Heisig, Jan Paul; Matthewes, Sönke Hendrik

No Evidence that Strict Educational Tracking Improves Student Performance through Classroom Homogeneity: A Critical Reanalysis of Esser and Seuring (2020)

Zeitschrift für Soziologie

Provided in Cooperation with:
WZB Berlin Social Science Center

Suggested Citation: Heisig, Jan Paul; Matthewes, Sönke Hendrik (2022) : No Evidence that Strict Educational Tracking Improves Student Performance through Classroom Homogeneity: A Critical Reanalysis of Esser and Seuring (2020), Zeitschrift für Soziologie, ISSN 2366-0325, De Gruyter, Berlin, Vol. 51, Iss. 1, pp. 99-111, https://doi.org/10.1515/zfsoz-2022-0001

This Version is available at:
https://hdl.handle.net/10419/262217

Jan Paul Heisig*, Sönke Hendrik Matthewes

No Evidence that Strict Educational Tracking Improves Student Performance through Classroom Homogeneity: A Critical Reanalysis of Esser and Seuring (2020)

Keine Belege für leistungsfördernde Effekte von strikter Leistungsdifferenzierung durch kognitive Homogenisierung: Eine kritische Reanalyse von Esser und Seuring (2020)

https://doi.org/10.1515/zfsoz-2022-0001

Abstract: In a recent contribution to this journal, Esser and Seuring (2020) draw on data from the National Educational Panel Study to attack the widespread view that tracking in lower secondary education exacerbates inequalities in student outcomes without improving average student performance. Exploiting variation in the strictness of tracking across 13 of the 16 German federal states (e. g., whether teacher recommendations are binding), Esser and Seuring claim to demonstrate that stricter tracking after grade 4 results in better performance in grade 7 and that this can be attributed to the greater homogeneity of classrooms under strict tracking. We show these conclusions to be untenable: Esser and Seuring's measures of classroom composition are highly dubious because the number of observed students is very small for many classrooms. Even when we adopt their classroom composition measures, simple corrections and extensions of their analysis reveal that there is no meaningful evidence for a positive relationship between classroom homogeneity and student achievement – the channel supposed to mediate the alleged positive effect of strict tracking. We go on to show that students from more strictly tracking states perform better already at the start of tracking (grade 5), which casts further doubt on the alleged positive effect of strict tracking on learning progress and leaves selection or anticipation effects as more plausible explanations. On a conceptual level, we emphasize that Esser and Seuring's analysis is limited to states that implement different forms of early tracking and cannot inform us about the relative performance of comprehensive and tracked systems that is the focus of most prior research.

Keywords: Ability Tracking; Secondary Education Systems; Peer Effects; Classroom Composition; Mediation Analysis; Replication.

Zusammenfassung: In einem kürzlich in dieser Zeitschrift veröffentlichten Artikel attackieren Esser und Seuring (2020) die verbreitete Auffassung, dass eine frühe Leistungsdifferenzierung in den ersten Jahren der Sekundarstufe Ungleichheiten zwischen Schüler*innen verstärkt, ohne sich positiv auf das durchschnittliche Leistungsniveau auszuwirken. Auf Basis einer Analyse von Daten des Nationalen Bildungspanels für 13 Bundesländer kommen die Autoren zu dem Ergebnis, dass sich eine strenge Leistungsdifferenzierung (z. B. durch bindende Grundschulempfehlungen) positiv auf das Leistungsniveau in Klasse 7 auswirkt und dass dies auf die homogenere Klassenzusammensetzung in strikt differenzierenden Ländern zurückgeführt werden kann. Der vorliegende Beitrag zeigt, dass diese Schlussfolgerungen nicht haltbar sind: Esser und Seurings Indikatoren für die Klassenzusammensetzung sind qualitativ fragwürdig, da die Anzahl gültiger Beobachtungen für viele Klassen sehr klein ist. Selbst bei Verwendung ihrer Indikatoren wird durch einfache Korrekturen und Ergänzungen ihrer Analyse schnell deutlich, dass es keine belastbaren empirischen Belege

***Corresponding author: Jan Paul Heisig**, Wissenschaftszentrum Berlin für Sozialforschung, Reichpietschufer 50, 10785 Berlin, E-Mail: jan.heisig@wzb.eu
Sönke Hendrik Matthewes, Universität Potsdam, Prof. f. VWL, insb. Wirtschaftspolitik, August-Bebel-Str. 89, 14482 Potsdam, E-Mail: matthewes@uni-potsdam.de

für den theoretisch zentralen positiven Zusammenhang zwischen homogener Klassenzusammensetzung und Leistungsniveau gibt. Zudem können wir zeigen, dass Schüler*innen in streng differenzierenden Ländern bereits zu Beginn der Sekundarstufe bessere Leistungen erzielen, ein weiteres Ergebnis, das gegen einen (kausalen) positiven Zusammenhang zwischen strenger Differenzierung und Lernfortschritt und für Alternativerklärungen wie Selektions- oder Antizipationseffekte spricht. In konzeptioneller Hinsicht heben wir hervor, dass sich die Analyse von Esser und Seuring auf verschiedene leistungsdifferenzierende Systeme beschränkt und insofern keine unmittelbaren Implikationen für den in der Literatur zentralen Vergleich zwischen differenzierenden und Gesamtschulsystemen (*comprehensive systems*) haben kann.

