" earnings differences.

[28] **?** shows that this condition is implied by a standard Roy-style selection model: both Z_V and Z_A act as costs shifting a single index that governs the relative attractiveness of vocational *vs.* academic education. Hence, students who switch their treatment choice in response to a marginal change in the index are the same regardless of whether this change is induced by a marginal decrease in Z_V or a marginal increase in Z_A (or *vice-versa*).

[29] If in A3, we replace Y_V with any characteristic that is not determined by D, e.g., a scalar covariate C from our control set, then also the left-hand side is directly identified. This is because, under A1, $C_V = C_A = C$. Under A2, we can thus estimate Z_V-induced vocational-academic compliers' average C by estimating equations (4.5) and (4.6) replacing YD_A with CD_A. Similarly, estimating equations (4.3) and (4.4) replacing YD_V with CD_V estimates the average C for the Z_A-induced vocational-academic compliers from the right-hand side of A3. In practice, we use covariate-predicted earnings for this exercise.

Table 4.5. Predicted earnings by complier type.

	V-A compliers induced by Z_V (1)	*V-A* compliers induced by Z_A (2)	Difference (1) – (2) (3)
Female sample	14,712	15,030	-318*
	(177)	(77)	(184)
Male sample	21,426	21,094	332*
	(186)	(80)	(189)

Notes: The table reports average covariate-predicted earnings for two groups of vocational-academic compliers: those induced by variation in distance to vocational college (Z_V) in column 1, and those induced by variation in distance to academic college (Z_A) in column 2. Column 3 reports the difference. The details of the estimation are explained in the text. Results are based on the estimation sample of students from schools without sixth form. Predicted earnings are calculated by gender-specific OLS regression of net annual earnings (at age 29) on the full control set. The number of observations is 303,608 (55,079 clusters) in the female sample and 315,217 (55,888 clusters) in the male sample. Standard errors are block bootstrapped at the LSOA×cohort level using 999 iterations. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.4 Results for Labour Market Outcomes

4.4.1 Main Results

Table 4.6 presents the main results for the alternative-specific effects of vocational education on students' labour market outcomes at age 29. We present all results separately by gender: columns 1–3 pertain to women and columns 4–6 to men.

Net effect.—The upper panel reports estimates for the net effect of vocational education, LATE_V, which pools the treatment effects for vocational-academic compliers and vocational-no education compliers into one weighted average. Not a single net effect estimate is statistically significant at conventional levels. Taken at face value, the point estimates suggest small positive effects on employment, null effects on earnings for women and moderate negative effects on earnings for men, though the level of imprecision is such that we cannot rule out large negative or even moderately positive effects. To move beyond these inconclusive results, the remainder of the table decomposes the net effects into their constituent margin-specific components, $\text{LATE}_{V\text{-}A}$ and $\text{LATE}_{V\text{-}N}$, as outlined in the previous section.

Academic education margin.—We first turn our attention to the vocational-academic education margin, where the vast majority of marginal vocational education students appear to be found: the complier shares in the second row of the table indicate that for 93% of female compliers and 82% of male compliers the alternative to the vocational track is academic education and not direct labour market entry. For these students, we find (near) zero effects of choosing vocational education on sustained employment at age 29 for both women (column 1) and men (column 4). For women, we further estimate a small and insignificant effect of vocational enrolment on earnings at age 29 in the full sample (column 2). However, once we restrict the sample to women with positive earnings to approximate those in the labour force, we find a

Table 4.6. IV estimates for margin-specific effects of vocational education.

	Female students			Male students		
	Sustained employment	Annual earnings (incl. 0s)	Annual earnings (excl. 0s)	Sustained employment	Annual earnings (incl. 0s)	Annual earnings (excl. 0s)
	(1)	(2)	(3)	(4)	(5)	(6)
Net effect vocational education						
$LATE_V$	0.007	22	-522	0.023	-979	-1,681
	(0.032)	(846)	(859)	(0.033)	(1,154)	(1,126)
Academic education margin						
Complier share	*0.896*	*0.896*	*0.910*	*0.804*	*0.804*	*0.811*
	(0.023)	(0.023)	(0.023)	(0.027)	(0.027)	(0.028)
$LATE_{V\text{-}A}$	0.003	-463	-1,252*	0.007	-2,347***	-2,837***
	(0.020)	(675)	(683)	(0.022)	(872)	(868)
No post-16 education margin						
Complier share	*0.104*	*0.104*	*0.090*	*0.196*	*0.196*	*0.189*
	(0.023)	(0.023)	(0.023)	(0.027)	(0.027)	(0.028)
$LATE_{V\text{-}N}$	0.045	4,201	6,875	0.086	4,629	3,278
	(0.296)	(6,737)	(8,364)	(0.155)	(5,098)	(4,916)
N students	303,608	303,608	225,772	315,217	315,217	262,373
N clusters	55,079	55,079	51,820	55,888	55,888	53,730

Notes: The table reports IV estimates of the net complier treatment effects of vocational education on the three indicated labour market outcomes (top panel), as well as its decomposition into the two margin-specific effects of vocational *vs.* academic education and vocational *vs.* no post-16 education, alongside the estimated complier share at each margin (bottom panels). The details of the estimation are explained in the text. Results are based on the estimation sample of respectively female students (columns 1–3) and male students (columns 4–6) from schools without sixth form. Standard errors are block bootstrapped at the LSOA×cohort level using 999 iterations. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

negative earnings effect of £1,100 that is significant at the 10% level.[30] This corresponds to a 5.5% reduction in annual earnings. For men, the negative effect of vocational enrolment is more pronounced: in the full sample, we estimate a highly significant and negative effect of £2,600, corresponding to an 11% reduction in annual earnings (column 5). The estimate is precise enough to rule out small negative effects. As labour force participation is much higher for men, restricting the sample to individuals with positive earnings matters less than in the case of women: with £3,150 the effect estimate in column 6 is substantially larger in absolute terms, identical in relative terms (11%).

No post-16 education margin.—The bottom panel of Table 4.6 reports results for the vocational-no education margin, i.e., for compliers whose alternative to vocational enrolment is no further education. As discussed above, students' decision to leave the education system at age 16 is much less responsive to vocational college proximity than the other two alternatives. Only 7% of female compliers and 18% of male compliers are to be found at this margin. The weaker first stages for no post-16 education, together with the fact that the expression for $LATE_{V\text{-}N}$ involves four different coefficient ratios, amount to rather imprecise results. For women, the first stage in question is so weak that the effect estimates in columns 1–3 are too noisy to be meaningfully

[30] Note that, given the absence of extensive margin effects, conditioning on positive earnings should not bias the estimates.

interpreted. For men, the first stage in question is strong at conventional levels and, indeed, their effect estimates in columns 4–6 are slightly more precise, hinting at large positive effects of vocational *vs.* no post-16 education on employment and earnings. However, we are unable to confidently rule out null or even negative effects. In the remainder of the paper, we therefore concentrate on results for the more relevant academic education margin, referring the interested reader to the appendix for some complementary results for the no post-16 margin.

Discussion.—The results in Table 4.6 carry a number of important insights. First, the vast majority of student whose choice to enrol in vocational education is responsive to incentives like distance, is choosing between vocational and academic education, not considering the option of no post-16 education. Accordingly, any policy that seeks to increase vocational enrolment by increasing the attractiveness of the vocational track will do so mainly at the expense of academic enrolment. It follows that understanding the alternative-specific return of vocational *vs.* academic education is of paramount importance for policy. Second, for these vocational-academic compliers, vocational education has no discernible effect on labour market attachment at age 29, but large negative effects on earnings, especially so for men. We note that a larger negative earnings effect for men than for women contrasts with the OLS findings of Table 4.2. Third, given the absence of extensive margin effects, negative earnings impacts must be due to intensive margin responses (i.e., working hours) or wages. Given that full-time employment is by far the most common working arrangement among men in England, differences in working hours are unlikely to be an important driver of these results, however.[31] This suggests that choosing vocational over academic education at age 16 substantially lowers students' wages 13 years later.

Fourth, the results reveal striking heterogeneity in the returns to vocational education: for male students at the margin with academic education, vocational enrolment unequivocally leads to a large reduction in earnings, whereas for those at the margin with no post-16 education, if anything, it appears to increase earnings. While less pronounced, the same pattern is visible for female students. The fact that the net effect is composed of two divergent alternative-specific effects highlights the importance of margin-specific identification in this context. The conventional single-instrument IV estimate is contaminated by large but very imprecise (unidentified) point estimates for a small group of compliers at the no post-16 education margin. This shrouds negative effects for most compliers, thus nurturing an ambiguous and more positive impression of vocational education in England than warranted. Incidentally, the fact that the two alternative-specific LATEs diverge also implies that a two-instrument IV would be no remedy. We illustrate this in Appendix Table A4.2, which shows the estimates obtained from 2SLS model that instruments D_A and D_N with the two distance instruments and decomposes those into their constituent effect and bias terms.

[31] Additionally, from a theoretical viewpoint, we expect education choices to affect the level and type of skills students acquire, and hence the type of job or occupation they can perform or how productively they can perform it. It is less clear how education choices would affect the intensive margin (or in which direction). Indeed, Birkelund and van de Werfhorst (2022) find that vocational education in Denmark increases the probability of being employed in lower-skilled occupations. Still, the question of intensive margin effects is an important that should be addressed by future research, especially for the female part of the sample.

4.4.2 Sensitivity Checks

This subsection presents a number of sensitivity checks probing the robustness of the main results.

Adjusting for complier differences.—We start by addressing concerns about potential violations of A3, which requires compliers shifted from academic to vocational education by a decrease in distance to vocational college (i.e., Z_V-compliers) to have the same mean vocational education PO than those shifted by a marginal increase in distance to academic college (i.e., Z_A-compliers). When assessing this assumption in Table 4.5, we found some small but marginally significant differences in these two complier types' covariate-predicted earnings. Accordingly, it is important to verify that this has no bearing on our conclusions.

To do so, we first repeat the exercise from Table 4.5 for all three labour market outcomes, i.e., for each one we calculate the differences in covariate-predicted mean outcomes between Z_V- and Z_A-compliers at the vocational-academic margin. Then, we adjust the Z_A-compliers-based vocational education PO estimate by this difference and recalculate $\text{LATE}_{V\text{-}A}$ with the adjusted PO. Table 4.7 reports the results from this exercise.[32] The odd columns report the original PO estimates, which underlie the main results in Table 4.6. The even columns report the adjusted POs. For example, we originally (i.e., using Z_A-compliers) estimated that, under vocational education, net annual earnings for male compliers are £21,029 (column 3). According to our procedure, taking into account differences in pre-determined characteristics and how they translate into outcome differences, this estimate should be upward-adjusted to better approximate the PO of Z_V-compliers, who are used to estimate the counterfactual academic education PO. Doing so yields an adjusted vocational education PO estimate of £21,361 which, in turn, yields a slightly smaller $\text{LATE}_{V\text{-}A}$ of -£2,014 (column 4).

Across outcomes, the adjustment marginally attenuates effects for males and marginally accentuates effects for females. The bottom row of Table 4.7 reports p-values testing the equality of the adjusted and unadjusted effect estimates. Statistically, we marginally reject equality in most cases, which is why in the remainder of the paper we use the adjusted estimates from the even columns when referring to our main results and also adjust all our subsequent estimates accordingly. Economically, however, our conclusions remain unchanged: we continue to find no effect of vocational education on employment but large negative effects on earnings. In fact, the latter are now more clearly visible also for females, with the effect in the positive earnings sample reaching statistical significance at the 95% confidence level.

Other sensitivity checks.—In Appendix Table A4.4 we inspect the sensitivity of the estimates to changing the definition of the distance instruments and to different sample restrictions. The effect estimates in columns 1–3 are based on distance instruments defined in terms of driving distance instead of geographical distance.[33] Columns 4–6 report results when abandoning all

[32] Appendix Table A4.3 repeats the exercise for the no post-16 education margin. Note that the net effect LATE_V is unaffected by this because adjustments at the two margins mechanically offset each other.

[33] Driving distances were obtained using the HERE Routing API to compute the shortest route between the population-weighted centroids of students' residential LSOAs and the coordinates of relevant educational institutions (link). See also Weber and Péclat (2017) on how to compute driving distance.

Table 4.7. Correcting the IV estimates at the academic education margin for complier differences.

Dependent variable:	Sustained employment		Annual earnings (incl. 0s)		Annual earnings (excl. 0s)	
	Raw IV (1)	Corrected (2)	Raw IV (3)	Corrected (4)	Raw IV (5)	Corrected (6)
A. Female students						
Vocational PO	0.744	0.735	15,715	15,398	19,505	19,308
	(0.010)	(0.010)	(227)	(275)	(218)	(276)
Academic PO	0.741		16,179		20,757	
	(0.018)		(626)		(655)	
$LATE_{V\text{-}A}$	0.003	-0.006	-463	-781	-1,252*	-1,449**
	(0.020)	(0.020)	(675)	(651)	(683)	(668)
	$[p_{(\Delta=0)} = 0.02]$		$[p_{(\Delta=0)} = 0.09]$		$[p_{(\Delta=0)} = 0.29]$	
B. Male students						
Vocational PO	0.839	0.844	21,029	21,361	23,861	24,194
	(0.010)	(0.010)	(332)	(379)	(322)	(361)
Academic PO	0.832		23,376		26,698	
	(0.019)		(811)		(791)	
$LATE_{V\text{-}A}$	0.007	0.013	-2,347***	-2,014**	-2,837***	-2,503***
	(0.022)	(0.022)	(872)	(857)	(868)	(843)
	$[p_{(\Delta=0)} = 0.07]$		$[p_{(\Delta=0)} = 0.08]$		$[p_{(\Delta=0)} = 0.08]$	

Notes: For each of the three outcomes, the table shows vocational-academic complier potential outcome (PO) and $LATE_{V\text{-}A}$ estimates, as they were originally estimated (odd columns) and when corrected for complier differences (even columns), separately by gender. The vocational education PO is corrected by adding to the original estimate the predicted outcome difference between compliers induced by variation in distance to vocational college and compliers induced by distance to academic college. The corrected $LATE_{V\text{-}A}$ is obtained by subtracting the original academic PO from the corrected vocational PO. In square brackets we report p-values testing against the null hypothesis that the corrected and original $LATE_{V\text{-}A}$ estimates are identical. The number of observations is the same as in Table 4.6. Standard errors are block bootstrapped at the LSOA×cohort level using 999 iterations. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

our sample restrictions and including all students from secondary schools without a sixth form (i.e., keeping students with missing KS2 scores or SEN status not found in the earnings data, as well as the 3% most remote students). Columns 7–9 take the opposite path and exclude all 6% of students who are never observed in, and thus cannot be matched to, the earnings data. None of this affects our conclusions at the vocational-academic margin. Unsurprisingly, the imprecise point estimates at the vocational-no post-16 education margin are more volatile.

4.4.3 Effects by Age

So far we have focused on labour market outcomes at age 29. In Figure 4.6 we inspect effect dynamics over students' early careers, plotting the estimated effects of vocational *vs.* academic education on the probability of sustained employment (panel A) and net annual earnings (i.e., including observations with zero earnings) (panel B) across all two-year age bands from 18–19

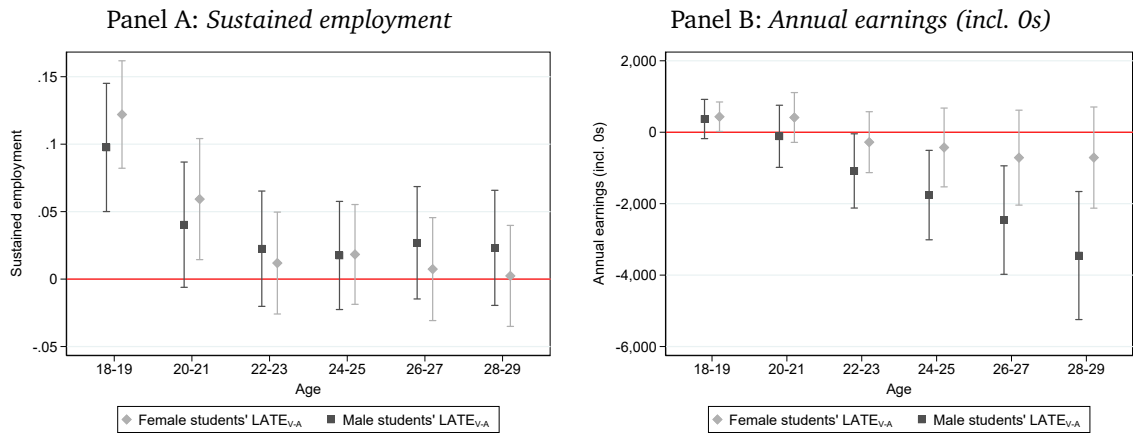


Figure 4.6. Age-effect profiles at the vocational-academic margin.

Notes: These figures plot estimates of the margin-specific effect of vocational *vs.* academic education, LATE$_{V\text{-}A}$, on sustained employment (panel A) and annual earnings (panel B) across different ages by gender. For each two-year age bin, we average annual earnings and sustained employment (if observed) over the two successive years. For comparability across the whole age range only earnings (and employment) from employed, but not from self-employed, work are included. The LATE$_{V\text{-}A}$ estimates are corrected for estimated differences in predicted outcomes between compliers groups as illustrated in 4.4.2. 95% confidence intervals are based on block bootstrapped standard errors at the LSOA×cohort level using 500 iterations.

to 28–29.[34] The figure reveals that vocational education confers an initial employment advantage: male and female vocational-track graduates are respectively 10 and 12 pp more likely to be in sustained employment at ages 18–19. This halves by ages 20–21 and essentially disappears by ages 22–23, after which the effect remains close to and indistinguishable from zero (positive point estimates suggest that, if anything, vocational education continues to confer a small employment advantage). Corresponding to this, vocational education also confers a small initial advantage in terms of (net) earnings. However, the earnings premium deteriorates close to linearly over the observed age range with point estimates for both genders turning negative by students' mid-twenties. The negative trend is more pronounced for men for whom the negative effect on net earnings becomes statistically significant at age 22-23.

Initial employment and earnings advantages for the vocationally educated are in line with what theory predicts: the occupational skills acquired in the vocational track facilitate the school-to-work transition, leading to earlier labour market entry and hence higher earnings early on. However, here these initial returns are extremely short-lived. Therefore, their erosion is unlikely to be driven by faster depreciation of specific human capital and lower labour market adaptability for the vocationally educated, as predicted by theory and found in other studies. Over the age range we observe, it is more plausible that students with academic education enter the labour market later but in higher-paying jobs with better wage progression. Note that the estimated effect dynamics make it likely that earnings differences continue to grow. Earlier obsolescence of occupation-specific skills will only add to these differences as students grow older.