Schlüsselwörter: Leistungsdifferenzierung; Sekundarbildungssysteme; Peer-Effekte; Klassenzusammensetzung; Mediationsanalyse; Replikation.

1 Introduction

How does educational tracking – the allocation of students to different educational programs on the basis of (perceived) ability – affect student outcomes? This question has been the subject of heated academic and public debates for decades. Advocates of tracking posit that more homogeneous classrooms allow for better tailoring of curricula and instruction style to students' abilities and should, therefore, boost competence development for all students.

Critics, in contrast, fear that only high track/ability students benefit from tracking, whereas students assigned to lower tracks lose out compared to a scenario with comprehensive schooling. This is because, first, assignment to lower tracks might be stigmatizing and undermine student self-perceptions through "stereotype threat". Second, classrooms dominated by low-performing students might create a number of challenges hampering effective instruction. Third, homogeneous learning environments might deprive low-performing students of valuable role models and peer support from higher-performing students. Crucially, as students from disadvantaged backgrounds are generally overrepresented in lower tracks (due to lower performance and potential biases in track placement), such impediments to student development would exacerbate (social) inequalities in educational success.

The empirical literature on the effects of tracking is not fully conclusive, but the predominant view appears to be that early and rigid forms of tracking (in particular, "external differentiation" between different school types or "between-school tracking", as opposed to internal and often subject-specific differentiation within educational programs) indeed reinforce educational inequalities by previous achievement (e. g., Guyon et al. 2012; Matthewes 2021; Roller & Steinberg 2020), by socio-economic background (e. g., Bol & Van de Werfhorst 2013; Brunello & Checchi 2007; Heisig et al. 2020; Horn 2009; Kerr et al. 2013; Marks 2005; Pfeffer 2015; Schütz et al. 2008; Werfhorst 2019), and by ethnicity/migration background (e. g., Ruhose & Schwerdt 2016). At the same time, most studies fail to find meaningful associations between tracking and average student performance or attainment, suggesting that apparent equity costs of tracking are not counterbalanced by gains in efficiency (e. g., Betts 2011; Hanushek & Wößmann 2006; Pfeffer 2015; Roller & Steinberg 2020).

In a recent article in this journal, Esser & Seuring (2020) – hereafter ES – challenge the predominant negative assessment of tracking, offering, what they call, a "correcting replication" (p. 279) of the literature's "standard position" (p. 278). ES make three main theoretical assertions and claim these to be supported by an empirical analysis that exploits differences in the strictness of tracking across German federal states:

1. Educational tracking, if strictly implemented, improves student achievement and reduces rather than reinforces social background effects.
2. This is achieved through a more homogeneous composition of classrooms in terms of cognitive abilities (*mediation hypothesis*): "cognitive homogenization" promotes learning and is particularly beneficial for low-ability students.
3. The positive effects of classroom homogeneity are stronger in school systems characterized by strict tracking (*moderation hypothesis*).

In this reply, we provide a fundamental reassessment of ES's analysis and conclusions by replicating and extending their empirical analysis using the same data from the German National Educational Panel Study (NEPS). We show that the first two claims cannot be upheld and that empirical support for the third is limited at best, thus rejecting their rejection of the previous literature.

We proceed as follows: Section 2 briefly summarizes key arguments from ES's article that are essential for understanding their predictions, empirical approach, and substantive conclusions. Section 3 contains our reassessment of ES's findings and conclusions, including a replication and extension of their empirical analysis. Section 4 concludes with a brief summary and an outlook.

2 Esser and Seuring's critique of the standard position

ES's contribution begins with a critical assessment of research supporting the standard position. Their main objection, especially regarding research based on cross-national comparisons, is that most studies do not account for differences in student ability. According to ES this leads to "classic omitted variable bias" (cf. Esser & Seuring 2020: 279): the better educational outcomes of students from higher social classes are (mis)attributed to their social origins whereas they really reflect their higher academic abilities. The obvious implication of this critique is that empirical analyses of the effects of ability tracking need to account for differences in student ability.

ES go on to present a positive theoretical account of why tracking in secondary education might increase student performance and reduce social inequalities in educational outcomes, the so-called "Model of Ability Tracking" (MoAbiT). The details of the model are beyond the scope of the present contribution and were partly developed in papers that precede ES (e. g., Esser 2016; Esser & Hoenig 2018). In keeping with classic arguments for educational tracking, their main argument for a positive effect of tracking on student performance is the formation of (ability-)homogeneous classrooms that allow curricula and methods of instruction to be tailored to the needs and abilities of each group of students. To maximize the alleged benefits of tracking it is essential that tracking follows student ability as closely as possible. This leads ES to predict that the inequality-reducing and performance-enhancing effects of tracking should be strongest when tracking is strict; that is, when the rules governing track allocation maximize the importance of student ability and minimize other considerations such as students' and parents' preferences and aspirations (e. g., by making track recommendations binding).