[34] Outcomes are the average over the two ages of each age band. Note that for this exercise we focus on earnings *including* zeros to keep the sample consistent across different ages. Further note, that (earnings from) self-employment are excluded to ensure comparability across the age range, because those we only observe from 2014 onward. Comparison with the estimates from Table 4.6 shows that this has little impact on the estimates for women but leads to an underestimate of the effect for men.

Consequently, we conceive of our effect estimates for the age of 29 as a lower bound of how the earnings differential between vocational and academic education will develop as students' careers progress.

4.4.4 Effects across the Distance Grid

The results presented thus far are based on global regression models for the reduced form and first stage equations that restrict the effect of the distance instruments to be constant and linear (in logs). This implies that any heterogeneity in the effect of vocational education among compliers is muted. In this section, we relax the parametric restrictions and let the coefficients vary across different values of the distance instruments, (Z_A, Z_V), thus allowing us to study how the effect of vocational education varies for marginal students who live at different distances from their closest academic and vocational colleges. This is interesting because compliers living farther away from a particular option (e.g., academic college) need to have higher underlying preferences for that option (e.g., academic education) for the higher costs of enrolling (travelling) to be offset—otherwise they would not be marginal (i.e., compliers). Accordingly, local estimates at different distance points correspond to returns for students with different unobserved preferences for vocational and academic education, allowing us to test whether preferences reflect individuals' comparative advantage for different types of education, i.e., whether students select into treatments based on (expected) gains.

To do so, we estimate the reduced form and first stage equations (4.3)–(4.8) as local linear regressions across a two-dimensional grid defined by the first, third, fifth, seventh and ninth deciles of the residualised distances to academic (Z_A) and vocational college (Z_V).[35] This yields 25 grid point-specific decompositions of net MTEs into their margin-specific components analogous to the decomposition of LATE_V from above. Here, we focus on the margin-specific MTEs of vocational *vs.* academic education (i.e., $\text{MTE}_{V\text{-}A}$) on net annual earnings, referring the interested reader to the appendix for more extensive results.

Figure 4.7 visualises the estimation results for males (panel A) and females (panel B). Each of the five diagrams plots $\text{MTE}_{V\text{-}A}$ estimates across different distances to vocational college, holding distance to academic college fixed at the first, third, fifth, seventh and ninth decile, respectively. Consistent with students selecting into tracks based on gains, we find that for males the returns to vocational education increase with distance to vocational college and decrease with distance to academic college. To exemplify this, first consider the rightmost effect estimate in the leftmost diagram, i.e., deciles $(z_V^d = 9, z_A^d = 1)$, pertaining to vocational-academic compliers who live closest to an academic college and furthest from a vocational college. Since they are willing to travel the longest distance to enrol in vocational education, these marginal students must

[35] We residualise the distance instruments with respect to the control set $\boldsymbol{X}$ to ensure that evaluation points only differ in education choices' costs due to differences in distance but not due to compositional changes. Note that, to keep the dimensionality manageable, our estimates are local with respect to the two distance instruments only: we estimate locally weighted regressions, where all variables enter additively but with coefficients that are allowed to vary arbitrarily across different $(Z_V = z_V, Z_A = z_A)$ evaluation points. We weight observations by their distance to the evaluation point using a two-dimensional Epanechnikov kernel function with bandwidth set to two standard deviations of the respective residualised distance.

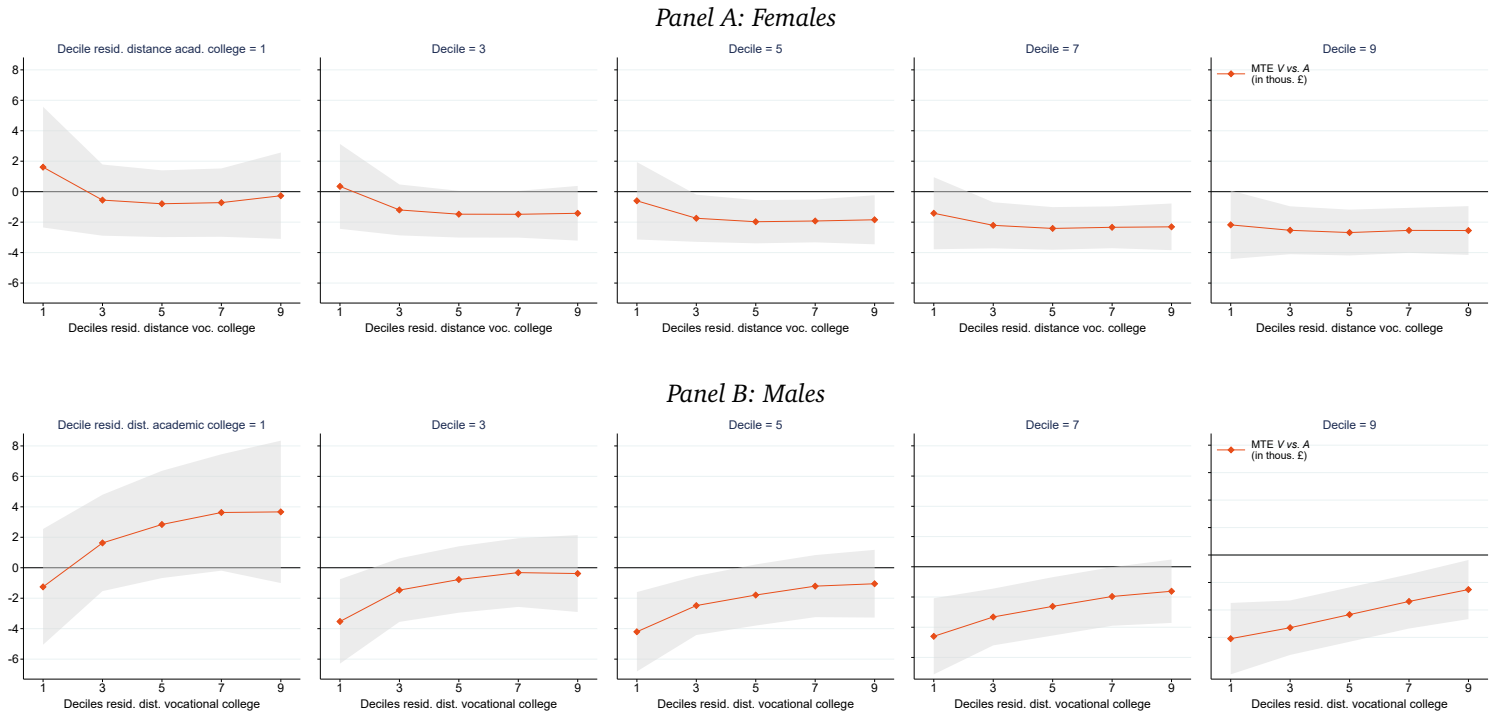


Figure 4.7. MTEs of vocational *vs.* academic education on earnings across the distance grid.

Notes: This figure plots, separately by gender, the $\text{MTE}_{V\text{-}A}$ estimated at different locations of the two-dimensional grid defined by the two residualised distance instruments. We first residualise the two distance instruments Z_V and Z_A with respect to the control set. Then, we define a 5-by-5 distance grid corresponding to the 1st, 3rd, 5th, 7th and 9th deciles of the two residualised distances. Finally, we estimate a series of local linear reduced form and first stage regressions for each gridpoint, weighting observations by their distance to the grid point using a two-dimensional Epanechnikov kernel function with a bandwidth of two standard deviations in either dimension. The MTE estimates are constructed from the local regression coefficients analogously to the main estimation. 90% confidence intervals are based on 500 block-bootstrap iterations clustering at the LSOA×cohort level.

have the strongest preferences for vocational education (or, equivalently, the strongest dislike for academic education). These compliers experience positive returns to vocational education of nearly £4,000. Staying within the same diagram, returns decrease as distance to vocational college becomes smaller, i.e., as we consider compliers with a weaker underlying preference for vocational education. Next, consider the opposite grid point, i.e., the leftmost effect estimate in the rightmost diagram, corresponding to deciles ($z_V^d = 1$, $z_A^d = 9$). This effect pertains to vocational-academic compliers with the strongest preferences for academic education, given that they would travel the farthest to enrol in academic college. These students experience large negative returns to vocational education of around -£6,000. Again, if within the same diagram, we consider larger distances to vocational college, the estimated returns become less negative. For females, effects also decrease with distance to academic college, though we do not find systematic variation in returns with respect to distance to vocational college.

Overall, this analysis confirms that for most students at the vocational-academic margin returns to vocational education are negative. However, it also reveals substantial treatment effect heterogeneity that systematically relates to students' selection into treatments: returns to vocational education are less negative (and even positive) for those with a stronger preference for vocational education. This suggests that students understand their comparative advantage and select into educational tracks based on their (expected) gains. Therefore, vocational education might not be detrimental for all students: non-marginal students with strong preferences for the vocational track (i.e., 'always-takers') might well benefit from it. Against this backdrop, the next section takes a closer look at the group of marginal students to whom our margin-specific effects estimates apply.

4.5 Understanding $LATE_{V\text{-}A}$

4.5.1 Characterising Compliers

Our IV estimates pertain to marginal vocational education students who would have made a different post-16 education choice had they lived at closer or further from a vocational college. We found that the vast majority of these students are choosing between vocational and academic education and subsequently focused on average effects for this group, i.e., on $LATE_{V\text{-}A}$. To better understand how representative the LATE is for effects in the wider population, this section characterises the group of compliers along several dimensions.

We first assess the overall size of the complier population. Dahl *et al.* (2014) show that it can be estimated by comparing treatment take-up at the extreme values of the instrument: the share of students who choose D_V at maximum distance to vocational college equals the share of always-takers, while the share of students who do not choose D_V at minimum distance to vocational college equals the share of never-takers. The rest are compliers would have chosen a different post-16 education at at least some point of the distance distribution. We estimate the share of always-takers as the mean of D_V at the 99^{th} percentile of the residualised Z_V distribution

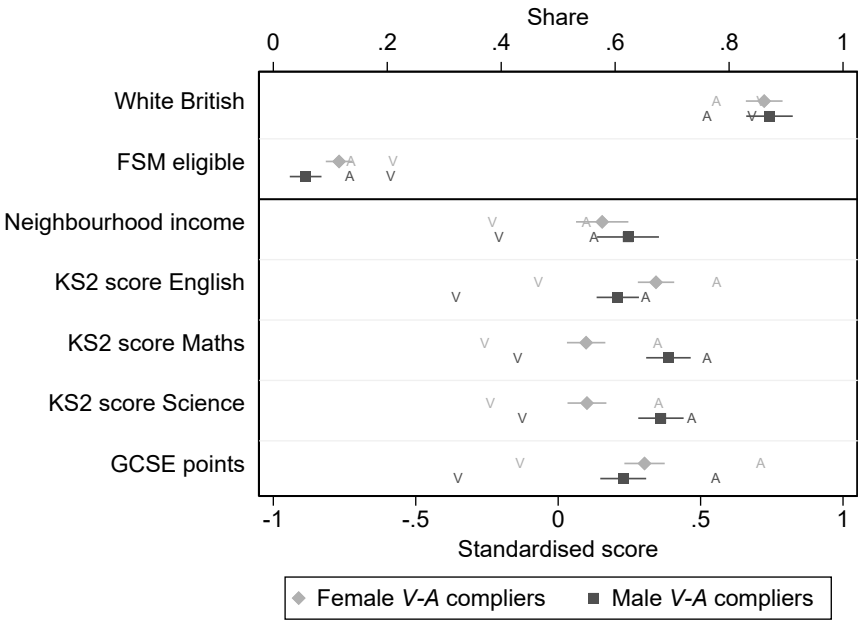


Figure 4.8. Complier characteristics.

Notes: The figure shows estimated mean characteristics of vocational-academic compliers with associated 95% CIs, alongside the sample means for the vocational (V) and academic (A) treatment groups, separately by gender. The figure is based on the estimation sample of students from schools without sixth form. Characteristics plotted in the lower part of the panel are standardised within the total student population. Neighbourhood income refers to the inverted index of neighbourhood income deprivation, so that higher values indicate less deprived neighbourhoods.

and the share of never-takers as the mean of $(1 - D_V)$ at the 1^{st} percentile. For females, they equal 0.48 and 0.39 respectively whereas for males the share of vocational always-takers is slightly higher at 0.50. Accordingly, compliers make up about 13% and 11% of our estimation samples for females and males respectively (i.e., of students from schools without sixth form). Of these, 90% and 80% are at the vocational-academic education margin, respectively.

Next, we characterise vocational-academic compliers by observables. Figure 4.8 plots estimates for *V-A*-compliers' mean predetermined characteristics along with means for all vocational- and academic-track students (separately for females and males).[36] Marginal students are more likely to be White British and, on average, of higher socio-economic status than both vocational- and academic-track students, as indicated by a lower prevalence of FSM eligibility and higher average neighbourhood income. In terms of previous achievement, marginal students lie in between the two tracks, though much closer to the academic than the vocational track: their end-of-primary (KS2) and end-of-secondary-school (GCSE) test scores are much higher than the vocational and only slightly lower than the academic average. This is unsurprising because students who are considering academic upper secondary education in Sixth Form Colleges need to meet the minimum GCSEs requirements to be eligible for admission, whereas many vocational students study courses at lower levels without strict entry requirements (i.e., Level 2 and

[36]Remember that *V-A*-compliers' average value in some scalar predetermined characteristic, C, can be estimated the same way we estimate the academic education PO for vocational-academic compliers (from equations (4.5) and (4.6)) after replacing YD_A with CD_A in equation (4.5).

Table 4.8. Comparing OLS and IV estimates for the vocational *vs.* academic effect.

Dependent variable:	Sustained employment			Annual earnings (incl. 0s)			Positive earnings (excl. 0s)		
Estimation method:	OLS (1)	rw-OLS (2)	IV (3)	OLS (4)	rw-OLS (5)	IV (6)	OLS (7)	rw-OLS (8)	IV (9)
Effect for females	-0.063	-0.057	-0.006	-3,754	-3,781	-781	-3,596	-3,650	-1,449
	(0.002)	(0.002)	(0.020)	(57)	(69)	(651)	(57)	(69)	(668)
Effect for males	-0.028	-0.024	0.013	-1,792	-1,834	-2,014	-1,617	-1,717	-2,503
	(0.002)	(0.002)	(0.022)	(71)	(91)	(857)	(68)	(87)	(843)

Notes: For each of the three outcomes of interest, the table reports, separately by gender, the effect of vocational *vs.* academic education under three different models. Columns 1,4 and 7 report the effects as estimated with a conventional OLS regression (same estimates as in Table 4.2); columns 2,5 and 8 report the effects as estimated by an OLS regression where observations are reweighted to be comparable with *V-A* compliers; finally, columns 3,6 and 9 report the corrected IV estimates of the $LATE_{V\text{-}A}$ from Table 4.7. Standard errors, reported in parentheses, are clustered at the LSOA×cohort level.

1 courses). Overall, the vocational-academic compliers for whom we estimate causal effects are academically apt students from socio-economically advantaged backgrounds, likely to do well in academic environments.

4.5.2 Comparison to OLS

From a policy perspective, returns for marginal students are important because their treatment choices are responsive to incentives—and thus likely also to a broad set of policies. While the previous section showed that this group represents a non-negligible portion of the population, it also showed that compliers are not representative of the average student. When stratifying effects along the distance grid, we found that within the subgroup of *V-A*-compliers there is heterogeneity in the return to vocational education. Accordingly, it would be premature to dismiss the efficacy of vocational *vs.* academic education for all students solely on the basis of the negative IV estimates for $LATE_{V\text{-}A}$.

OLS has potential for learning about returns in the broader population because, absent remaining unobserved confounders, it estimates alternative-specific ATEs. Accordingly, it is worthwhile to compare our plausibly causal but local IV estimates with the possibly biased but global OLS results from Table 4.2. If treatment effects were homogeneous, comparison of IV and OLS estimates would allow one to directly infer the severity of selection bias. However, if treatment effects are heterogeneous, like shown to be the case here, the effect estimates can differ even in the absence of selection bias, simply because the average causal effect among compliers differs from the average effect in the overall population. Part of this heterogeneity can be associated with differences in observed characteristics: in the previous section we have seen that *V-A*-compliers are of above-average academic ability and socio-economic status. To enable a meaningful comparison, we reweight the OLS estimates to account for these differences following the procedure outlined in Bhuller *et al.* (2020): first, we split both gender-specific samples into eight mutually exclusive and collectively exhaustive subgroups based on quartiles of previous achievement and neighbourhood income. Next, we estimate the relevant first-stage equation (4.6) separately for each subsample, allowing us to calculate the proportion of *V-A*-compliers by subgroup. Then, we

reweight the estimation sample so that the proportion of compliers in a given subgroup matches the share of the estimation sample for that subgroup.

Table 4.8 compares regular controlled OLS estimates, OLS estimates using the complier-reweighted sample ('rw-OLS') and our margin-specific IV estimates for the effect of vocational *vs.* academic education on our three labour market outcomes of interest, separately by gender. The results suggest that the differences between the IV and the OLS estimates cannot be accounted for by heterogeneity in effects, at least due to observables. The effect estimates from reweighted OLS are generally very close to those from regular OLS. The most striking differences between OLS and IV are to be found in terms of employment: OLS suggests that vocational education substantially reduces students' probability to be in sustained employment, but IV shows that this is an artefact of selection as causal employment effects are null for both genders. In terms of earnings, the selection patterns differ by gender: for females, OLS substantially overestimates the earnings penalty, while for males, if anything, OLS underestimates it. This suggests that self-selection into the vocational track is more negative for females.

4.5.3 Mechanisms: Educational Attainment and Progression

While standard human capital (and signalling) models would predict that vocational education is an improvement over no upper secondary education—a prediction fully consistent with the large (but imprecise) earnings premium we find at the vocational *vs.* no post-16 education margin—the relative merits of vocational *vs.* academic upper secondary education are much less clear in theory. They depend crucially on the types of skills that students acquire in the respective educational tracks and on the opportunities for higher education they face afterwards. To get a better understanding of the mechanisms behind the large (for males) and moderate (for females) earnings penalties from vocational education we find at this margin, in this section we investigate how vocational enrolment affects students' educational attainment and progression.

First, we study the types of courses compliers choose when enrolled in vocational *vs.* academic upper secondary education. Figure 4.9 plots the margin-specific vocational and academic PO estimates for the teaching hours-weighted share of Level 3 and of academic courses, respectively. For male compliers, academic education increases the share of Level 3 courses from 92% to 98%. For female compliers, a tiny increase from 83% to 84% is insignificant. Beyond the positive effect for males, these numbers show that vocational-academic compliers mainly study Level 3 courses *regardless of track choice*, highlighting their positive selection compared to the average vocational-track student who predominantly studies courses below Level 3 (see Figure 4.3). Similarly, at approximately 60%, the share of academic courses is high even when these students attend a vocational college (the average for vocational-track students is 18%; see Figure 4.1). Nevertheless, academic enrolment substantially increases this share to approximately 77% for females and to almost 90% for males. The greater difference in curricula among males offers a first potential explanation for the larger earnings penalty they experience from vocational education.

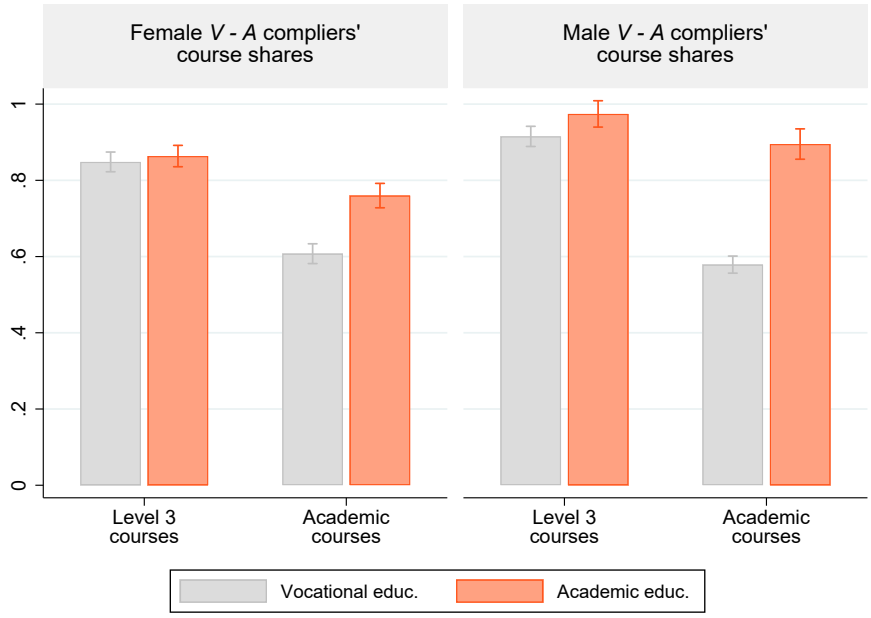


Figure 4.9. Compliers' course enrolment potential outcomes.

Notes: This figure plots, separately by gender, the estimated mean academic and vocational education potential outcomes (POs) for the share of Level 3 courses and the share of academic courses for vocational-academic compliers. The shares are based on all courses (of more than one month) started by students within 24 months or their first studying spell and are weighted by the recorded number of teaching hours. Academic courses refer to A(AS)-level and GCSEs qualifications. The figure is based on the estimation sample of students from schools without sixth form, restricted to individuals with positive earnings at age 29. The vocational education POs are corrected for differences in predicted outcomes between complier groups (as illustrated in 4.4.2). 90% CIs are based on block bootstrapped standard errors at the LSOA×cohort level using 999 iterations.

Second, we study how these education choices translate into upper-secondary attainment. We consider two outcomes: attainment of at least one qualification at Level 3 and a full Level 3. The former is the expected educational level by age 18 and required by employers for the majority of jobs. The latter is crucial for entering higher education. The first two rows of Figure 4.10 plot complier PO estimates, next to the implied margin-specific treatment effect, $LATE_{V\text{-}A}$, for these outcomes, separately for females and males. We find that females are equally likely to achieve these outcomes in either track, while male compliers are about 5–6 pp more likely to do so if they enrol in the academic track. This confirms that male compliers' upper secondary educational experience is more sensitive to track than that of females.

Finally, we consider students' educational progression to post-secondary education. We first consider apprenticeships which are commonly perceived as a positive outcome to complement the vocational route. Apprenticeships combine workplace training with classroom-based learning and have been shown yield substantial returns in the English labour market (Cavaglia *et al.*, 2020). For females, the effect of vocational education on starting an apprenticeship is small and insignificant, but for males it is significant and substantial in size at about 3.5 pp, or 23%. Next, we consider the canonical academic route of higher education (HE). For females, we find no significant difference in the probability of starting or completing a university degree although, importantly, we can also rule out large negative effects. Still, vocational education seems to channel them into less selective (non 'pre-1992') universities. For males, vocational education

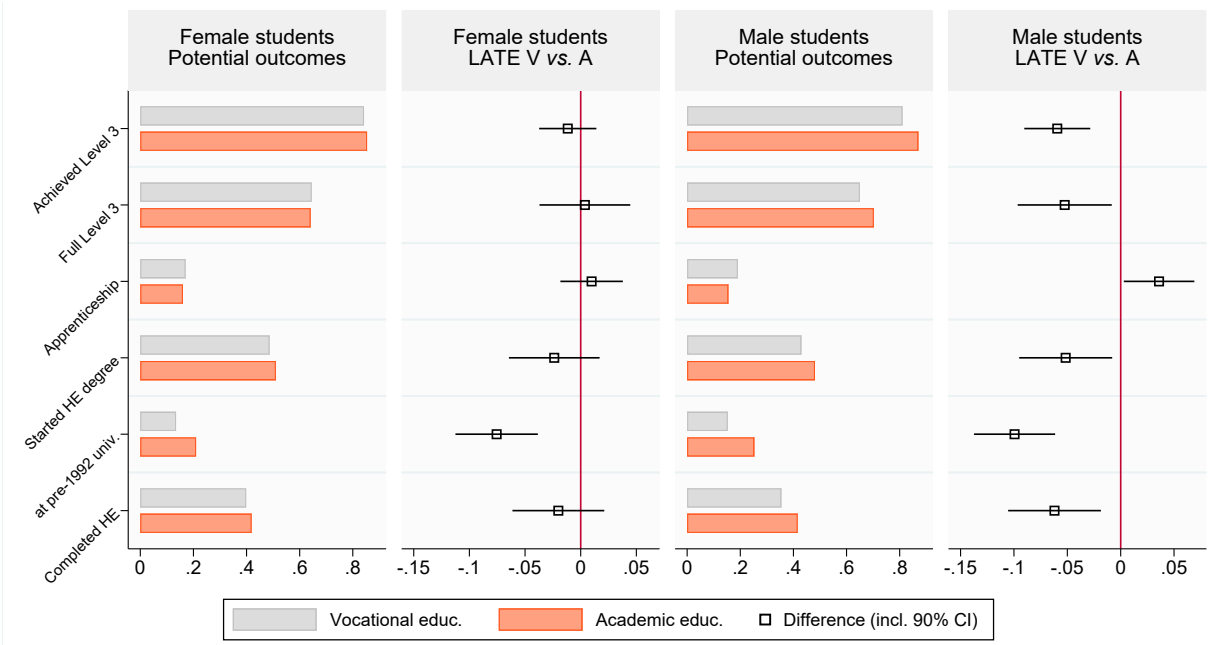


Figure 4.10. Effects and potential outcomes for education outcomes at the vocational-academic margin.

Notes: This figure plots estimates for vocational-academic compliers' mean potential outcomes (POs) and the associated LATEs$_{V\text{-}A}$ for a range of indicators of educational attainment, separately by gender. The figure is based on the estimation sample of students from schools without sixth form, restricted to individuals with positive earnings at age 29. The vocational education POs are corrected for differences in predicted outcomes between complier groups (as illustrated in 4.4.2). 90% CIs are based on block bootstrapped standard errors at the LSOA×cohort level using 999 iterations.

substantially decreases the probability to start or complete an HE degree by 5 and 6 pp (or 10% and 14%), respectively. Moreover, male compliers are far less likely to attend more selective universities.

Overall, these results show that vocational education harms the educational attainment of male marginal students (with the exception of progression to apprenticeships). For females the effects are largely null, which is likely explained by the fact that their upper secondary education curriculum is less affected by track choice. Some arguments for expanding access to vocational programmes assume that students at the margin between vocational and academic education do not end up in university anyway and would therefore benefit from a more occupation-specific curriculum. Our findings suggest that, at least for England—a country where university participation has greatly expanded over the last 25 years—this argument is not particularly salient. The average marginal student is relatively apt academically and, while far from all marginal students who attend the academic track go on to complete university, more than 40% do—half of them even study in more selective pre-1992 universities. It has to be stressed that marginal students attain relatively high levels of education even when attending the vocational track; but, at least for males, lower HE completion rates mean that many students may end up missing out on graduate jobs opportunities through choosing the vocational over the academic track.

To quantify how these differences in HE completion could contribute to the negative earnings effect of vocational education we perform a partial Oaxaca-Blinder decomposition. In particular,

Table 4.9. Decomposition of vocational *vs.* academic earnings effect.

	Female students (1)	Male students (2)
Earnings effect ($\text{LATE}_{V\text{-}A}$)	-1,449	-2,503
Effect explained by HE completion	-472	-1,661
HE completion effect ($\text{LATE}^{\text{HE}}_{V\text{-}A}$)	-0.020	-0.062
Vocational $\wedge$ HE PO (Y_V^{HE})	23,531	26,798

Notes: The upper panel of the table reports the estimated earnings effect at the vocational *vs.* academic education margin (row 1), as well as the part of the effect that is explained by estimated differences in HE completion (holding fixed returns to HE completion) (row 2). This is computed by multiplying the estimated effect on HE completion (row 3) by the average earnings of *V-A*-compliers who choose the vocational track and go on to complete a HE degree (row 4). Results are based on the estimation sample of students from schools without sixth form, restricted to individuals with positive earnings at age 29.

we multiply the differences in HE completion plotted in Figure 4.10 by the average potential earnings of compliers in vocational education who go on to complete a university degree[37]: this product tells us how much more compliers in vocational education would earn, on average, if they were equally likely to complete a university degree (while holding fixed the earnings level). Table 4.9 reports the results for both genders: for females, taking the point estimates at face value, we find that the negative difference in HE completion of 2.3 pp corresponds to an average reduction of earnings of 532 GBP, equivalent to 40% of the main effect on earnings; for males, the negative difference of 6.6 pp corresponds to a reduction in earnings of 1760 GBP, or 63% of the earnings effect. These results indicate that marginal vocational students' more limited progression to HE can play a substantial role in explaining the earnings loss they suffer when diverted from academic education.

4.5.4 External Validity

Because our identification strategy relies on cross-instrumenting education choices, we restricted attention to students from secondary schools without sixth form for whom both distance instruments play a role. In this section, we explore to what extent our results are informative of

[37] These are identified following the same procedure as for the identification of the potential outcomes in the main analysis. In particular, we estimate the equivalent of the reduced form and first stage equations (4.3) and (4.4) except that in both equations we replace D_V with an indicator for enrolling in vocational education *and* completing a HE degree D_V^{HE}. Analogously to the main identification results, we have that the ratio between the two coefficients on Z_V identifies the relevant potential outcome. More formally, and abstracting from the linear parameterisation, it can be shown that under the usual assumptions A1, A2 and A3, the average PO for vocational-academic compliers at point $(Z_V, Z_A, \boldsymbol{X}) = (z_V, z_A, \mathrm{x})$ who enrol in the vocational track and subsequently complete a HE degree is identified as a ratio of partial derivatives as follows:

$$\lim_{z'_V \uparrow z_V} \mathbb{E}\left[Y_V^{HE} \mid D(z'_V, z_A, \mathrm{x}) = V,\, D(z_V, z_A, \mathrm{x}) = A\right]$$
$$= \frac{\partial \mathbb{E}\left[Y D_V^{HE} \mid Z_V = z_V, Z_A = z_A, \boldsymbol{X} = \mathrm{x}\right]}{\partial Z_A} \Bigg/ \frac{\partial \mathbb{E}\left[D_V^{HE} \mid Z_V = z_V, Z_A = z_A, \boldsymbol{X} = \mathrm{x}\right]}{\partial Z_A}.$$

Table 4.10. Net effect of vocational education across secondary school samples.

	Female students			Male students		
	Sustained employment	Annual earnings (incl. 0s)	Annual earnings (excl. 0s)	Sustained employment	Annual earnings (incl. 0s)	Annual earnings (excl. 0s)
	(1)	(2)	(3)	(4)	(5)	(6)
A. Students from schools without sixth form (main analysis sample)						
Share *V*-compliers (both margins)	0.13	0.13	0.14	0.11	0.11	0.12
Of those at *V-A* margin	0.90	0.90	0.94	0.80	0.80	0.81
Implied share *V-A*-compliers	0.12	0.12	0.13	0.09	0.09	0.10
Net effect vocational educ. ($LATE_V$)	0.007	22	-540	0.023	-979	-1,681
	(0.032)	(846)	(893)	(0.033)	(1,154)	(1,126)
B. Students from schools with sixth form (excluded from main analysis)						
Share *V*-compliers (both margins)	0.07	0.07	0.06	0.07	0.07	0.07
Of those at *V-A* margin	0.79	0.79	0.82	0.63	0.63	0.66
Implied share *V-A*-compliers	0.05	0.05	0.05	0.04	0.04	0.04
Net effect vocational educ. ($LATE_V$)	0.117	-1,341	-4,657	-0.012	-4,109	-5,276
	(0.049)	(1,432)	(1,564)	(0.043)	(1,661)	(1,627)

Notes: This table compares the net effect of vocational education ($LATE_V$) and complier shares between the estimation sample containing only students from secondary schools without a sixth form (panel A) and students from secondary schools with a sixth form that are excluded from the main analysis (panel B). The first row of each panel reports the overall share of vocational education compliers in the sample; the second rows report the share of those at the vocational *vs.* academic margin (λ); the third rows report the implied proportion of *V-A* compliers in the sample obtained from multiplying the first two rows. Finally, the fourth rows report $LATE_V$ estimated by a 2SLS regression that instruments the vocational treatment indicator D_V with Z_V (while controlling linearly for Z_A and the control set). Standard errors for $LATE_V$ are reported in parentheses.

the returns to vocational education for the remaining population of English students attending secondary school with sixth form. Because these students can attend the academic track on their own secondary school, their post-16 education choice does not respond to distance to academic college and, hence, for them we cannot recover alternative-specific effects. However, their decision to enrol in the vocational track is still affected by distance to vocational college, Z_V. Accordingly, also for them the *net* complier treatment effect of vocational education, $LATE_V$, is identified by the univariate IV that instruments vocational education, D_V, with Z_V. Also identification of the share of compliers at the vocational-academic margin, λ, does not rely on Z_A, so that it is identified in this sample, as well. Hence, to gauge in how far our estimates for the margin-specific returns of vocational *vs.* academic education extrapolate to students from schools with sixth form, we compare estimated net returns and margin-specific complier shares between the two samples to at least infer plausible magnitudes for the margin-specific effect that we cannot estimate directly.

Table 4.10 reports the results from this exercise: panel A refers to students from schools without sixth form (i.e., the main analysis sample) and panel B refers to students from schools with sixth form (i.e., those excluded from the main analysis). The first three rows in each panel quantify the size of the complier population: the first reports the overall share of vocational education compliers in the population, the second reports estimates of share of compliers at the

vocational-academic margin, λ, and the third reports the product of those two shares, i.e. the implied population share of *V-A*-compliers. There are substantially more marginal students in our analysis sample (13% and 11% of females and males, respectively) than in the excluded part of the population (7% of both genders). As expected, the majority of students whose post-16 education choice is responsive to changes altering the (relative) attractiveness of the vocational track attends schools without sixth form, where the option of vocational education is more salient.

For sustained employment, the net complier treatment effect is much larger for female students from schools with sixth form than in our estimation sample, while for males it is close to zero in either case. For earnings, in contrast, the net effect of vocational education is much more negative for (female and male) students from schools with sixth form, especially when conditioning on positive earnings. At the same time, the share of compliers at the vocational-academic margin, for whom we would expect a negative effect of vocational education given the previous results, is smaller in this part of the population. Because we deem it highly unlikely that the effect of vocational *vs.* no post-16 education is strongly negative for these students, this suggests that the negative earnings effect of vocational *vs.* academic education is even more pronounced for students from schools with sixth form. This is plausible because the quality of the academic track might be higher in secondary schools than in (academic) Sixth Form Colleges. Accordingly, our main estimates are likely to be a lower bound for the earnings penalty from vocational education for all marginal students in English student population.

4.6 Discussion and Conclusions

In recent years, many countries have witnessed a renewed policy interest in expanding and improving vocational education to make education systems more inclusive and fit for changing economic needs. Internationally, systems with widespread firm-based vocational education provision, like Germany and Switzerland, are examples of the merits of vocational education. It is less clear how effective such policies would be in more market-oriented economies with weaker traditions of vocational education, like the UK and the US. Our paper contributes to this debate by delivering plausibly causal estimates of alternative-specific returns to vocational education in England.

Compelling evidence on returns to vocational education is scant because comparisons of vocational with other students are contaminated by strong self-selection. Additionally, identifying more policy-relevant effects among marginal students (i.e., students whose education choices are responsive to changes in incentives) is challenging when they can choose among more than two (unordered) alternatives. Conventional IV methods designed to circumvent self-selection and recover effects among marginal students cannot identify alternative-specific returns unless strong assumptions are imposed. We address these challenges in the English setting where, upon completing compulsory education at age 16, students can choose to enrol in an academic track, enrol in a vocational track or leave education. We exploit that the academic and the vocational

track are linked to distinct post-16 institutions to construct two alternative-specific IVs based on students distance to the nearest respective provider. Thus equipped, we can apply an identification strategy proposed by **?** to estimate, separately by gender, the return of vocational *vs.* academic education and the return of vocational *vs.* no post-16 education among compliers.

Our analysis shows that the vast majority of marginal vocational students is choosing between vocational and academic education, not considering the option of no post-16 education. For these students, we find large negative effects of vocational education on earnings at age 29, especially among males, and null effects on the probability of employment. Given that returns are negative from students' early twenties onwards, they are not due to faster depreciation of occupation-specific skills but due to students entering lower-wage jobs with weaker wage progression. Characterising the group of marginal vocational-academic students by observables reveals that they are on average higher achieving and from more advantage backgrounds than typical vocational students and, hence, likely to do well in academic environments. To explore the mechanisms behind the negative earnings effects, we study effects on education outcomes and find that vocational education impairs these students' educational attainment and progression. A decomposition exercise suggests that for males around 60% of the negative earnings effect can be explained by a lower probability of obtaining a university degree. We find that returns are negative for most marginal students but detect effect heterogeneity that is consistent with comparative advantage: marginal students living further away from vocational colleges (who must have higher unobserved preferences for the vocational track) exhibit more modest negative returns. Among males, we even find positive returns to vocational education among those with the highest relative preferences for the vocational track, suggesting that it may well be beneficial for a large share of non-marginal students (i.e., so-called 'always-takers').