3 Extension and reassessment of Esser and Seuring's analysis

We now turn to our replication and extension of ES's analysis. We begin by stating our main criticisms and then elaborate them in the remainder of this section:

1. In contrast to the previous literature which compares school systems that track with systems that do not, ES compare states that track with different levels of strictness. Accordingly, they estimate the effect of different sorting mechanisms *conditional* on having an early-tracking system but not the effect of tracked *vs.* comprehensive schooling.
2. The NEPS data are not well-suited for studying the role of classroom composition because the number of students observed per classroom is often very small. This raises serious concerns about the quality of the classroom composition measures.
3. ES's conclusions are not well supported even by their own analysis. In particular, they fall prey to a common mistake in the interpretation of interaction effects and interpret the main effects of the classroom homogeneity measures unconditionally. Correct interpretation of their regressions implies that the alleged positive effects of classroom homogeneity on student performance are restricted to classrooms with very low levels of average student ability (or SES) – and even this conditional positive effect of homogeneity disappears when we use improved measures of classroom composition (see Criticism 2). In addition, our extension of ES's analysis shows that their data provide essentially no evidence for a mediation of the effect of strict tracking through classroom homogeneity – the MoAbiT's most central claim.
4. While ES's critique of the previous literature emphasizes a lack of adjustment for (pre-existing) differences in student ability, their own analysis falls short in this regard. ES's only ability control is a short general cognition test administered at the start of tracking in grade 5. When we additionally control for baseline differences in *achievement* at the start of tracking, the positive effect of strictness of tracking disappears. This result is reinforced by placebo tests that use grade 5 achievement scores as the outcome and show that the alleged effects of tracking on achievement are visible already before it has really started. Taken together, this suggests that the relationship between strictness of tracking and achievement is either spurious or that strict tracking affects student performance before it actually occurs (e. g., through incentive effects).
5. ES also propose an innovative moderation hypothesis that predicts the effect of classroom homogeneity on student performance to be particularly strong in strictly tracking systems. According to ES, such moderation might occur because strict tracking facilitates the design of optimally tailored curricula and creates a meritocratic "educational climate" (cf. Esser and Seuring 2020: 293). Yet, ES's assessment of

this hypothesis is based on overly complex models with three-way interactions that make it difficult to draw definitive conclusions. Our reanalysis based on simpler specifications suggests that there is some limited evidence for the hypothesis, but that it is not robust to controlling for baseline achievement differences and by no means strong enough to warrant ES's conclusion that the MoAbiT receives strong and unambiguous support.

3.1 Criticism 1: Esser and Seuring's treatment definition differs from previous studies

A first limitation of ES's analysis concerns the scope of institutional variation covered in their analysis. Their analysis exploits institutional variation across 13 of the 16 German federal states (*Bundesländer*).[1] Notably, all of these states have *between-school* tracked secondary education systems that start to track in grade 5, when students are about 10 years of age. Hence, all of the 13 education systems included in ES's analysis fall into the most strongly or early-tracking systems by international standards. Their key institutional variable accordingly is neither the timing of tracking nor the number of school tracks – the two features that have received the greatest attention in (cross-national) research – but the *strictness* of tracking, as captured by a qualitative grouping of the German federal states into three groups based on whether track recommendations at the end of primary schools are binding (or can be overridden by parents) and the extent to which the rules for these recommendations are standardized. As such, ES's analysis covers only a small portion of the international variation in secondary education systems and cannot provide direct evidence on the performance of comprehensive education with no tracking in secondary education.

While this point may seem obvious, we believe that it does not receive sufficient attention in ES's contribution – particularly since the authors frame their work as a rather comprehensive and wholesale refutation of the previous literature and the standard position, including work that takes a broad cross-national perspective.

1 Three states – Berlin, Brandenburg, Mecklenburg-Western Pomerania – are excluded because tracking occurs later for most students in these states, usually after grade 6.

3.2 Criticism 2: Esser and Seuring's measures of classroom composition are problematic

Given their emphasis on classroom homogeneity as the key variable mediating the effect of strict tracking on student achievement, direct measures of classroom composition are central to ES's analysis. ES characterize classroom composition in terms of both socio-economic status (SES), proxied by the maximum parental ISEI-88 score, and student ability (ABL), measured using the NEPS-MAT test, a short test of reasoning capabilities designed for the NEPS. For both dimensions, a first measure captures the *classroom average*, labelled NSES/NABL (with the prefix "N" for "niveau"), and a second measure captures *homogeneity as the additive inverse of the within-classroom standard deviation*, labelled HSES/HABL. We plot their distributions in Fig. A3 in the online appendix.

While these classroom characteristics are conceptually straightforward, there are two major issues with their implementation in ES. First, ES compute one unique value of each measure per classroom, using all students with complete information. This means that a student's own SES or ABL score is used in calculating the respective classroom averages assigned to her. This can lead to serious bias by creating spurious correlations between individual characteristics and classroom composition (Angrist 2014). It is generally preferable to construct peer averages using a "leave-i-out" approach, where a student's own value in question is not included in the calculation of the classroom average (Angrist 2014).

Second, the number of students with complete information is very (almost certainly too) small for many classrooms. We illustrate this in Fig. A1 in the online appendix, which plots the number of students with valid information on SES and ABL for each classroom: in more than half of the classrooms the composition measures are calculated using fewer than 10 students (*average* class size in Germany is 22). Not only does this create problems of measurement error, failure to implement a "leave-i-out" approach will also be particularly consequential for small classrooms. Both issues become very obvious when one considers the non-negligible number of classroom where there is only *one* student with valid information. In these cases, the classroom averages, NSES and NABL, are identical to the individual-level measures, SES and ABL. Even more problematically, ES set the within-classroom standard deviations of SES and ABL to zero for these classrooms, so the homogeneity

measures (HSES and HABL) are maximal by construction.[2]

Even though these are serious issues and, at the very least, classrooms with only one student observation should be excluded from an analysis focusing on the role of classroom composition, we stick to ES's variable and sample definitions in our replications below to avoid differences due to different samples. We have, however, re-run the main regressions with improved "leave-i-out" measures of the classroom averages and excluding classrooms with only one student observation from the sample (as it is impossible to construct reasonable estimates of classroom homogeneity in these cases and ES's assumption of maximal homogeneity is clearly untenable). These additional regressions, reported in Tab. B1 to B3 in the online appendix, provide even less support for ES's claims than the regressions that we present in the main article (see in particular Section 3.3).[3]

3.3 Criticism 3: Classroom homogeneity shows no unambiguous relationship with student performance and does not mediate the effect of strict tracking

We begin our reanalysis of ES with the results pertaining to the hypothesized mediation of the effect of (strict) tracking through classroom homogeneity, which are presented in Tab. 4 of their paper. Our reanalysis, presented in Tab. 1, deviates from ES only by adding several model specifications that are missing from their analysis, despite being crucial for testing their claims.[4] Reassuringly, all models presented in ES are exactly reproduced in our reanalysis.

As discussed above, ES emphasize the importance of controlling for potential selection effects (i. e., differences in student composition between states with different strictness of tracking), yet curiously their models only include student-level ability and SES when they also include the corresponding classroom-level aggregates. We agree that good selectivity controls are important and therefore include individual-level SES and ABL in all specifications (save for baseline Model 1). This means that we report a key specification missing from ES's analysis: Model 2 is a controlled baseline model that includes individual ABL and SES as selection controls but does not yet include any of the compositional measures, thus allowing the reader to inspect the effect of tracking before moving to the mediation analysis. Comparison of Models 1 and 2 shows that the inclusion of ABL and SES alone noticeably attenuates the strictness of the tracking coefficients, leaving the contrast between the least strict (T1, the reference category) and the medium strict (T2) states significant at a 10 % level only. ES misattribute this attenuation to the classroom composition measures because they add the classroom- and student-level measures jointly in one step.

We now turn to the potential mediation of the (remaining) effect of tracking through classroom composition. Somewhat unconventionally, ES's assessment of the mediation hypothesis is based exclusively on "interactive" specifications where the effects of homogeneity (i. e., HABL/HSES) are allowed to vary by the corresponding classroom average (i. e., NABL/NSES). We follow a more standard approach and enter the composition measures sequentially and additively before including interactions. ES's central claim is that the positive effect of strict tracking operates through the creation of more homogeneous classrooms. As the most straightforward test of this claim, Model 3 in Tab. 1 only adds the measure for homogeneity in terms of ability, HABL, to the controlled baseline model. This key specification is missing from ES. Strikingly, the addition of HABL leaves the coefficients on the strictness of tracking indicators virtually unchanged (if anything, the effect of T2 *increases*), providing strong evidence *against* mediation of the effect of strictness of tracking through HABL. The inclusion of further compositional measures in the remaining columns does lead to some attenuation

2 In Fig. A2 in the online appendix we plot the number of observed students against the homogeneity measure for both ABL and SES. Two patterns stand out: (i) the homogeneity measures are negatively correlated with classroom size, mainly due to the (erroneously imputed) single-observation classrooms, and (ii) due to measurement error both measures also have higher variance among classrooms with few observations.

3 Note that for these regressions we have z-standardised the classroom composition variables to have a mean of zero and a standard deviation (SD) of one in order to ease coefficient interpretability (as the effect of a one SD increase of the independent variable). Moreover, classroom- and student-level sample sizes are slightly lower than in the main analysis due to the exclusion of single-student classrooms.