Overall, these results stand in marked contrast with other recent quasi-experimental studies focusing on returns to vocational *vs.* academic education which tend to find either positive or at least non-negative effects for average marginal students at comparable ages. However, with the exception of Brunner *et al.* (2021), these studies focus on Nordic European countries characterised by very different education and labour market institutions from the UK (Bertrand *et al.*, 2021; Birkelund and van de Werfhorst, 2022; Silliman and Virtanen, 2022). In these countries, upper-secondary vocational education is better integrated with firm-based training (the predominant learning mode in Denmark) and offers more equal pathways into tertiary education. For example, in contrast to our analysis, Silliman and Virtanen (2022) find that the Finnish vocational track does not reduce university enrolment compared to the academic track, potentially explaining part of the positive returns they identify.

Differences in labour market institutions likely compound these gaps. Birkelund and van de Werfhorst (2022), for example, find that vocational students diverted from the academic track in Denmark enter occupations with lower prestige, without this translating into significant earnings differences. They attribute this to widespread collective wage agreements and strong trade unions, which reduce earnings differences between occupations. For a comparison, collective bargaining coverage in the UK in 2015 stood at 28% compared to 83% and 89% in Denmark

and Finland respectively, with trade union density figures similarly far apart.[38] Accordingly, one interpretation of our results is that in less egalitarian countries with high returns to high-skilled occupations, differences in skills acquisition across tracks, exacerbated by differences in education progression, translate into large labour market penalties very early in students' careers. Importantly, if vocational skills depreciate faster in the longer-term, these penalties are set to increase with our estimates likely being a lower bound of the life-time loss in earnings caused by vocational education.

Another possible explanation is that the comparison between vocational and academic education masks differences in education quality across tracks. After all, studying the effect of enrolment in selective stand-alone technical high schools in Connecticut, where labour market institutions are arguably more similar to England than to Nordic Europe, Brunner *et al.* (2021) find positive returns for men. These technical high schools are typically better-resourced (e.g., lower teacher-student ratios) than counterfactual schools in the district. Note, however, that our findings are unlikely to be confounded by institutional differences: by focusing on students from secondary schools without post-16 provision, we mainly compare individuals enrolled in vocational 'Further Education Colleges' with those in academic 'Sixth Form Colleges', which tend to have similar size, organisational structure and funding arrangements. Indeed, when we estimate (net) effects in the full sample, thus including smaller and typically better-resourced *schools* in the academic counterfactual to vocational education, our results become even more negative. Of course, average peer achievement differs dramatically between vocational and academic tracks and absent further data, we cannot fully rule out differences in other inputs, but simple quality differences are unlikely to fully explain our findings.

Finally, beyond any objective difference in education quality, vocational education in England arguably suffers from a particularly unfavourable reputation, resulting from deeply rooted social (mis-)perceptions, as well as long-standing dysfunctionalities successive governments have been slow or reluctant to correct (see Wolf, 2011; Musset and Field, 2013). Correspondingly, it is possible that students are penalised in the labour market because of a negative signal associated with vocational education; although this is unlikely to fully explain our results and cannot explain the effect heterogeneity we uncover.

Our findings at the margin between vocational and no post-16 education are more inconclusive. This is the result of the involved nature of the identification procedure and the small proportion of marginal students at this margin (particularly among females), which make results at this margin are arguably less relevant from a policy perspective. Nevertheless, we find an indication that students who enrol in vocational education as opposed to dropping out of education may experience positive returns, with significant results for marginal male students with highest unobserved preference for continuing into vocational education. Perhaps more importantly, the fact that the two separate alternative-specific returns appear to be divergent warns against relying on a conventional IV strategy that only identifies the *net* effect of vocational

[38] Source: OECD Database on Institutional Characteristics of Trade Unions, Wage Setting, State Intervention and Social Pacts, available at https://stats.oecd.org/Index.aspx?DataSetCode=CBC.

education (a weighted average of the two margin-specific effects). We show that this would yield a smaller and insignificant estimate of the returns to vocational education, thus shrouding the large negative effects for the majority of marginal students and nurturing an ambiguous and more positive impression of vocational education in England than warranted.

Our alternative-specific estimates dispel this ambiguity and can help formulate more suitable, targeted policy interventions. In particular, our findings dissuade from an outright expansion of vocational education in England as this would mostly result in students being diverted from the academic track with an associated large loss in their earnings. Efforts to recruit people into vocational education should instead be targeted at individuals at risk of dropping out. For example, students who enrolled into vocational education as a result of recent increases in the minimum school leaving age from age 16 to 18 are likely to have benefited, even if the wider implications might be limited due their small proportion.

More importantly, strengthening the emphasis and availability of apprenticeships in the vocational track and facilitating pathways into tertiary education would likely boost returns to vocational education across the board, while potentially attracting those marginal students most likely to benefit from vocational education (even at the margin with academic education). While these changes would help to reduce inequality of opportunities across secondary school tracks, the discussion above suggest that effects of vocational education are crucially mediated by labour market institutions, so that education policy alone can be no silver bullet.

Appendix A: Additional Tables and Figures

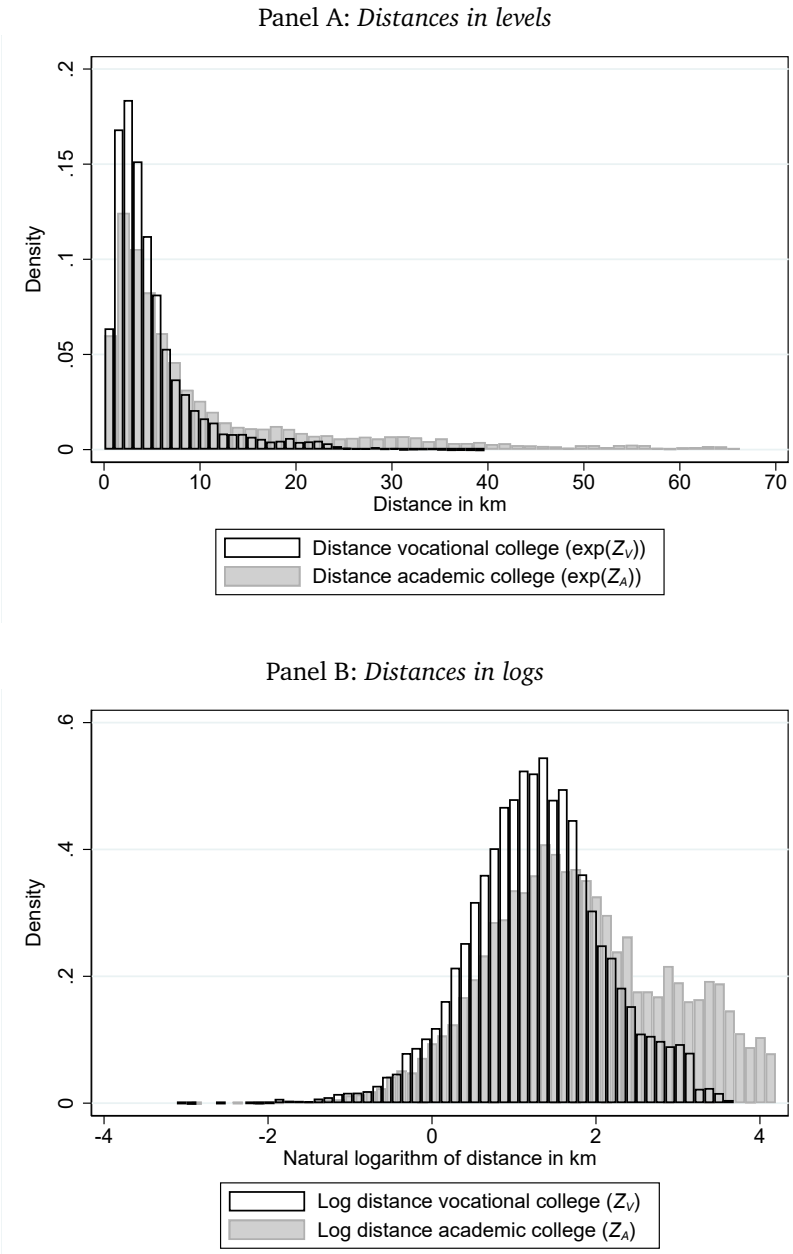


Figure A4.1. Distributions of the distance instruments in levels and logs.

Notes: This figure plots histograms of the distribution of students' distance (in km) to their closest vocational and academic college in levels (Panel A) and natural logarithms (Panel B).

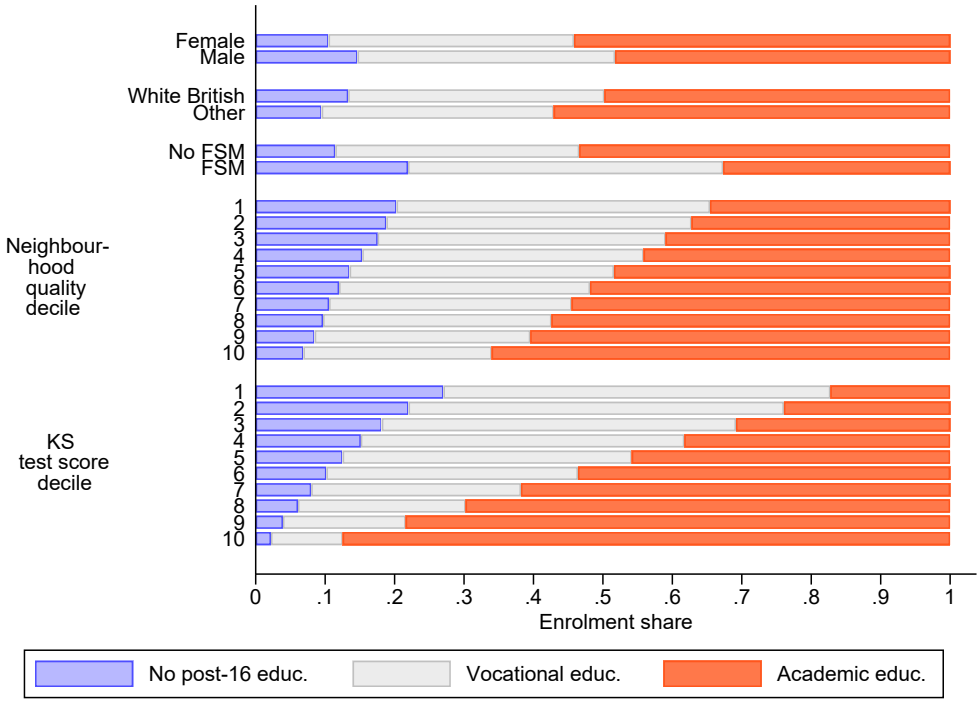
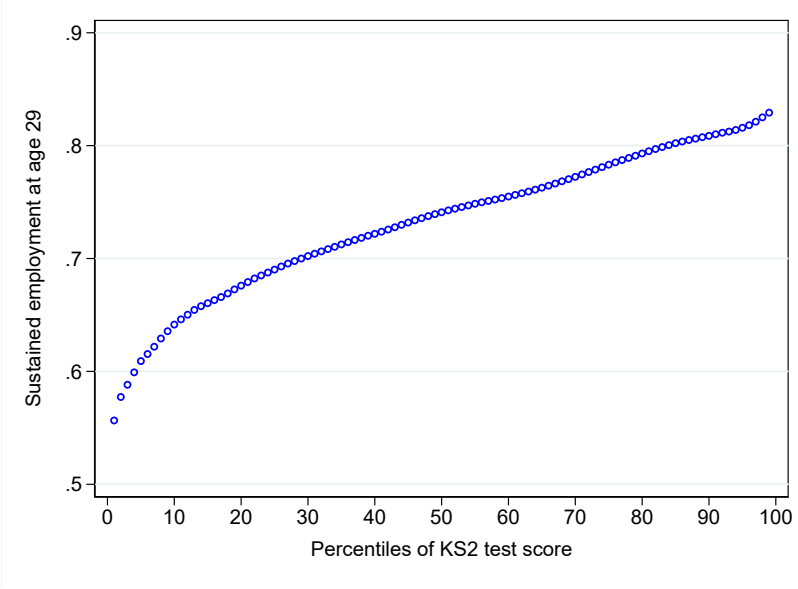


Figure A4.2. Track choices by observable characteristics.

Notes: This figure shows the distribution of students over treatments by observable characteristics. The figure is based on students from schools with sixth form (i.e., those excluded from the estimation sample). FSM stands for free school meal eligibility. Neighbourhood quality deciles are deciles of the first principal component (PC) of all seven (inverted) IoDs. KS2 test score deciles are deciles of the first PC of all three end-of-primary-school (KS2) test scores. PCs are extracted (and their deciles calculated) in the full sample, so that the deciles refer to the same categories in the estimation sample as for students from schools with sixth form (see Figure 4.2.)

Panel A: *Sustained employment*



Panel B: *Annual earnings (incl. 0s)*

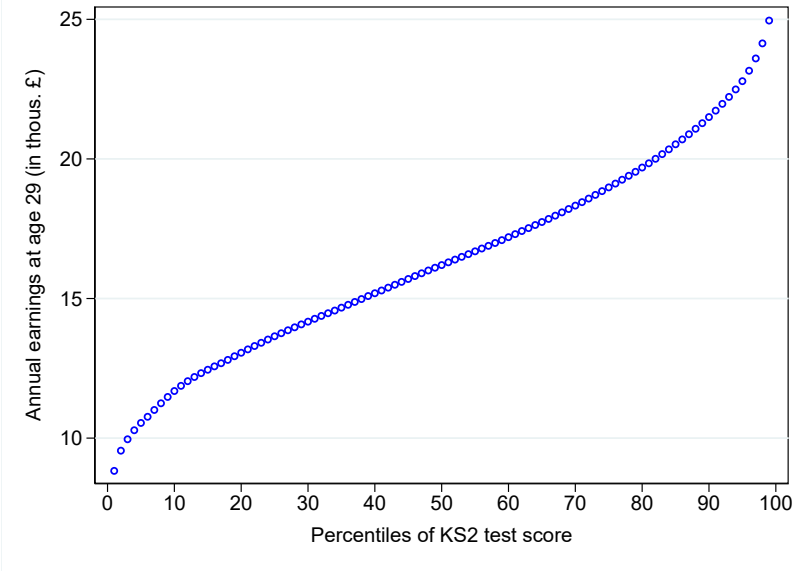
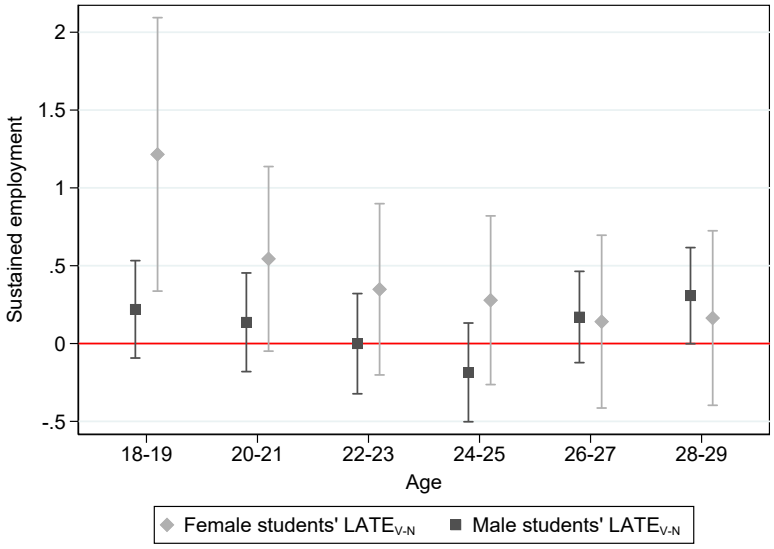


Figure A4.3. Labour market outcomes by KS2 test score percentile.

Notes: This figure shows fitted values for average employment and annual earnings at age 29 (in 2010 pounds) across percentiles of the first principal component of the KS2 English, Maths and Science scores, estimated via local linear regressions.

Panel A: *Sustained employment*



Panel B: *Annual earnings (incl. 0s)*

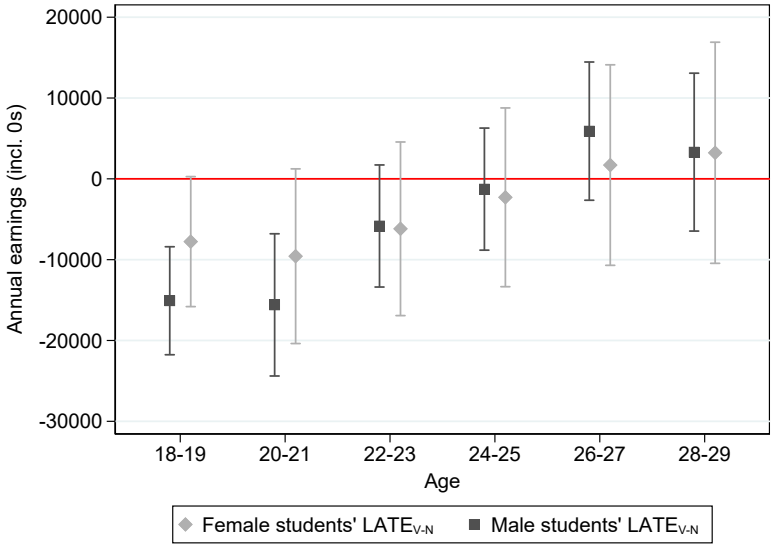


Figure A4.4. Age-effect profiles at the vocational-no post-16 education margin.

Notes: This set of figures plot the estimates of LATEs of vocational *vs.* no post-16 education on sustained employment (Panel A) and annual earnings (Panel B) and the associated 95 % CIs at different age points by gender. For increased precision and computational ease we combine outcomes from two successive age points; we do so by taking the average of annual earnings (if observed) between two successive years and the average of whether students were in sustained employment. For comparability across the whole age range only earnings (and employment) from employed, but not from self-employed, work are included. LATE$_{V\text{-}N}$ estimates are corrected for estimated differences in predicted outcomes across compliers groups as illustrated in 4.4.2. Confidence intervals are based on block bootstrapped standard errors at the LSOA×cohort level using 500 iterations.

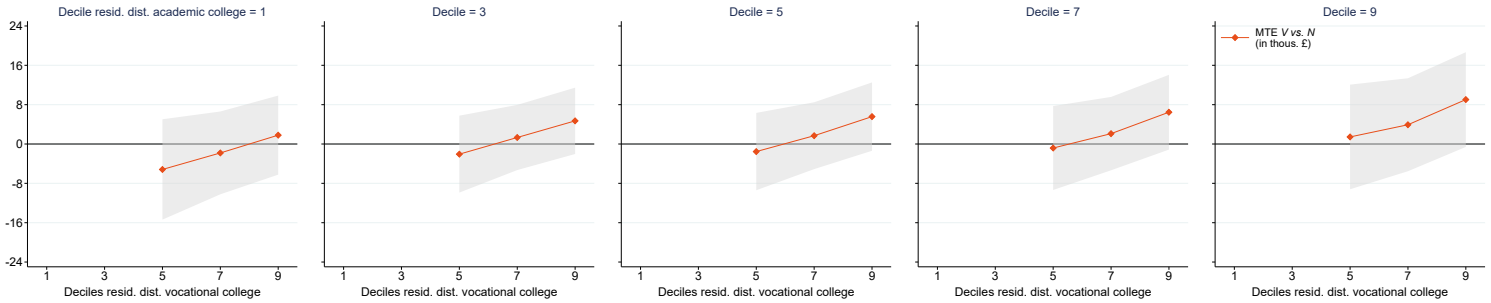


Figure A4.5. MTEs of vocational *vs.* no post-16 education on earnings across the distance grid for males.