4 Among other things, this means that we estimate simple-linear mixed-effects models with random intercepts at the class level. Like ES, we neither include an additional random intercept at the state level nor state-level random slopes for the classroom composition measures, and we assume normally distributed test statistics. While it is beyond the scope of our contribution to systematically explore the impact of these specification choices, extant methodological work suggests that they will result in anti-conservative statistical inference (underestimation of standard errors and overrejection of null hypotheses; see, for example, Elff et al. 2021; Heisig & Schaeffer 2019).

Tab. 1. Mediation analysis.

	Baseline		Ability composition				SES composition		Both	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Model no. in ES:	1							4		5
Strictness of tracking										
T2	0.22**	0.15*	0.17**	0.09	0.09	0.09	0.09	0.09	0.08	0.08
	(0.11)	(0.08)	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.06)	(0.06)
T3	0.29***	0.23***	0.23***	0.18***	0.17***	0.17***	0.15***	0.15***	0.15***	0.14***
	(0.09)	(0.07)	(0.07)	(0.05)	(0.05)	(0.05)	(0.06)	(0.06)	(0.05)	(0.05)
Classroom ability composition										
NABL				2.00***	1.99***	2.84***			1.40***	2.19***
				(0.14)	(0.14)	(0.46)			(0.17)	(0.46)
HABL			0.53**		0.07	0.96*			0.14	1.01**
			(0.21)		(0.18)	(0.49)			(0.17)	(0.49)
NABL × HABL						−1.62*				−1.58*
						(0.83)				(0.81)
Classroom SES composition										
NSES							2.49***	4.56***	1.37***	2.39***
							(0.20)	(0.79)	(0.23)	(0.78)
HSES							−0.06	1.42**	−0.09	0.58
							(0.20)	(0.58)	(0.19)	(0.56)
NSES × HSES								−3.06***		−1.48
								(1.12)		(1.08)
Individual controls										
ABL		1.70***	1.70***	1.43***	1.43***	1.43***	1.60***	1.59***	1.43***	1.43***
		(0.07)	(0.07)	(0.08)	(0.08)	(0.08)	(0.07)	(0.07)	(0.08)	(0.08)
SES		0.65***	0.65***	0.54***	0.54***	0.53***	0.40***	0.40***	0.40***	0.40***
		(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)	(0.07)	(0.07)
Gender	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Migration	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ECEC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
N classes	349	349	349	349	349	349	349	349	349	349
N students	2662	2662	2662	2662	2662	2662	2662	2662	2662	2662

Notes: Shown are coefficients and, in parentheses, standard errors from multilevel mixed-effects linear regressions with random intercepts at the class level. The dependent variable is the average of grade 7 math and reading test scores, standardized to mean zero and unit standard deviation. T2 = Medium strictness of tracking; T3 = High strictness of tracking; ABL = cognitive abilities; SES = socio-economic status; NABL = classroom average cognitive abilities; HABL = classroom homogeneity cognitive abilities; NSES = classroom average socio-economic status; HSES = classroom homogeneity socio-economic status; ECEC = early-childhood education and care attendance in months. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

of the strictness of tracking coefficients, but this is driven by the classroom *average*s (NABL and NSES). The homogeneity measures (HABL and HSES) emphasized by ES play essentially no role in this attenuation.[5]

More than that, the results in Tab. 1 do not even show a clear-cut positive relation between performance and classroom homogeneity. While the coefficient on HABL is positive in Model 3, Model 5 shows that when HABL and NABL are added jointly, the coefficient on homogeneity drops to zero, indicating that the positive association between HABL and performance in Model 3 is spurious and comes from more homogeneous classroom typically being classrooms with higher average ability levels. Model 6 adds the interaction between NABL and HABL (as noted above, ES only show such interactive specifications and do not con-

5 In addition, Fig. A4 in the online appendix shows that differences in (measured) classroom homogeneity between the three state groups T1, T2, and T3, are much less clear-cut than one would expect. In particular, the least strictly tracking states (T1) appear to have more homogeneous classrooms on average than the medium group (T2).

sider the kinds of additive specifications that we report in columns 3–5, 7, and 9). While the coefficient on HABL is indeed positively signed and statistically significant in Model 6, the presence of the interaction term implies that it no longer corresponds to a general or average "effect". As is well-known (see, for example, Kam and Franzese 2007), it now has to be interpreted as the effect of HABL for classrooms with NABL = 0, that is, for classrooms *with minimal average ability* (because, following ES, the classroom composition measures are rescaled to range from 0 to 1). The strongly negative interaction term implies that the predicted effect of homogeneity starts to turn negative in the middle ranges of the classroom ability distribution (more precisely, when NABL is around .59 because .96/1.62 ≈ .59), which is consistent with the near-zero average effect from Model 5. ES do not make these important qualifications when discussing the corresponding estimates, nurturing the misleading impression that their results are consistent with the *general* achievement-enhancing effect of homogeneity suggested by the MoAbiT.[6]