Notes: This figure plots, only for males, the $\text{MTE}_{V\text{-}N}$ estimated at different locations of the two-dimensional grid defined by the two residualised distance instruments. We first residualise the two distance instruments Z_V and Z_A with respect to the control set. Then, we define a 5-by-5 distance grid corresponding to the 1st, 3rd, 5th, 7th and 9th deciles of the two residualised distances. Finally, we estimate a series of local linear reduced form and first stage regressions for each gridpoint, weighting observations by their distance to the grid point using a two-dimensional Epanechnikov kernel function with a bandwidth of two standard deviations in either dimension. The MTE estimates are constructed from the local regression coefficients analogously to the main estimation. 90% confidence intervals are based on 500 block-bootstrap iterations clustering at the LSOA×cohort level. To improve the visualisation of the estimates at the gridpoints when they are more relevant, we choose to consistently omit estimates at the 1st and 3rd deciles of residualised Z_V: at close distance to vocational college, students at the *V-N* margin are not responsive to changes in distance as shown by a too weak firt stage resulting in inflated estimates and CIs. The implication is that estimates of $\text{MTE}_{V\text{-}N}$ for students leaving close to vocational colleges are not as relevant since the proportion of compliers at the *V-N* margin is particularly low as shown by A4.6. Results for females are not reported for similar reasons: in line with the global specification, first stages are too weak across all gridpoints so that ensuing estimates cannot be meaningfully interpreted or visualised. They remain available on request.

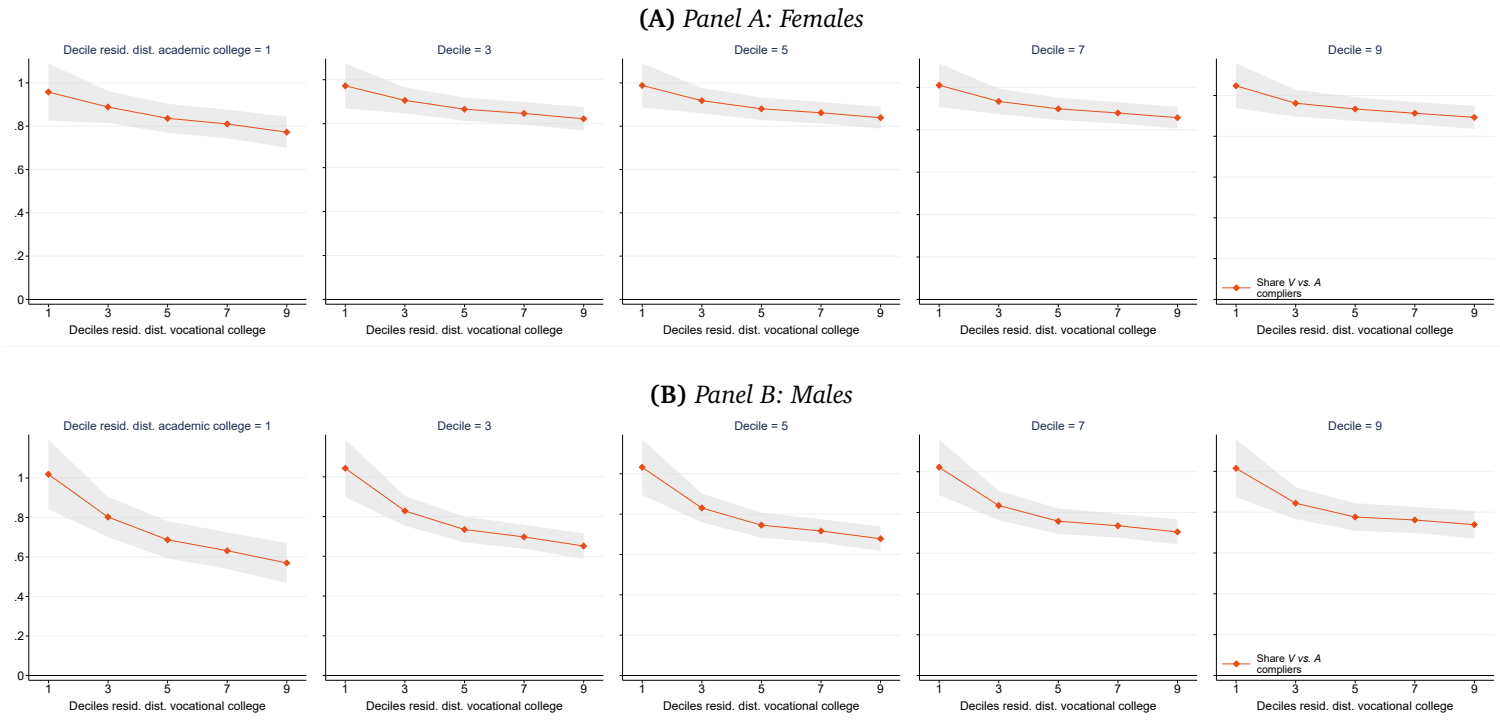


Figure A4.6. Compliers shares across the distance grid.

Notes: This figure plots, separately by gender, the share of $V - A$ compliers estimated at different locations of the two-dimensional grid defined by the two residualised distance instruments. We first residualise the two distance instruments Z_V and Z_A with respect to the control set. Then, we define a 5-by-5 distance grid corresponding to the 1st, 3rd, 5th, 7th and 9th deciles of the two residualised distances. Finally, we estimate a series of local linear first stage regressions at each gridpoint, weighting observations by their distance to the grid point using a two-dimensional Epanechnikov kernel function with a bandwidth of two standard deviations in either dimension. The share estimates are constructed from the local regression coefficients analogously to the main estimation. 90% confidence intervals are based on 500 block-bootstrap iterations clustering at the LSOA×cohort level.

Table A4.1. First stages in different subsamples: testing monotonicity.

First stage(s) for:	Net LATE	Margin-specific LATEs		
	D_V wrt Z_V (1)	D_V wrt Z_A (2)	D_A wrt Z_V (3)	D_N wrt Z_V (4)
Female	-0.0393***	0.0822***	0.0353***	0.0041***
	(0.0015)	(0.0013)	(0.0014)	(0.0010)
	303,608	303,608	303,608	303,608
Male	-0.0341***	0.0700***	0.0274***	0.0067***
	(0.0015)	(0.0013)	(0.0013)	(0.0011)
	315,217	315,217	315,217	315,217
White British	-0.0406***	0.0753***	0.0338***	0.0068***
	(0.0013)	(0.0011)	(0.0011)	(0.0008)
	498,228	498,228	498,228	498,228
Other ethnicity	-0.0142***	0.0823***	0.0133***	0.0009
	(0.0028)	(0.0023)	(0.0026)	(0.0017)
	120,597	120,597	120,597	120,597
Free school meal (FSM)	-0.0217***	0.0526***	0.0152***	0.0065***
	(0.0025)	(0.0021)	(0.0019)	(0.0021)
	107,607	107,607	107,607	107,607
No FSM	-0.0395***	0.0809***	0.0344***	0.0051***
	(0.0013)	(0.0011)	(0.0012)	(0.0008)
	511,218	511,218	511,218	511,218
Bottom 25% KS2	-0.0244***	0.0373***	0.0177***	0.0068***
	(0.0019)	(0.0015)	(0.0012)	(0.0016)
	183,738	183,738	183,738	183,738
Second 25% KS2	-0.0326***	0.0584***	0.0236***	0.0090***
	(0.0020)	(0.0017)	(0.0016)	(0.0015)
	145,027	145,027	145,027	145,027
Third 25% KS2	-0.0422***	0.0946***	0.0378***	0.0044***
	(0.0021)	(0.0018)	(0.0019)	(0.0013)
	145,033	145,033	145,033	145,033
Top 25% KS2	-0.0520***	0.1232***	0.0504***	0.0016*
	(0.0023)	(0.0019)	(0.0023)	(0.0010)
	145,027	145,027	145,027	145,027
Bottom 25% IoD	-0.0147***	0.0577***	0.0133***	0.0015
	(0.0025)	(0.0021)	(0.0020)	(0.0019)
	154,739	154,739	154,739	154,739
Second 25% IoD	-0.0187***	0.0682***	0.0170***	0.0018
	(0.0024)	(0.0020)	(0.0020)	(0.0016)
	154,680	154,680	154,680	154,680
Third 25% IoD	-0.0446***	0.0825***	0.0346***	0.0101***
	(0.0023)	(0.0019)	(0.0021)	(0.0014)
	154,700	154,700	154,700	154,700
Top 25% IoD	-0.0624***	0.0981***	0.0555***	0.0069***
	(0.0025)	(0.0023)	(0.0024)	(0.0012)
	154,706	154,706	154,706	154,706

Notes: The table reports the relevant first stage coefficients as estimated in different covariates sub-sample. Standard errors are in parentheses. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4.2. Comparison and decomposition of associated Multivariate IV estimates.

	Female students	Male students
Multivariate 2SLS estimates		
β^{2SLS}_{V-A}	-623	-1116
	(397)	(527)
β^{2SLS}_{V-N}	5578	-420
	(5116)	(4188)
Margin-specific LATE estimates		
$LATE_{V-A}$	-463	-2347
$LATE_{V-N}$	4201	4629
$LATE_{A-N}$	8648	-11673
First-stages derived weights		
θ_A	0.960	0.934
θ_N	0.654	0.729

Notes: The upper panel of the table reports, separately by gender, the estimated coefficients of a 2SLS regression where choice indicators for being in the academic track and in no Post-16 education are instrumented with both distance instruments simultaneously. These estimates cannot be normally interpreted as returns to education choices as they represent a mixture of effects from multiple margins of treatment. It can be shown that: $-\beta^{2SLS}_{V\text{-}A} = \theta_A$ LATE$_{V\text{-}A}$ + $(1 - \theta_A)$(LATE$_{V\text{-}N}$ − LATE$_{A\text{-}N}$) and that $-\beta^{2SLS}_{V\text{-}N} = \theta_N$ LATE$_{V\text{-}N}$ + $(1 - \theta_N)$(LATE$_{V\text{-}A}$ + LATE$_{A\text{-}N}$) where LATE$_{A\text{-}N}$ is the effect at the margin of academic *vs.* no post-16 education and θ_A and θ_N depend on the multivariate 2SLS first-stage equations (**?**). The bottom panel of the table reports the relevant elements of this decomposition, explaining why empirically the 2SLS estimates would appear to be considerably different from our estimated effects.

Table A4.3. Correcting the IV estimates at the no post-16 education margin for complier differences.

Dependent variable:	Sustained employment		Annual earnings (incl. 0s)		Annual earnings (excl. 0s)	
	Raw IV (1)	Corrected (2)	Raw IV (3)	Corrected (4)	Raw IV (5)	Corrected (6)
A. Female students						
Vocational PO	0.721	0.794	14,635	17,372	22,928	24,923
	(0.257)	(0.258)	(6,270)	(6,270)	(8,213)	(8,563)
Academic PO	0.676		10,434		16,053	
	(0.152)		(2,740)		(3,451)	
$LATE_{V\text{-}N}$	0.045	0.118	4,201	6,938	6,875	8,871
	(0.296)	(0.290)	(6,737)	(6,498)	(8,364)	(8,351)
	$p_{(\Delta=0)} = 0.08$		$p_{(\Delta=0)} = 0.13$		$p_{(\Delta=0)} = 0.34$	
B. Male students						
Vocational PO	0.865	0.842	24,647	23,283	27,004	25,574
	(0.130)	(0.129)	(4,622)	(4,508)	(4,507)	(4,421)
Academic PO	0.779		20,017		23,726	
	(0.092)		(2,629)		(2,588)	
$LATE_{V\text{-}N}$	0.086	0.062	4,629	3,266	3,278	1,848
	(0.155)	(0.154)	(5,098)	(5,022)	(4,916)	(4,868)
	$p_{(\Delta=0)} = 0.09$		$p_{(\Delta=0)} = 0.09$		$p_{(\Delta=0)} = 0.08$	

Notes: For each of the three outcomes, the table shows, separately by gender, the original and corrected potential outcomes (PO) of vocational and no post-16 education for compliers at the margin between vocational and no post-16 education as well as the resulting original and corrected $LATE_{V\text{-}N}$. The vocational PO is corrected by taking into account the estimated difference in the associated predicted outcome between the two groups of compliers induced by conditional variation in the two separate instruments. The corrected $LATE_{V\text{-}N}$ is then obtained by subtracting the original no post-16 PO from the corrected vocational PO. The table also reports the p-values for the test of the null hypothesis that the corrected $LATE_{V\text{-}N}$ is equal to the original one. The number of observations is identical to Table 4.6. Standard errors are block bootstrapped at the LSOA×cohort level using 999 iterations. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4.4. Robustness checks for the margin-specific labour market effects estimates.

	Instruments ≡ driving distance			No sample restrictions			Excluding unmatched students		
	Sustained employment	Annual earnings (incl. 0s)	Annual earnings (excl. 0s)	Sustained employment	Annual earnings (incl. 0s)	Annual earnings (excl. 0s)	Sustained employment	Annual earnings (incl. 0s)	Annual earnings (excl. 0s)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
A. Female students									
$\text{Share}_{V\text{-}A}$	0.87	0.87	0.90	0.94	0.94	0.95	0.93	0.93	0.94
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
$\text{LATE}_{V\text{-}A}$	-0.002	-513	-1101*	-0.008	-587	-1370**	-0.025	-1114*	-1361**
	(0.021)	(672)	(673)	(0.019)	(590)	(630)	(0.019)	(635)	(670)
$\text{LATE}_{V\text{-}N}$	0.156	6302	6860	0.062	8304	17889	0.205	10700	14010
	(0.237)	(5135)	(7348)	(0.747)	(28511)	(138997)	(0.447)	(11373)	(26700)
B. Male students									
$\text{Share}_{V\text{-}A}$	0.79	0.79	0.79	0.81	0.81	0.83	0.81	0.81	0.83
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)
$\text{LATE}_{V\text{-}A}$	0.018	-1617*	-2082**	0.001	-2221***	-2875***	0.008	-2262***	-2798***
	(0.023)	(894)	(866)	(0.022)	(860)	(834)	(0.021)	(889)	(803)
$\text{LATE}_{V\text{-}N}$	-0.079	-1474	-2237	0.173	6363	3732	0.128	5027	3435
	(0.145)	(4767)	(4456)	(0.183)	(5679)	(5358)	(0.159)	(5256)	(5315)

Notes: This table reports, separately by gender and for all the three outcomes of interest, three different robustness checks. In columns (1–3) we redefine the instruments to measure the (log) driving (instead of geographical) distance between students' residential LSOA centroid and the closest academic/vocational college; in columns (4–6) we drop all sample restrictions, thus retaining groups of students unlikely to participate in the labour market in the estimation sample (see 4.2.2); in columns (7–9) we exclude all students who cannot be matched to a single entry in the tax records throughout the 13 year period we observe. All regressions include the same set of controls as in the main estimation. Standard errors are block bootstrapped at the LSOA×cohort level using 500 iterations. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Chapter 5

Conclusion

This dissertation studies the consequences of institutional differentiation in the secondary school systems. The empirical analyses focus on Germany and England, exemplifying two countries which take radically different approaches in this regard. Across chapters, the focus shifts from *vertical* differentiation, i.e., ability tracking between different school types, to *horizontal* differentiation, i.e., curricular tracking between vocational and academic education. Guided by the different motivations behind these two modes of tracking, the chapters examine effects on student achievement during school, educational attainment after school and employment and earnings in the labour market. The findings demonstrate that differentiation profoundly impacts on these student outcomes. Accordingly, school policy has important implications for the welfare of individuals and economies as a whole. In this concluding chapter, I discuss, in turn, limitations of the analyses and ensuing directions for future research, as well as implications for school policy that derive from my findings.

5.1 Limitations and Scope for Future Research

To estimate the effect early between-school ability tracking on student achievement ***Chapter 2*** exploits differences in tracking between German federal states: in all states, about 40% of students transition to the academic track after comprehensive primary school. Depending on the state, the remaining student body is either divided between low- and intermediate-track schools or taught comprehensively for another two years. I estimate the effects of these two years of comprehensive instead of tracked schooling in a triple-differences framework: the estimator compares achievement growth of comprehensively taught non-academic-track students with that of tracked ones, while controlling for state-specific trends using unaffected students in the academic track. This reveals an positive average effect of prolonged comprehensive schooling on mathematics and reading achievement. The average effects are almost entirely driven by large effects for low achievers, while effects for high achievers are null.

The main limitation of this research is that the treatment effect identified in this paper pertains to a population of students that excludes the highest-achieving students in academic-track schools. Hence, one cannot directly extrapolate from these results to the effects of fully comprehensive school systems. However, the fact that I find null effects for the highest-achieving

non-academic-track students who, given considerable overlap in pre-tracking achievement between tracks, would be medium-high achievers also in the academic track, suggests that the null effects are likely to extend to most academic-track students. Still, future research should empirically test these conjectures directly and estimate (heterogeneous) effects of teaching all students comprehensively for another two years. It should be noted that three states, by virtue of primary school covering the first six grades (Berlin and Brandenburg) or a mandatory comprehensive 'orientation stage' (Mecklenburg-orpommern), do indeed keep schooling fully comprehensive during these grades.[1] However, identification would be considerably more challenging because general differences between primary and secondary schools, as well as state specificities (e.g., all of them are in East Germany) would hamper the comparison. Moreover, these states are small in terms of student numbers, so that sample size would be an issue in most conventional data sets. Of course, evidence may come from other settings than Germany.

Other limitations relate to the data. First, the sample size in the NEPS data is modest, thus preventing estimation of precise effects sizes. While this is tackled by reverting to the IQB data, these come with their own limitations, in particular missing achievement measures in seventh grade, right after the window with clear-cut treatment differences between states, and no panel structure, limiting their potential for inspecting effect heterogeneity. Second, it would be interesting and highly relevant to inspect long-term outcomes. I am prevented from inspecting effects beyond ninth grade because this the latest point of testing in the IQB data and, while the NEPS continued to follow students beyond this point, sample sizes become very small.

To study the effect of opening up alternative pathways towards university eligibility in the context of Germany's tracked school system, ***Chapter 3*** exploits between-county differences in the expansion of comprehensive schools and vocational high schools. In particular, the analysis regresses county×cohort-level attainment rates on county and cohort fixed effects and measures of local school supply for the two mentioned school types and traditional academic-track schools. School supply is measured either continuously, by the number of schools per student, or binary, indicating the availability of at least a single school of the respective type. This reveals robust positive effects on university-entrance certificate attainment rates for vocational high school supply, but not for comprehensive and academic-track school supply.

While the use of aggregate administrative data allows us to observe the population of counties, schools and graduates, thereby enabling this analysis which would not be possible with survey data, it comes with some serious limitations. First, our measures of school supply are crude because they merely count the number of schools and not the number of slots. As explained above, this is preferable to using more granular enrolment counts if the aim is to identify supply-side effects. However, it means that our treatment variables are measured with substantial error, attenuating the effect estimates towards zero. This strengthens our conclusions regarding the effect of vocational high schools, given that it is, nonetheless, sizeable and significant across model specifications. However, it means that the absence of significant effects for the other two

[1] For this reason these states are excluded from the analysis in chapter 2.

school types must be taken with caution. For comprehensive school supply, over-time variation in our measure is quite limited. For these schools, but likely even more so for academic-track schools, which are available in every county throughout our study period, expansion might have happened not through the creation of new schools but through an expansion of slots in existing schools, which we cannot capture.

Second, the aggregate data does not allow us to observe (or follow) individual students. This means that we are forced to restrict attention to attainment of the university-entrance certificate. Clearly, this is an important outcome, but a broader focus on achievement, educational progression and labour market performance would be crucial to evaluate these schools more comprehensively. We hope future research will augment the scant evidence base on this topic, currently consisting of only a few descriptive studies on achievement (e.g., Köller *et al.*, 2004; Vieluf *et al.*, 2014) and one study on educational progression and early career labour market outcomes (Zimmermann, 2019). Another implication of not being able to observe individual students is that our analysis remains mute regarding the type of students driving the attainment effects we find. This would be crucial to assess the impact of vocational high schools on educational inequality, about which we can only speculate.

Chapter 4 estimates returns to vocational education in England for students at the margin between vocational and academic education and, separately, for students at the margin between vocational and no post-16 education by leveraging two alternative-specific instrumental variables (IVs) in a novel identification framework. The IVs respectively measure students' proximity to the nearest vocational provider and the nearest academic provider and are plausibly exogenous conditional on detailed student-, school- and neighbourhood-level characteristics. The estimations reveal that the vast majority of marginal vocational students are at the margin with academic education (instead of no further education) and, for them, vocational education leads to substantial losses in earnings at age 29, especially among males.

While the administrative education datasets used in this research (NPD, ILR and HESA) are exceptionally wide in scope and fine in detail, the labour market dataset they are linked to (HMRC) is more crude. The latter is collected for tax purposes and merely contains information on employment spells and earnings but no information on employers or occupation, let alone tasks. Accordingly, we cannot inspect if vocational students enter systematically different occupations or firms, and if this explains part of the negative earnings effects we find. Hence, investigating labour market mechanisms for the effects of vocational education is an important avenue for future research.

Additionally, there are inherent limits to IV estimation. Any IV design is only as strong as its first stage—not only because it determines the estimator's precision, but also because it determines how broadly the causal effects IV estimates (i.e., LATE) apply. In our case, the share of compliers in the population is non-negligible at 12%. Still, for the remaining 88%, strictly speaking, we cannot estimate causal returns. To learn more about how externally valid our effects are, we estimate marginal treatment effects and inspect the pattern of effect heterogeneity with respect to students' (unobserved) preferences. This reveals that, while returns to vocational

education (*vs.* academic education) are negative for most marginal students, for males with particularly high preferences they are positive. Students seem to have some understanding of their comparative advantage and sort accordingly. Therefore, we can credibly speculate about how our effects might extrapolate. In particular, we conjecture that for many non-marginal students returns to vocational education might be positive. Ideally, future research would test these conjectures more directly.

Similarly, we have to revert to informed speculation regarding the effects for students from secondary schools with upper-secondary provision ('sixth form'), because for those the second instrument, distance to academic college, is not relevant. Our informal bounding exercise suggests that returns to vocational education are even more negative for these students. Clearly, direct evidence on these students would be preferable though.

With respect to precision, our goal of margin-specific identification makes estimation particularly demanding because it means that we rely on multiple first stages per margin, especially at the *V-N* margin between vocational and no post-16 education where identification is particularly involved. Indeed, we cannot retrieve precise estimates at this margin, also because a weak first stage for no post-16 education compounds the problem. However, note that the latter is a finding in itself: the strong overall first stage for vocational education, together with the weak first stage for no post-16 education implies that very few of the students, whose education choice is responsive to incentives like distance, are at this margin of indifference. Accordingly, effect at this margin are less important from a policy perspective: they only apply to 20% of 11%, i.e., 2.1% of all male and 10% of 13%, i.e., 1.3% of all female students (from schools without sixth form). Nevertheless, precise estimation of these effects would be interesting from a theoretical perspective because a central argument for vocational education concerns this margin.

5.2 Policy Implications

Despite these limitations, the findings presented in this dissertation carry a number of lessons for school policy.

The results in chapter 2 offer important insights for the long-standing debate surrounding ability tracking. Regardless of the fact that the institutional differences between states I exploit for identification only concern non-academic-track students, the findings clearly disprove the rationale that classroom homogeneity is always good for teaching. Homogeneous, but strongly negatively selected, school environments impair the skill acquisition of students. Of course, this does not imply that homogeneity is irrelevant. Instead, it points to a trade-off: there are limits to efficiency gains from classroom homogeneity because other mechanisms, such as peer effects, motivation and aspirations, start to depress achievement at the bottom once the system becomes too vertically differentiated. Accordingly, policy-makers need to carefully balance these forces when determining the degree of ability tracking in their school systems and the age at which it starts.

Here it is important to stress that deferring or abandoning ability grouping altogether is not the only option: the evidence for streaming students *within* schools appears much more promising than that for tracking *between* schools. Next to counteracting the emergence of strongly negatively selected school environments, within-school streaming offers much greater flexibility: students can move more easily between levels and, by grouping students *per subject*, teaching should become more efficient in accordance with the original rationale of ability grouping. An additional benefit of grouping by ability within schools might be to contain between-school sorting through parental school choice dynamics. Comprehensive school systems have been shown to make native and higher socio-economic status (SES) parents, whose children tend to be higher achieving, select specific (e.g., private) schools to separate their children from low-SES and migrant children, who tend to be lower achieving (e.g., Kruse, 2019; Kosunen *et al.*, 2020). Within-school streaming might plausibly curb incentives to do so.

Concretely for Germany, my findings support the recent reform push towards a two-tiered school structure. Several states abolished separate low- and intermediate-track schools, establishing comprehensive schools as the sole alternative besides academic-track schools. While, given the above, a general deferring of sorting between schools, paired with differentiation within schools, might be even more efficient, my results confirm that these reforms lead to improvements in terms of equity and efficiency, in line with policymakers' intentions. Note that my research design estimates treatment effects at the state level, which is where sovereignty about education policy lies in Germany. Accordingly, these results have immediate relevance for policymakers, serving as reform evaluations for states that implemented such changes and as an evidence base for those that did not (Bavaria, Baden-Wurttemberg, Lower Saxony, North Rhine Westphalia and Hesse).

Further, this dissertation offer some insights for German policymakers aiming to increase university enrolment. With the caveats of limited treatment variation and attenuation bias in mind, the results in chapter 3 at least show that comprehensive schools do not have *large* effects on the attainment of university eligibility. While comprehensive schools do bring benefits by increasing achievement at the bottom if they come to replace low- and intermediate-track schools in the two-tiered system, these schools do not prove particularly effective in increasing attainment at the top. This might not be surprising given that their typical entry point is right after primary school, where academic-track schools cream skim most higher-achieving students. In contrast, the purely upper-secondary vocational high schools seem to do a good job at attracting students who otherwise would not have pursued academic-upper secondary education and to, thereby, increase attainment. One reason for their success might be that providing a slightly less demanding, and for many students certainly less daunting, upgrading option after lower-secondary school than traditional academic-track schools, fits much better into the institutional logic of the tripartite system than the comprehensive school model.

Another reason for the effectiveness of vocational high schools is to be found in their hybrid nature, combining an academic upper-secondary curriculum with specialised vocational courses. The idea of combining academic and vocational curricula in hybrid institutions has garnered

policy interest recently (see Machin *et al.*, 2020) and my findings are consistent with other evidence that this can be a promising model (e.g., Kreisman and Stange, 2020; Machin *et al.*, 2020; Bertrand *et al.*, 2021). The results in chapter 4 show that oft-voiced worries about 'diversion' from academic to vocational education are warranted: even in England, where vocational education is not particularly reputable, a substantial share of students is to be found at this margin and diversion leads to large penalties in terms of educational progression to university and later earnings. Therefore, hybrid schools can be an attractive alternative for these students, alleviating some of the concerns that come with curricular tracking between vocational and academic education, especially if they keep the university route open, as vocational high schools in Germany do.

More generally, regarding the long-standing debate on the merits of vocational education, this dissertation makes the important conceptual point that effect heterogeneity is key. In contrast to ability tracking where the goal is to teach a given content more efficiently, the rationale of curricular tracking is to diversify the contents taught in order to generate better matches between the talents and interests of students and the knowledge and skills they learn for the labour market. The very aim of curricular tracking is that students select into different curricula based on comparative advantage. Accordingly, average comparisons of academic *vs.* vocational students are of limited informational value. Policy-relevant parameters are either system-level effects (e.g., 'How would the population distribution of skills/earnings change if we had more or less vocational content?') or alternative-specific effects for marginal students. Those can give guidance about where improvements or re-allocations are necessary for improved efficiency.

A very recent literature that takes these issues seriously, has reached encouraging conclusions regarding the merits of vocational education. These studies show that despite unfavourable average age-earnings profiles, marginal students benefit from vocational education: on net in Norway (Bertrand *et al.*, 2021), at the margin to no further education, with zero effects at the margin to academic education, in Denmark (Birkelund and van de Werfhorst, 2022) and even at the margin to academic education in Finland (Silliman and Virtanen, 2022). The evidence in chapter 4 curbs this enthusiasm for England, where most marginal students are at the margin with academic education and clearly lose out from choosing the vocational track instead. This highlights the importance of exercising caution when extrapolating findings across contexts with different educational and labour market institutions. Nevertheless, also here, we find evidence for positive effects of vocational education for students with high preferences for this track, confirming that there is merit to the general principle of labour force diversification through offering different curricula in upper secondary education.

In terms of concrete implications for education policy in England, our results suggest that the current allocation of students over vocational and academic tracks is inefficient and policy should aim to sway more eligible students to choose the academic route. Alternatively (and additionally), the vocational track should be improved through establishing clearer pathways towards higher education, as is the case in the Nordic countries. Incidentally, this mirrors one of the central pieces of advice for vocational education reform stressed in the famous Wolf

(2011) report. Finally, though imprecise and insignificant, our estimates for the second margin are tentative evidence for positive returns to vocational education against the alternative of no further education. This suggests that the decision to extend compulsory education to age 18, which the English government has taken since the cohorts that we study have graduated, might have generated substantial societal returns.

Altogether, this dissertation demonstrates that school policy, in general, and rules regarding ability and curricular tracking, in particular, profoundly impact on the the welfare of individuals and economies as a whole. Policymakers ought to use this potential to reform our school systems towards higher efficiency and equity, so that all students are well-prepared for the challenges of twenty-first century labour markets.

Bibliography

Aakvik, A., Salvanes, K.G. and Vaage, K. (2010). 'Measuring heterogeneity in the returns to education using an education reform', *European Economic Review*, vol. 54(4), pp. 483–500.

Abadie, A. (2002). 'Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models', *Journal of the American Statistical Association*, vol. 97(457), pp. 284–292.

Abadie, A., Athey, S., Imbens, G.W. and Wooldridge, J. (2017). 'When should you adjust standard errors for clustering?', NBER Working Paper No. 24003.

Abdulkadiroğlu, A., Angrist, J. and Pathak, P. (2014). 'The elite illusion: Achievement effects at Boston and New York exam schools', *Econometrica*, vol. 82(1), pp. 137–196.

Acemoglu, D. and Autor, D. (2011). 'Skills, Tasks and Technologies: Implications for Employment and Earnings', in (D. Card and O. Ashenfelter, eds.), *Handbook of Labor Economics*, pp. 1043–1171, vol. 4, Elsevier.

Acemoglu, D. and Pischke, J.S. (1998). 'Why Do Firms Train? Theory and Evidence', *The Quarterly Journal of Economics*, vol. 113(1), pp. 79–119.

Acemoglu, D. and Restrepo, P. (2018). 'The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment', *American Economic Review*, vol. 108(6), pp. 1488–1542.

Agan, A.Y., Doleac, J.L. and Harvey, A. (2021). 'Misdemeanor Prosecution', National Bureau of Economic Research.

Ainsworth, J.W. (2002). 'Why does it take a village? The mediation of neighborhood effects on educational achievement', *Social Forces*, vol. 81(1), pp. 117–152.

Ammermüller, A. (2013). 'Institutional features of schooling systems and educational inequality: Cross-country evidence from PIRLS and PISA', *German Economic Review*, vol. 14(2), pp. 190–213.

Ammermüller, A. and Pischke, J.S. (2009). 'Peer effects in european primary schools: Evidence from the Progress in International Reading Literacy Study', *Journal of Labor Economics*, vol. 27(3), pp. 315–348.

Angrist, J.D. and Krueger, A.B. (1991). 'Does Compulsory School Attendance Affect Schooling and Earnings?', *The Quarterly Journal of Economics*, vol. 106(4), pp. 979–1014.

Angrist, J.D. and Lavy, V. (1999). 'Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement*', *The Quarterly Journal of Economics*, vol. 114(2), pp. 533–575.

Angrist, J.D. and Pischke, J.S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press.

Angrist, J.D. and Pischke, J.S. (2010). 'The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics', *Journal of Economic Perspectives*, vol. 24(2), pp. 3–30.

Arum, R., Gamoran, A. and Shavit, Y. (2007). 'More Inclusion Than Diversion: Expansion, Differentiation, and Market Structure in Higher Education', in (Y. Shavit, R. Arum and A. Gamoran, eds.), *Stratification in Higher Education: A Comparative Study*, pp. 1–35, Stanford: Stanford University Press.

Autor, D.H. (2019). 'Work of the Past, Work of the Future', *AEA Papers and Proceedings*, vol. 109, pp. 1–32.

Autorengruppe Bildungsberichterstattung (2020). *Bildung in Deutschland 2020*, Bielefeld: Bertelsmann Verlag.

Baethge, M. and Wolter, A. (2015). 'The German Skill Formation Model in Transition', *Journal for Labour Market Research*, vol. 48, pp. 97–112.

Bald, H. (2011). 'Realschule–Erweiterte Realschule–Mittelschule Usw.–Eine Problemanzeige', *Theo-Web. Zeitschrift für Religionspädagogik*, vol. 10, pp. 80–102.

Becker, G. (1964). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*, New York: National Bureau of Economic Research.

Becker, R. and Blossfeld, H.P. (2022). 'Changes in the returns to education at entry into the labour market in West Germany', *Longitudinal and Life Course Studies*, vol. 13(1), pp. 61–86.

Bellenberg, G. (2005). 'Wege durch die Schule - Zum Zusammenhang zwischen institutionalisierten Bildungswegen und individuellen Bildungsverläufen im deutschen Schulsystem', *Bildungsforschung*, vol. 2(2).

Bellenberg, G. (2012). *Schulformwechsel in Deutschland. Durchlässigkeit und Selektion in den 16 Schulsystemen der Bundesländer innerhalb der Sekundarstufe I*, Gütersloh: Bertelsmann-Stiftung.

Berg, I. (1971). *Education and Jobs: The Great Training Robbery*, Boston: Beacon Press.

Bertrand, M., Duflo, E. and Mullainathan, S. (2004). 'How much should we trust differences-in-differences estimates?', *The Quarterly Journal of Economics*, vol. 119(1), pp. 249–275.

Bertrand, M., Mogstad, M. and Mountjoy, J. (2021). 'Improving Educational Pathways to Social Mobility: Evidence from Norway's Reform 94', *Journal of Labor Economics*, vol. 39(4), pp. 965–1010.

Betts, J.R. (2011). 'The economics of tracking in education', pp. 341–381, vol. 3 of *Handbook of the Economics of Education*, Amsterdam: Elsevier.

Bhuller, M., Dahl, G.B., Løken, K.V. and Mogstad, M. (2020). 'Incarceration, Recidivism, and Employment', *Journal of Political Economy*, vol. 128(4), pp. 1269–1324.

Bietenbeck, J. (2020). 'The Long-Term Impacts of Low-Achieving Childhood Peers: Evidence from Project STAR', *Journal of the European Economic Association*, vol. 18(1), pp. 392–426.

Birkelund, J.F. and van de Werfhorst, H.G. (2022). 'Long-term labor market returns to upper secondary school track choice: Leveraging idiosyncratic variation in peers' choices', *Social Science Research*, vol. 102, p. 102629.

Bishop, J.H. (1997). 'The Effect of National Standards and Curriculum-Based Exams on Achievement', *The American Economic Review*, vol. 87(2), pp. 260–264.

Black, S.E., Devereux, P.J. and Salvanes, K.G. (2005). 'Why the Apple Doesn't Fall Far: Understanding Intergenerational Transmission of Human Capital', *American Economic Review*, vol. 95(1), pp. 437–449.

Blossfeld, H.P., Rossbach, H.G. and von Maurice, J. (2011). 'Education as a lifelong process: The German National Educational Panel Study (NEPS)', *Zeitschrift für Erziehungswissenschaft*, vol. 14 (special issue).

Blundell, R., Joyce, R., Norris Keiller, A. and Ziliak, J.P. (2018). 'Income inequality and the labour market in Britain and the US', *Journal of Public Economics*, vol. 162, pp. 48–62.

Boshoff, J., Moore, J. and Speckesser, S. (2019). 'Inequality in education and labour market participation of young people across English localities: An exploration based on Longitudinal Education Outcomes (LEO) data', CVER Briefing Notes 010.

Brännström, L. (2008). 'Making Their Mark: The Effects of Neighbourhood and Upper Secondary School on Educational Achievement', *European Sociological Review*, vol. 24(4), pp. 463–478.