Altogether, these results – at most – support a performance-enhancing effect of homogeneity for a subset of (low-ability) classrooms, with essentially no evidence that this effect can account for the relationship between strict tracking and performance. Importantly, even the conditional positive effect of homogeneity for low-ability classrooms becomes highly questionable once we address the problems with ES's classroom composition measures noted above (see Section 3.2): Tab. B1 in the online appendix shows that not only the average effect of HABL (in Models 5 and 9) but also its main effect and interaction with NABL (in Models 6 and 10) are essentially zero when the classroom composition measures are calculated correctly (i. e., when NABL is constructed using the "leave-i-out" approach and single-student classrooms are excluded instead of adopting ES's assumption that these classrooms are maximally homogeneous).

3.4 Criticism 4: The relationship between strict tracking and student performance is largely accounted for by initial performance

While the analysis of the previous section casts serious doubt on the mediating role of classroom homogeneity, it does not explain why seventh-graders tend to perform better in states with stricter tracking. In this section, we show that students in strictly tracking states perform better already at the start of secondary school (grade 5) and that this initial advantage explains the largest portion of their performance advantage in grade 7. Hence, the achievement advantage of students in strictly tracking systems appears to reflect selection and/or anticipation effects rather than the positive effects of strict tracking on competence development in secondary school emphasized by ES.

As student bodies differ substantially between German federal states, the key challenge for any study aiming to identify causal effects of schooling policy using between-state policy variation is to convincingly control for these differences. ES rely on a rather small control set that consists only of indicators for gender, migration background and early childcare attendance, next to the two mentioned continuous measures of family background (SES) and ability (ABL). The latter is based on a very short general cognition (reasoning) test administered at the beginning of grade 5.[7] The outcome variable, ACH, measures grade 7 achievement by averaging two elaborate domain-specific competence tests in mathematics and reading. Importantly, similar competence tests in mathematics and reading were also administered at the beginning of grade 5.[8] This makes it possible to control for achievement differences at baseline – when there has been only minimal exposure to tracked secondary education – in the regressions for grade 7 achievement. Somewhat surprisingly, ES do not make use of this information in their paper.

Tab. 2 presents this natural robustness check.[9] Models 2, 3, and 4 repeat key specifications from Tab. 1 (Models 2, 5, and 10, respectively) with grade 5 achievement, Gr. 5 ACH, added as an individual-level control for initial dif-

6 As an aside, we note that ES's presentation and interpretation of their results is rendered even more confusing by their listing of the MoAbiT's alleged implications in the first column of their Tab. 4 (where "+" is meant to indicate that the MoAbiT predicts a positive coefficient and "0" or "≥0" are meant to indicate that the implications of the model are weaker or unclear). Here, ES indicate that the main effects of NABL and HABL (as well as those of HSES and NSES) are all predicted to be positive, but it is difficult to see how such a prediction in terms of the direction of the *conditional effect of homogeneity for classrooms with minimal average ability* (and *vice versa*) follows from the MoAbiT, except as a special case of a general performance-enhancing effect that our additional specifications show not to be supported by the data.

7 Students have 9 minutes to answer 12 questions and are simply scored by the number of correct answers (i. e., on a discrete 13-point scale; see Fig. A3 in the online appendix). See Haberkorn & Pohl (2013) for further details.

8 More precisely, the competence tests (including the general cognition test underlying ABL) were administered during the first months of the school year. See footnote 11 for further discussion.

9 Note that the sample size is slightly lower than in Tab. 1 due to missing data on grade 5 achievment.

Tab. 2. Selection as an alternative explanation.

Dependent variable	Grade 7 achievement				Grade 5 achievement		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Strictness of tracking							
T2	0.15*	0.03	0.03	0.03	0.16**	0.10*	0.10*
	(0.08)	(0.05)	(0.05)	(0.04)	(0.08)	(0.06)	(0.06)
T3	0.24***	0.08**	0.06*	0.05	0.25***	0.20***	0.17***
	(0.07)	(0.04)	(0.04)	(0.04)	(0.06)	(0.05)	(0.05)
Classroom ability composition							
NABL			0.86***	1.05***		2.00***	1.97***
			(0.11)	(0.34)		(0.14)	(0.43)
HABL			0.17	0.63*		−0.22	0.56
			(0.13)	(0.37)		(0.17)	(0.46)
NABL × HABL				−0.78			−1.26
				(0.61)			(0.76)
Classroom SES composition							
NSES				0.95			2.57***
				(0.59)			(0.74)
HSES				0.26			0.58
				(0.43)			(0.53)
NSES × HSES				−0.63			−1.53
				(0.82)			(1.02)
Individual controls							
Gr. 5 ACH (z-score)		0.59***	0.56***	0.55***			
		(0.02)	(0.02)	(0.02)			
ABL (z-score)	0.36***	0.16***	0.12***	0.13***	0.39***	0.32***	0.32***
	(0.02)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
SES (z-score)	0.14***	0.09***	0.08***	0.06***	0.11***	0.09***	0.05***
	(0.02)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Gender	✓	✓	✓	✓	✓	✓	✓
Migration	✓	✓	✓	✓	✓	✓	✓
ECEC	✓	✓	✓	✓	✓	✓	✓
N classes	349	349	349	349	349	349	349
N students	2659	2659	2659	2659	2659	2659	2659