Brauckmann, S. and Neumann, M. (2004). 'Berufliche Gymnasien in Baden-Württemberg: Geschichte und heutige Ausgestaltung', in (O. Köller, R. Watermann, U. Trautwein and O. Lüdtke, eds.), *Wege Zur Hochschulreife in Baden-Württemberg. TOSCA – Eine Untersuchung an Allgemein Bildenden Und Beruflichen Gymnasien*, pp. 69–111, Opladen: Leske+Budrich.

Breen, R. and Jonsson, J.O. (2005). 'Inequality of Opportunity in Comparative Perspective: Recent Research on Educational Attainment and Social Mobility', *Annual Review of Sociology*, vol. 31(1), pp. 223–243.

Breen, R., Luijkx, R., Müller, W. and Pollak, R. (2009). 'Nonpersistent inequality in educational attainment. Evidence from eight European countries', *American Journal of Sociology*, vol. 114(5), pp. 1475–1521.

Breen, R., van de Werfhorst, H.G. and Jæger, M.M. (2014). 'Deciding under Doubt: A Theory of Risk Aversion, Time Discounting Preferences, and Educational Decision-making', *European Sociological Review*, vol. 30(2), pp. 258–270.

Brunello, G. and Checchi, D. (2007). 'Does school tracking affect equality of opportunity? New international evidence', *Economic Policy*, vol. 22(52), pp. 782–861.

Brunello, G., Giannini, M. and Ariga, K. (2007). 'The optimal timing of school tracking: A general model with calibration for Germany', in (L. Woessmann and P. Peterson, eds.), *Schools and the Equal Opportunity Problem*, pp. 129–156, Cambridge: MIT Press.

Brunello, G. and Rocco, L. (2017). 'The Labor Market Effects of Academic and Vocational Education over the Life Cycle: Evidence Based on a British Cohort', *Journal of Human Capital*, vol. 11(1), pp. 106–166.

Brunner, E.J., Dougherty, S.M. and Ross, S.L. (2021). 'The Effects of Career and Technical Education: Evidence from the Connecticut Technical High School System', *The Review of Economics and Statistics*, pp. 1–46.

Brzinsky-Fay, C. (2007). 'Lost in Transition? Labour Market Entry Sequences of School Leavers in Europe', *European Sociological Review*, vol. 23(4), pp. 409–422.

Buchholz, S. and Schier, A. (2015). 'New Game, New Chance? Social Inequalities and Upgrading Secondary School Qualifications in West Germany', *European Sociological Review*, vol. 31(5), pp. 603–615.

Burke, M.A. and Sass, T.R. (2013). 'Classroom peer effects and student achievement', *Journal of Labor Economics*, vol. 31(1), pp. 51–82.

Bursztyn, L. and Jensen, R. (2015). 'How does peer pressure affect educational investments?', *The Quarterly Journal of Economics*, vol. 130(3), pp. 1329–1367.

Busemeyer, M.R. and Trampusch, C. (2011). 'The Comparative Political Economy of Collective Skill Formation', in (M. R. Busemeyer and C. Trampusch, eds.), *The Political Economy of Collective Skill Formation*, pp. 3–38, Oxford University Press.

Cameron, A.C., Gelbach, J.B. and Miller, D.L. (2008). 'Bootstrap-Based Improvements for Inference with Clustered Errors', *The Review of Economics and Statistics*, vol. 90(3), pp. 414–427.

Card, D. (1999). 'The Causal Effect of Education on Earnings', in (O. C. Ashenfelter and D. Card, eds.), *Handbook of Labor Economics*, pp. 1801–1863, vol. 3, Elsevier.

Card, D. (2001). 'Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems', *Econometrica*, vol. 69(5), pp. 1127–1160.

Card, D. and Giuliano, L. (2016). 'Can tracking raise the test scores of high-ability minority students?', *American Economic Review*, vol. 106(10), pp. 2783–2816.

Carneiro, P., Heckman, J.J. and Vytlacil, E.J. (2011). 'Estimating Marginal Returns to Education', *American Economic Review*, vol. 101(6), pp. 2754–2781.

Carrell, S.E., Fullerton, R.L. and West, J.E. (2009). 'Does your cohort matter? Measuring peer effects in college achievement', *Journal of Labor Economics*, vol. 27(3), pp. 439–464.

Carrell, S.E., Hoekstra, M. and Kuka, E. (2018). 'The long-run effects of disruptive peers', *American Economic Review*, vol. 108(11), pp. 3377–3415.

Cattaneo, M.D., Crump, R.K., Farrell, M.H. and Feng, Y. (2021). 'On Binscatter', *arXiv:1902.09608*.

Cavaglia, C., McNally, S. and Ventura, G. (2020). 'Do Apprenticeships Pay? Evidence for England', *Oxford Bulletin of Economics and Statistics*, vol. 82(5), pp. 1094–1134.

Checchi, D. (2006). *The Economics of Education: Human Capital, Family Background and Inequality*, Cambridge: Cambridge University Press.

Chetty, R. and Hendren, N. (2018). 'The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects', *The Quarterly Journal of Economics*, vol. 133(3), pp. 1107–1162.

Coleman, J.S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F. and York, R. (1966). *Equality of Educational Opportunity*, Washington, D.C.: U.S. Government Printing Office.

Collins, R. (1979). *The Credential Society: An Historical Sociology of Education and Stratification*, New York: Academic Press.

Conti, G. and Heckman, J.J. (2014). 'Economics of Child Well-Being', in (A. Ben-Arieh, F. Casas, I. Frønes and J. E. Korbin, eds.), *Handbook of Child Well-Being: Theories, Methods and Policies in Global Perspective*, pp. 363–401, Dordrecht: Springer Netherlands.

Contini, D. and Cugnata, F. (2016). 'Learning inequalities between primary and secondary school. Difference-in-Difference with international assessments', University of Turin 'Cognetti de Martiis' Working Paper No. 07/16.

Cunha, F. and Heckman, J.J. (2008). 'Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation', *Journal of Human Resources*, vol. 43(4), pp. 738–782.

Cunha, F., Heckman, J.J., Lochner, L. and Masterov, D.V. (2006). 'Interpreting the Evidence on Life Cycle Skill Formation', in (E. Hanushek and F. Welch, eds.), *Handbook of the Economics of Education*, pp. 697–812, vol. 1, Elsevier.

Dahl, G., Rooth, D.O. and Stenberg, A. (2020). 'High School Majors, Comparative (Dis)Advantage, and Future Earnings', National Bureau of Economic Research Working Paper 27524.

Dahl, G.B., Kostøl, A.R. and Mogstad, M. (2014). 'Family Welfare Cultures', *The Quarterly Journal of Economics*, vol. 129(4), pp. 1711–1752.

Dauth, W., Findeisen, S., Suedekum, J. and Woessner, N. (2021). 'The Adjustment of Labor Markets to Robots', *Journal of the European Economic Association*, vol. 19(6), pp. 3104–3153.

Dee, T.S. (2004). 'Are there civic returns to education?', *Journal of Public Economics*, vol. 88(9), pp. 1697–1720.

Dewey, J. (1916). *Democracy and Education: An Introduction to the Philosophy of Education*, New York: Macmillan.

Dobbie, W., Goldin, J. and Yang, C.S. (2018). 'The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges', *American Economic Review*, vol. 108(2), pp. 201–240.

Dollmann, J. (2016). 'Less Choice, Less Inequality? A Natural Experiment on Social and Ethnic Differences in Educational Decision-Making', *European Sociological Review*, vol. 32(2), pp. 203–215.

Duflo, E., Dupas, P. and Kremer, M. (2011). 'Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya', *American Economic Review*, vol. 101(5), pp. 1739–74.

Dumont, H., Klinge, D. and Maaz, K. (2019). 'The Many (Subtle) Ways Parents Game the System: Mixed-method Evidence on the Transition into Secondary-school Tracks in Germany', *Sociology of Education*, vol. 92(2), pp. 199–228.

Dumont, H., Protsch, P., Jansen, M. and Becker, M. (2017). 'Fish swimming into the ocean: How tracking relates to students' self-beliefs and school disengagement at the end of schooling', *Journal of Educational Psychology*, vol. 109(6), pp. 855–870.

Durazzi, N. (2019). 'The political economy of high skills: Higher education in knowledge-based labour markets', *Journal of European Public Policy*, vol. 26(12), pp. 1799–1817.

Dustmann, C. (2004). 'Parental background, secondary school track choice, and wages', *Oxford Economic Papers*, vol. 56(2), pp. 209–230.

Dustmann, C., Fitzenberger, B. and Machin, S. (2008). 'The economics of education and training', in (C. Dustmann, B. Fitzenberger and S. Machin, eds.), *The Economics of Education and Training*, pp. 1–5, Heidelberg: Physica.

Dustmann, C., Ludsteck, J. and Schönberg, U. (2009). 'Revisiting the German Wage Structure', *The Quarterly Journal of Economics*, vol. 124(2), pp. 843–881.

Dustmann, C., Puhani, P. and Schoenberg, U. (2017). 'The long-term effects of early track choice', *The Economic Journal*, vol. 127(603), pp. 1348–1380.

Edelstein, B. and Nikolai, R. (2013). 'Strukturwandel im Sekundarbereich. Determinanten schulpolitischer Reformprozesse in Sachsen und Hamburg', *Zeitschrift fuer Pädagogik*, vol. 59(4), pp. 482–494.

Edelstein, B. and Veith, H. (2017). 'Schulgeschichte nach 1945. Von der Nachkriegszeit bis zur Gegenwart', in (Bundeszentrale für politische Bildung and Wissenschaftszentrum Berlin für Sozialforschung, eds.), *Dossier Bildung*, Bonn: Bundeszentrale für politische Bildung.

Eisenkopf, G. (2007). 'Tracking and incentives', Thurgau Institute of Economics Research Paper No. 22.

R. Erikson and J. O. Jonsson, eds. (1996). *Can Education Be Equalized?: The Swedish Case in Comparative Perspective*, Boulder: Westview.

Ertl, H. and Phillips, D. (2000). 'The Enduring Nature of the Tripartite System of Secondary Schooling in Germany: Some Explanations', *British Journal of Educational Studies*, vol. 48(4), pp. 391–412.

European Commission (2014). 'European Vacancy and Recruitment Report', *Directorate-General for Employment, Social Affairs and Inclusion*.

Fend, H. (1982). *Gesamtschule Im Vergleich: Bilanz Der Ergebnisse Des Gesamtschulversuchs*, Weinheim: Beltz.

Fersterer, J., Pischke, J.S. and Winter-Ebmer, R. (2008). 'Returns to Apprenticeship Training in Austria: Evidence from Failed Firms', *The Scandinavian Journal of Economics*, vol. 110(4), pp. 733–753.

Figlio, D.N. (2007). 'Boys named Sue: Disruptive children and their peers', *Education Finance and Policy*, vol. 2(4), pp. 376–394.

Figlio, D.N. and Page, M.E. (2002). 'School choice and the distributional effects of ability tracking: Does separation increase inequality?', *Journal of Urban Economics*, vol. 51(3), pp. 497–514.

Fischer, L., Rohm, T., Gnambs, R. and Carstensen, C. (2016). 'Linking the data of the competence tests', NEPS Survey Paper No. 1.

Freeman, R. (1976). *The Overeducated American*, New York: Academic Press.

Galindo-Rueda, F. and Vignoles, A.F. (2004). 'The heterogeneous effect of selection in secondary schools: Understanding the changing role of ability', IZA Discussion Paper No. 1245.

Gamoran, A. and Mare, R.D. (1989). 'Secondary school tracking and educational inequality: Compensation, reinforcement, or neutrality?', *American Journal of Sociology*, vol. 94(5), pp. 1146–1183.

Garlick, R. (2018). 'Academic Peer Effects with Different Group Assignment Policies: Residential Tracking Versus Random Assignment', *American Economic Journal: Applied Economics*, vol. 10(3), pp. 345–369.

Goldin, C. (2001). 'The Human-Capital Century and American Leadership: Virtues of the Past', *The Journal of Economic History*, vol. 61(2), pp. 263–292.

Goldin, C. (2016). 'Human Capital', in (C. Diebolt and M. Haupert, eds.), *Handbook of Cliometrics*, pp. 55–86, Berlin, Heidelberg: Springer.

Goldin, C. and Katz, L.F. (2010). *The Race between Education and Technology*, Cambridge, MA: Harvard University Press.

Goodman-Bacon, A. (2021). 'Difference-in-differences with variation in treatment timing', *Journal of Econometrics*, vol. 225(2), pp. 254–277.

Graf, L. (2013). *The Hybridization of Vocational Training and Higher Education in Austria, Germany, and Switzerland*, Opladen, Berlin & Toronto: Budrich UniPress Ltd.

Graf, L. (2018). 'Combined modes of gradual change: The case of academic upgrading and declining collectivism in German skill formation', *Socio-Economic Review*, vol. 16(1), pp. 185–205.

Grossman, M. (2006). 'Education and Nonmarket Outcomes', in (E. Hanushek and F. Welch, eds.), *Handbook of the Economics of Education*, pp. 577–633, vol. 1, Elsevier.

Guyon, N., Maurin, E. and McNally, S. (2012). 'The effect of tracking students by ability into different schools a natural experiment', *Journal of Human Resources*, vol. 47(3), pp. 684–721.

Hall, C. (2016). 'Does more general education reduce the risk of future unemployment? Evidence from an expansion of vocational upper secondary education', *Economics of Education Review*, vol. 52, pp. 251–271.

Hall, P. and Kang, K.H. (2001). 'Bootstrapping nonparametric density estimators with empirically chosen bandwidths', *Annals of Statistics*, pp. 1443–1468.

Hampf, F. and Woessmann, L. (2017). 'Vocational vs. General Education and Employment over the Life Cycle: New Evidence from PIAAC', *CESifo Economic Studies*, vol. 63(3), pp. 255–269.

Hanushek, E.A. (1986). 'The Economics of Schooling: Production and Efficiency in Public Schools', *Journal of Economic Literature*, vol. 24(3), pp. 1141–1177.

Hanushek, E.A. (2003). 'The Failure of Input-based Schooling Policies', *The Economic Journal*, vol. 113(485), pp. F64–F98.

Hanushek, E.A., Schwerdt, G., Woessmann, L. and Zhang, L. (2017). 'General Education, Vocational Education, and Labor-Market Outcomes over the Lifecycle', *Journal of Human Resources*, vol. 52(1), pp. 48–87.

Hanushek, E.A. and Woessmann, L. (2006). 'Does educational tracking affect performance and inequality? Differences-in-Differences evidence across countries', *The Economic Journal*, vol. 116(510), pp. C63–C76.

Hanushek, E.A. and Woessmann, L. (2011). 'The Economics of International Differences in Educational Achievement', in (E. A. Hanushek, S. Machin and L. Woessmann, eds.), *Handbook of the Economics of Education*, pp. 89–200, vol. 3, Elsevier.

Hanushek, E.A. and Woessmann, L. (2015). *The Knowledge Capital of Nations: Education and the Economics of Growth*, MIT Press.

Havnes, T. and Mogstad, M. (2015). 'Is universal child care leveling the playing field?', *Journal of Public Economics*, vol. 127, pp. 100–114.

Heckman, J.J. and Pinto, R. (2018). 'Unordered Monotonicity', *Econometrica*, vol. 86(1), pp. 1–35.

Heckman, J.J. and Urzúa, S. (2010). 'Comparing IV with structural models: What simple IV can and cannot identify', *Journal of Econometrics*, vol. 156(1), pp. 27–37.

Heckman, J.J. and Vytlacil, E. (2005). 'Structural Equations, Treatment Effects, and Econometric Policy Evaluation', *Econometrica*, vol. 73(3), pp. 669–738.

Helbig, M. and Nikolai, R. (2015). *Die Unvergleichbaren. Der Wandel Der Schulsysteme in Den Deutschen Bundesländern Seit 1949*, Bad Heilbrunn: Julius Klinkhardt.

Helbig, M. and Nikolai, R. (2017). 'Alter Wolf im neuen Schafspelz? Die Persistenz sozialer Ungleichheiten im Berliner Schulsystem', WZB Discussion Paper No. P 2017-001.

Hirschl, N. and Smith, C.M. (2020). 'Well-Placed: The Geography of Opportunity and High School Effects on College Attendance', *Research in Higher Education*, vol. 61(5), pp. 567–587.

Huebener, M. (2020). 'Parental education and children's health throughout life', in (S. Bradley and C. Green, eds.), *The Economics of Education (Second Edition)*, pp. 91–102, Academic Press.

Hupkau, C., McNally, S., Ruiz-Valenzuela, J. and Ventura, G. (2017). 'Post-Compulsory Education in England: Choices and Implications', *National Institute Economic Review*, vol. 240(1), pp. R42–R57.

Hupkau, C. and Ventura, G. (2017). 'Further education in England: Learners and institutions', CVER Briefing Notes 001.

Imbens, G.W. and Angrist, J.D. (1994). 'Identification and Estimation of Local Average Treatment Effects', *Econometrica*, vol. 62(2), pp. 467–475.

Imbens, G.W. and Rubin, D.B. (1997). 'Estimating Outcome Distributions for Compliers in Instrumental Variables Models', *The Review of Economic Studies*, vol. 64(4), pp. 555–574.

Independent Panel on Technical Education (2016). 'Report of the Independent Panel on Technical Education (Sainsbury Review)', Department for Business, Innovation & Skills and Department for Education, London.

Jackson, C.K., Johnson, R.C. and Persico, C. (2016). 'The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms', *The Quarterly Journal of Economics*, vol. 131(1), pp. 157–218.

Jackson, M. (2021). 'Expansion, Enrollment, and Inequality of Educational Opportunity', *Sociological Methods & Research*, vol. 50(3), pp. 1215–1242.

Jacoby, T. and Dougherty, S.M. (2016). 'The new CTE: New York City as laboratory for America', Manhattan Institute Report No. 6.

Jerrim, J. and Micklewright, J. (2014). 'Socio-economic Gradients in Children's Cognitive Skills: Are Cross-Country Comparisons Robust to Who Reports Family Background?', *European Sociological Review*, vol. 30(6), pp. 766–781.

Kamhöfer, D.A., Schmitz, H. and Westphal, M. (2019). 'Heterogeneity in Marginal Non-Monetary Returns to Higher Education', *Journal of the European Economic Association*, vol. 17(1), pp. 205–244.

Kennedy, E.H., Lorch, S. and Small, D.S. (2019). 'Robust causal inference with continuous instruments using the local instrumental variable curve', *Journal of the Royal Statistical Society: Series B*, vol. 81(1), pp. 121–143.

A. C. Kerckhoff, K. Fogelman, D. Crook and D. Reeder, eds. (1996). *Going Comprehensive in England and Wales: A Study of Uneven Change*, London: Routledge.

Kerr, S.P., Pekkarinen, T. and Uusitalo, R. (2013). 'School tracking and development of cognitive skills', *Journal of Labor Economics*, vol. 31(3), pp. 577–602.

Kirkeboen, L.J., Leuven, E. and Mogstad, M. (2016). 'Field of Study, Earnings, and Self-Selection', *The Quarterly Journal of Economics*, vol. 131(3), pp. 1057–1111.

Klasik, D., Blagg, K. and Pekor, Z. (2018). 'Out of the Education Desert: How Limited Local College Options are Associated with Inequity in Postsecondary Opportunities', *Social Sciences*, vol. 7(9), p. 165.

Kline, P. and Walters, C.R. (2016). 'Evaluating Public Programs with Close Substitutes: The Case of Head Start', *The Quarterly Journal of Economics*, vol. 131(4), pp. 1795–1848.

Köller, O. (2008). 'Gesamtschule - Erweiterung statt Alternative', in (K. S. Cortina, J. Baumert, A. Leschinsky, K. U. Mayer and L. Trommer, eds.), *Das Bildungswesen in der Bundesrepublik Deutschland: Strukturen und Entwicklungen im Überblick*, Reinbek bei Hamburg: Rowohlt.

O. Köller, R. Watermann, U. Trautwein and O. Lüdtke, eds. (2004). *Wege Zur Hochschulreife in Baden-Württemberg. TOSCA – Eine Untersuchung an Allgemein Bildenden Und Beruflichen Gymnasien*, Opladen: Leske+Budrich.

Kosunen, S., Bernelius, V., Seppänen, P. and Porkka, M. (2020). 'School choice to lower secondary schools and mechanisms of segregation in urban Finland', *Urban education*, vol. 55(10), pp. 1461–1488.

Kreisman, D. and Stange, K. (2020). 'Vocational and Career Tech Education in American High Schools: The Value of Depth Over Breadth', *Education Finance and Policy*, vol. 15(1), pp. 11–44.

Krueger, A.B. (1999). 'Experimental Estimates of Education Production Functions', *The Quarterly Journal of Economics*, vol. 114(2), pp. 497–532.

Kruse, H. (2019). 'Between-school ability tracking and ethnic segregation in secondary schooling', *Social Forces*, vol. 98(1), pp. 119–146.

Kultusministerkonferenz (2014). *The Education System in the Federal Republic of Germany 2012/2013*, Bonn: Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Laender in the Federal Republic of Germany.

Lavy, V., Silva, O. and Weinhardt, F. (2012). 'The good, the bad, and the average: Evidence on ability peer effects in schools', *Journal of Labor Economics*, vol. 30(2), pp. 367–414.

Lee, J. and Barro, R.J. (2001). 'Schooling Quality in a Cross–Section of Countries', *Economica*, vol. 68(272), pp. 465–488.

Leschinsky, A. (2008). 'Die Realschule - Ein zweischneidiger Erfolg', in (K. S. Cortina, J. Baumert, A. Leschinsky, K. U. Mayer and L. Trommer, eds.), *Das Bildungswesen in der Bundesrepublik Deutschland: Strukturen und Entwicklungen im Überblick*, Reinbek bei Hamburg: Rowohlt.

A. Leschinsky and K. U. Mayer, eds. (1999). *The Comprehensive School Experiment Revisited: Evidence from Western Europe*, Frankfurt a. M.: Lang.

Lochner, L. and Moretti, E. (2004). 'The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports', *American Economic Review*, vol. 94(1), pp. 155–189.

Lochner, L.J. and Monge-Naranjo, A. (2011). 'The Nature of Credit Constraints and Human Capital', *American Economic Review*, vol. 101(6), pp. 2487–2529.

Machin, S., McNally, S., Terrier, C. and Ventura, G. (2020). 'Closing the Gap Between Vocational and General Education? Evidence from University Technical Colleges in England', CVER Research Paper 031.

Mackinnon, J.G. and Webb, M.D. (2017). 'Wild bootstrap inference for wildly different cluster sizes', *Journal of Applied Econometrics*, vol. 32(2), pp. 233–254.

MacKinnon, J.G. and Webb, M.D. (2018). 'The wild bootstrap for few (treated) clusters', *The Econometrics Journal*, vol. 21(2), pp. 114–135.

Malamud, O. and Pop-Eleches, C. (2011). 'School tracking and access to higher education among disadvantaged groups', *Journal of Public Economics*, vol. 95(11), pp. 1538–1549.

Mane, F. (1999). 'Trends in the payoff to academic and occupation-specific skills: The short and medium run returns to academic and vocational high school courses for non-college-bound students', *Economics of Education Review*, vol. 18(4), pp. 417–437.

Marcus, J. and Zambre, V. (2019). 'The effect of increasing education efficiency on university enrollment evidence from administrative data and an unusual schooling reform in Germany', *Journal of Human Resources*, vol. 54(2), pp. 468–502.

Matthewes, S.H. (2021). 'Better Together? Heterogeneous Effects of Tracking on Student Achievement', *The Economic Journal*, vol. 131(635), pp. 1269–1307.

Mayer, K.U., Müller, W. and Pollak, R. (2007). 'Germany: Institutional change and inequalities of access in higher education', in (Y. Shavit, R. Arum and A. Gamoran, eds.), *Stratification in Higher Education*, pp. 240–265, Standford: Stanford University Press.

Meghir, C. and Palme, M. (2005). 'Educational reform, ability, and family background', *American Economic Review*, vol. 95(1), pp. 414–424.

Michaels, G., Natraj, A. and Van Reenen, J. (2014). 'Has ICT Polarized Skill Demand? Evidence from Eleven Countries over Twenty-Five Years', *The Review of Economics and Statistics*, vol. 96(1), pp. 60–77.

Mincer, J.A. (1974). *Schooling, Experience, and Earnings*, National Bureau of Economic Research.

Moretti, E. (2004). 'Estimating the social return to higher education: Evidence from longitudinal and repeated cross-sectional data', *Journal of Econometrics*, vol. 121(1), pp. 175–212.

Mummolo, J. and Peterson, E. (2018). 'Improving the Interpretation of Fixed Effects Regression Results', *Political Science Research and Methods*, vol. 6(4), pp. 829–835.

Murphy, R. and Weinhardt, F. (2020). 'Top of the Class: The Importance of Ordinal Rank', *The Review of Economic Studies*, vol. 87(6), pp. 2777–2826.

Musset, P. and Field, S. (2013). *A Skills beyond School Review of England*, OECD Reviews of Vocational Education and Training, Paris: OECD Publishing.

Neugebauer, M. (2010). 'Bildungsungleichheit und Grundschulempfehlung beim Übergang auf das Gymnasium. Eine Dekomposition primärer und sekundärer Herkunftseffekte.', *Zeitschrift für Soziologie*, vol. 39(3), pp. 202–214.

Neugebauer, M., Reimer, D., Schindler, S. and Stocké, V. (2013). 'Inequality in Transitions to Secondary and Tertiary Education in Germany', in (M. Jackson, ed.), *Determined to Succeed? Performance versus Choice in Educational Attainment*, Standford: Stanford University Press.

Neumann, M., Becker, M., Baumert, J., Maaz, K., Köller, O. and Jansen, M. (2017). 'Das zweigliedrige Berliner Sekundarschulsystem auf dem Prüfstand: Ein Zwischenresümee', in (*Zweigliedrigkeit Im Deutschen Schulsystem – Potenziale Und Herausforderungen in Berlin*Münster: Waxmann.

Neumann, M., Becker, M. and Maaz, K. (2013). 'Die Abkehr von der traditionellen Dreigliedrigkeit im Sekundarschulsystem: Auf unterschiedlichen Wegen zum gleichen Ziel?', *Recht der Jugend und des Bildungswesens*, vol. 61(3), pp. 274–292.

Neumark, D. and Rothstein, D. (2007). 'Do School-To-Work Programs Help the 'Forgotten Half'?', in (D. Neumark, ed.), *Improving School-to-Work Transitions*, New York: Russell Sage Foundation.

Neumark, D., Schweitzer, M. and Wascher, W. (2004). 'Minimum wage effects throughout the wage distribution', *Journal of Human Resources*, vol. 39(2), pp. 425–450.

Oakes, J. (1985). *Keeping Track: How Schools Structure Inequality*, Yale University Press.

OECD (2017). *Getting Skills Right: Skills for Jobs Indicators*.

Oesch, D. (2013). *Occupational Change in Europe: How Technology and Education Transform the Job Structure*, Oxford: Oxford University Press.

Oesch, D. and Rodríguez Menés, J. (2011). 'Upgrading or polarization? Occupational change in Britain, Germany, Spain and Switzerland, 1990–2008', *Socio-Economic Review*, vol. 9(3), pp. 503–531.

Olea, J.L.M. and Pflueger, C. (2013). 'A Robust Test for Weak Instruments', *Journal of Business & Economic Statistics*, vol. 31(3), pp. 358–369.

Oosterbeek, H. and Webbink, D. (2007). 'Wage effects of an extra year of basic vocational education', *Economics of Education Review*, vol. 26(4), pp. 408–419.

Oreopoulos, P. and Salvanes, K.G. (2011). 'Priceless: The Nonpecuniary Benefits of Schooling', *Journal of Economic Perspectives*, vol. 25(1), pp. 159–184.

Owens, A. (2010). 'Neighborhoods and Schools as Competing and Reinforcing Contexts for Educational Attainment', *Sociology of Education*, vol. 83(4), pp. 287–311.

Pant, H.A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T. and Pöhlmann, C. (2013). *IQB-Ländervergleich 2012 - Zusatzmaterialien*, Münster: Waxmann.

Papageorge, N.W., Gershenson, S. and Kang, K.M. (2020). 'Teacher Expectations Matter', *The Review of Economics and Statistics*, vol. 102(2), pp. 234–251.

Paulus, W. and Blossfeld, H.P. (2007). 'Schichtspezifische Präferenzen oder sozioökonomisches Entscheidungskalkül? Zur Rolle elterlicher Bildungsaspirationen im Entscheidungsprozess beim übergang von der Grundschule in die Sekundarstufe', *Zeitschrift für Pädagogik*, vol. 53(4), pp. 491–508.

Pfeffer, F.T. (2008). 'Persistent Inequality in Educational Attainment and its Institutional Context', *European Sociological Review*, vol. 24(5), pp. 543–565.

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

Pischke, J.S. and Manning, A. (2006). 'Comprehensive versus selective schooling in England and Wales: What do we know?', NBER Working Paper No. 12176.

Powell, J.J. and Solga, H. (2011). 'Why are Participation Rates in Higher Education in Germany so Low? Institutional Barriers to Higher Education Expansion', *Journal of Education and Work*, vol. 24(1), pp. 49–68.

Psacharopoulos, G. and Patrinos, H.A. (2018). 'Returns to investment in education: A decennial review of the global literature', *Education Economics*, vol. 26(5), pp. 445–458.

Raffe, D., Brannen, K., Fairgrieve, J. and Martin, C. (2001). 'Participation, Inclusiveness, Academic Drift and Parity of Esteem: A comparison of post-compulsory education and training in England, Wales, Scotland and Northern Ireland', *Oxford Review of Education*, vol. 27(2), pp. 173–203.

Roodman, D., Nielsen, M.Ø., MacKinnon, J.G. and Webb, M.D. (2019). 'Fast and wild: Bootstrap inference in Stata using boottest', *The Stata Journal*, vol. 19(1), pp. 4–60.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Ryan, P. (2001). 'The School-to-Work Transition: A Cross-National Perspective', *Journal of Economic Literature*, vol. 39(1), pp. 34–92.

Sacerdote, B. (2001). 'Peer effects with random assignment: Results for Dartmouth roommates', *The Quarterly Journal of Economics*, vol. 116(2), pp. 681–704.

Sacerdote, B. (2011). 'Peer effects in education: How might they work, how big are they and how much do we know thus far?', pp. 249–277, vol. 3 of *Handbook of the Economics of Education*, Amsterdam: Elsevier.

Schindler, S. (2015). 'Soziale Ungleichheit im Bildungsverlauf. Alte Befunde und neue Schlüsse?', *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, vol. 67(3), pp. 509–537.

Schindler, S. (2017). 'School tracking, educational mobility and inequality in German secondary education: Developments across cohorts', *European Societies*, vol. 19(1), pp. 28–48.

Schindler, S. and Bittmann, F. (2021). 'Diversion or Inclusion? Alternative Routes to Higher Education Eligibility and Inequality in Educational Attainment in Germany', *European Sociological Review*, vol. 37(6), pp. 972–986.

Schober, P.S. and Spiess, C.K. (2013). 'Early childhood education activities and care arrangements of disadvantaged children in Germany', *Child Indicators Research*, vol. 6(4), pp. 709–735.

Schütz, G., Ursprung, H.W. and Woessmann, L. (2008). 'Education policy and equality of opportunity', *Kyklos*, vol. 61(2), pp. 279–308.

Schwerdt, G. and Ruhose, J. (2016). 'Does early educational tracking increase migrant-native achievement gaps? Differences-in-Differences evidence across countries', *Economics of Education Review*, vol. 52, pp. 134–154.

Y. Shavit and H.-P. Blossfeld, eds. (1993). *Persistent Inequality: Changing Educational Attainment in Thirteen Countries*, Boulder: Westview.

Shavit, Y. and Müller, W. (1998). *From School to Work. A Comparative Study of Educational Qualifications and Occupational Destinations*, Oxford: Clarendon Press.

Shavit, Y. and Müller, W. (2000). 'Vocational Secondary Education. Where diversion and where safety net?', *European Societies*, vol. 2(1), pp. 29–50.

Silliman, M. and Virtanen, H. (2022). 'Labor Market Returns to Vocational Secondary Education', *American Economic Journal: Applied Economics*, vol. 14(1), pp. 197–224.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Slavin, R.E. (1990). 'Achievement Effects of Ability Grouping in Secondary Schools: A Best-Evidence Synthesis', *Review of Educational Research*, vol. 60(3), pp. 471–499.

Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*, New York: Penguin Books 1982 reprint.

Solga, H., Protsch, P., Ebner, C. and Brzinsky-Fay, C. (2014). 'The German Vocational Education and Training System: Its Institutional Configuration, Strengths, and Challenges', WZB Discussion Paper No. SP I 2014-502, Berlin.

Spiess, C.K. and Wrohlich, K. (2010). 'Does distance determine who attends a university in Germany?', *Economics of Education Review*, vol. 29(3), pp. 470–479.

Spitz-Oener, A. (2006). 'Technical Change, Job Tasks, and Rising Educational Demands: Looking outside the Wage Structure', *Journal of Labor Economics*, vol. 24(2), pp. 235–270.

Stanat, P., Böhme, K., Schipolowski, S. and Haag, N. (2016). *IQB-Bildungstrend 2015: Sprachliche Kompetenzen Am Ende Der 9. Jahrgangsstufe Im Zweiten Ländervergleich*, Münster: Waxmann.

Stanat, P., Pant, H.A., Böhme, K. and Richter, D. (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik: Ergebnisse des IQB-Ländervergleichs 2011*, Münster: Waxmann.

Statistisches Bundesamt (2001a). *Allgemeinbildende Schulen, Fachserie 11, Reihe 1 (Yearly Statistical Publication for School Years 1995/96–2016/17)*, Wiesbaden: Statistisches Bundesamt.

Statistisches Bundesamt (2001b). *Berufliche Schulen, Fachserie 11, Reihe 2 (Yearly Statistical Publication for School Years 1995/96–2016/17)*, Wiesbaden: Statistisches Bundesamt.

Statistisches Bundesamt (2012). *Allgemeinbildende Schulen, Fachserie 11 Reihe 1, Schuljahr 2011/12*, Wiesbaden.

W. Streeck and K. Thelen, eds. (2005). *Beyond Continuity. Institutional Change in Advanced Political Economies*, New York: Oxford University Press.

Thelen, K. (2004). *How Institutions Evolve: The Political Economy of Skills in Germany, Britain, the United States and Japan*, Cambridge: Cambridge University Press.

Thurow, L.C. (1975). *Generating Inequality - Mechanisms of Distribution in the U.S. Economy*, New York: Basic Books.

Turley, R.N.L. (2009). 'College Proximity: Mapping Access to Opportunity', *Sociology of Education*, vol. 82(2), pp. 126–146.

US Department of Education (2012). *Investing in America's Future - A Blueprint for Transforming Career and Technical Education*, Washington, D.C.: US Department of Education.

van de Werfhorst, H.G. (2019). 'Early Tracking and Social Inequality in Educational Attainment: Educational Reforms in 21 European Countries', *American Journal of Education*, vol. 126(1), pp. 65–99.

van Ewijk, R. (2011). 'Same work, lower grade? Student ethnicity and teachers' subjective assessments', *Economics of Education Review*, vol. 30(5), pp. 1045–1058.

van Reenen, J. (2022). 'Innovation and human capital policy', in (A. Goolsbee and B. F. Jones, eds.), *Innovation and Public Policy*, University of Chicago Press.

Vieluf, U., Ivanov, S. and Nikolova, R. (2014). 'KESS 12/13 - Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der gymnasialen Oberstufe', Hamburger Schulbehörde, Hamburg.

Waldinger, F. (2007). 'Does tracking affect the importance of family background on students' test scores?', London School of Economics Working Paper.

Walters, P.B. (2000). 'The Limits of Growth. School Expansion and School Reform in Historical Perspective', in (M. T. Hallinan, ed.), *Handbook of the Sociology of Education*, pp. 241–261, New York: Springer.

Webb, M.D. (2014). 'Reworking wild bootstrap based inference for clustered errors', Queen's Economics Department Working Paper No. 1315.

Weber, S. and Péclat, M. (2017). 'A Simple Command to Calculate Travel Distance and Travel Time', *The Stata Journal*, vol. 17(4), pp. 962–971.

Whitmore, D. (2005). 'Resource and peer impacts on girls' academic achievement: Evidence from a randomized experiment', *American Economic Review*, vol. 95(2), pp. 199–203.

Winkler, O. (2017). *Aufstiege Und Abstiege Im Bildungsverlauf. Eine Empirische Untersuchung Zur Öffnung von Bildungswegen*, Halle: Springer.

Woessmann, L. (2003). 'Schooling Resources, Educational Institutions and Student Performance: The International Evidence', *Oxford Bulletin of Economics and Statistics*, vol. 65(2), pp. 117–170.

Woessmann, L. (2016). 'The importance of school systems: Evidence from international differences in student achievement', *Journal of Economic Perspectives*, vol. 30(3), pp. 3–32.

Wolf, A. (2011). 'Review of vocational education: The Wolf report', UK Department for Education and Department for Business Innovation and Skills.

Zilic, I. (2018). 'General versus vocational education: Lessons from a quasi-experiment in Croatia', *Economics of Education Review*, vol. 62, pp. 1–11.

Zimmermann, M. (2019). *Four Essays on the Economics of Education and Inequality*, Berlin: Humboldt-Universität zu Berlin.