Notes: Shown are coefficients and, in parentheses, standard errors from multilevel mixed-effects linear regressions with random intercepts at the class level. The dependent variable is the average of grade 7 (Models 1–4) or grade 5 (Models 5–7) math and reading test scores, standardized to mean zero and unit standard deviation. T2 = Medium strictness of tracking; T3 = High strictness of tracking; Gr. 5 ACH = Grade 5 achievement; ABL = cognitive abilities; SES = socio-economic status; NABL = classroom average cognitive abilities; HABL = classroom homogeneity cognitive abilities; NSES = classroom average socio-economic status; HSES = classroom homogeneity socio-economic status; ECEC = early-childhood education and care attendance in months. Grade 5 scores are missing for three students, explaining the difference in the number of observations compared to Tab. 1. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

ferences in student achievement. For reference, in Model 1 we repeat the controlled baseline model without grade 5 achievement from above (i. e., Model 2 in Tab. 1). Note that we now z-standardize the individual-level measures ABL and SES to ease comparisons with Gr. 5 ACH.[10] Comparing Models 1 and 2, we see that grade 5 achievement turns out highly predictive of grade 7 achievement – in fact, much more so than ES's ability control, ABL. More importantly, controlling for grade 5 achievement leads to a drastic attenuation of the tracking coefficients: the difference in grade 7 achievement between the most and least strictly tracking states drops to .08 standard deviations (SD) – merely one third of the difference before controlling for initial achievement (see Model 1) – and the difference between the medium and least strictly tracking group declines to a negligible and statistically insignificant .03 SDs. This indicates that the largest portion of the relationship

10 This is without loss of generality. In particular, it does not affect the coefficient estimates for the other variables in the model.

between tracking and grade 7 performance is attributable to students in strictly tracking states being stronger performers already at the start of secondary school.

This conclusion is reinforced by a direct (placebo) test for pre-treatment differences between more and less strictly tracking states in Models 5 to 7. Specifically, we now use grade 5 rather than grade 7 achievement as the dependent variable. If strict tracking enhances student performance by creating homogeneous learning environments, as stipulated by the MoAbiT, we should see performance advantages emerge over the course of secondary education because of continued exposure to the supposedly beneficial homogeneous learning environments. However, Models 5 to 7 show that the positive relationship between tracking and achievement already exists in grade 5 when exposure to tracked programs has been minimal, and that it is about as strong as in grade 7 (in terms of SDs of the outcome variable).[11]

These findings seem to leave two (non-mutually exclusive) possibilities. The first is that the association between strictness of tracking and student performance is spurious in the sense of being attributable to unobserved factors that affect student achievement both in grade 5 and in grade 7. A second possibility is that tracking does exert a causal effect on student performance but that this effect occurs already prior to the onset of actual tracking. In particular, strict tracking might create performance incentives and induce greater educational investments among primary school students, their teachers, and/or their parents. In a recent study based on NEPS data, Bach and Fischer (2020) provide evidence supporting this interpretation. Clearly, such anticipation effects are very different from the homogeneity mechanism emphasized by ES.

3.5 Criticism 5: Evidence for interaction effects between homogeneity and strict tracking is weak at best

In this section, we revisit ES's moderation hypothesis that predicts stronger effects of classroom homogeneity on achievement in more strictly tracking systems. In a nutshell, ES's theoretical argument for such a moderation effect is that strict tracking facilitates the optimal tailoring of curricula, teacher education, and other factors to the different ability levels of students.

Empirically, such moderation should manifest in a positive interaction between the strictness of tracking and classroom homogeneity. Unfortunately, the results presented in Tab. 5 in ES make it extremely difficult to tell whether such an interaction exists, as the authors only present specifications that include three-way interactions between tracking and the two measures of classroom composition. In their discussion of the results on p. 293, ES point to the (positive and statistically significant) two-way interactions between strict tracking and classroom composition in terms of ability (i. e., T3*NABL and T3*HABL). However, the immediate theoretical significance of these two-way interactions is very limited, because due to the included three-way interaction (T3*NABL*HABL), they again refer to specific conditional effects: the coefficient on T3*HABL tells us how the "effect" of ability homogeneity *for classrooms with minimal average ability* differs between the most and least strictly tracking states. By the same token, T3*NABL captures how the effect of average ability *for classrooms with minimal homogeneity* differs between the two systems. These highly specific effects for classrooms located at the very extremes of the classroom composition distributions tell us little about a possible general moderating effect of tracking.

To see whether ES's moderation hypothesis is actually supported by the data, we estimate simpler specifications that do not include three-way interactions. Tab. 3 presents two blocks of models, each comprising four specifications. Models 1 to 4 use the same of set of controls as ES, so results can be directly compared. Models 5 to 8 include grade 5 achievement as an additional control, which, as shown above, is important to control for selectivity. The first model in each sequence is a baseline specification that does not include any interactions between strictness of tracking and the classroom composition measures. We then add the interactions between tracking and ability

11 An anonymous referee pointed out that the measures of grade 5 achievement and ABL are not strictly pre-treatment, as the corresponding tests were conducted during the first half of the school year (42.0 % were administered in November, 55.5 % in December, and 2.5 % in January). Moreover, the number of school days students had experienced when taking the tests – including not only those in grade 5, but also those in grade 7 – might vary systematically across states due to differences in the start of the school year (which starts somewhere between August and September depending on the state). In supplementary analyses, we therefore calculated timing-corrected ABL and achievement scores by regressing each one of them on the number of days between the start of the school year and the time of the test (though for the latter we do not have daily information but only monthly). We then re-ran the regressions in Tables 1–3 with the residuals from these regressions as our ability/achievement measures instead of the original scores (of course, also constructing the classroom composition measures using residualized ABL). The results, presented in Tab. C1 to C3 in the online appendix, are very similar to those in the main article and do not alter any of our conclusions. This is despite a small reduction in sample size relative to the main analysis due to missing data on test timing.

Tab. 3. Moderation analysis.

	Controls as in ES				Controlling for grade 5 achievement			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Strictness of tracking								
T2	0.08	−0.08	0.18	0.06	0.03	−0.07	−0.12	−0.17
	(0.06)	(0.26)	(0.48)	(0.49)	(0.04)	(0.19)	(0.35)	(0.36)
T3	0.15***	−0.56**	0.29	−0.13	0.05	−0.39**	0.08	−0.15
	(0.05)	(0.25)	(0.33)	(0.37)	(0.04)	(0.18)	(0.24)	(0.27)
Classroom ability composition								
NABL	1.40***	1.16***	1.43***	1.16***	0.66***	0.49***	0.68***	0.51***
	(0.17)	(0.22)	(0.17)	(0.25)	(0.13)	(0.16)	(0.13)	(0.18)
T2 × NABL		0.10		0.03		0.11		0.02
		(0.37)		(0.50)		(0.27)		(0.37)
T3 × NABL		0.59**		0.66*		0.41**		0.41
		(0.28)		(0.35)		(0.20)		(0.26)
HABL	0.14	−0.07	0.12	−0.13	0.19	0.09	0.18	0.06
	(0.17)	(0.24)	(0.17)	(0.25)	(0.13)	(0.18)	(0.13)	(0.19)
T2 × HABL		0.21		0.28		0.05		0.07
		(0.46)		(0.47)		(0.34)		(0.35)
T3 × HABL		0.67*		0.72*		0.35		0.38
		(0.40)		(0.40)		(0.29)		(0.29)
Classroom SES composition								
NSES	1.37***	1.41***	1.23***	1.50***	0.52***	0.55***	0.37*	0.54**
	(0.23)	(0.23)	(0.29)	(0.33)	(0.17)	(0.17)	(0.22)	(0.25)
T2 × NSES			0.09	0.03			0.15	0.13
			(0.48)	(0.66)			(0.34)	(0.48)
T3 × NSES			0.36	−0.22			0.35	−0.01
			(0.38)	(0.47)			(0.28)	(0.35)
HSES	−0.09	−0.11	0.13	0.17	−0.03	−0.06	0.08	0.09
	(0.19)	(0.19)	(0.26)	(0.27)	(0.14)	(0.14)	(0.20)	(0.20)
T2 × HSES			−0.24	−0.25			0.12	0.14
			(0.64)	(0.64)			(0.47)	(0.47)
T3 × HSES			−0.55	−0.62			−0.36	−0.38
			(0.40)	(0.40)			(0.30)	(0.30)
Individual controls								
Gr. 5 ACH					✓	✓	✓	✓
ABL	✓	✓	✓	✓	✓	✓	✓	✓
SES	✓	✓	✓	✓	✓	✓	✓	✓
Gender	✓	✓	✓	✓	✓	✓	✓	✓
Migration	✓	✓	✓	✓	✓	✓	✓	✓
ECEC	✓	✓	✓	✓	✓	✓	✓	✓
N classes	349	349	349	349	349	349	349	349
N students	2662	2662	2662	2662	2659	2659	2659	2659

Notes: Shown are coefficients and, in parentheses, standard errors from multilevel mixed-effects linear regressions with random intercepts at the class level. The dependent variable is the average of grade 7 math and reading test scores, standardized to mean zero and unit standard deviation. T2 = Medium strictness of tracking; T3 = High strictness of tracking; Gr. 5 ACH = Grade 5 achievement; ABL = cognitive abilities; SES = socio-economic status; NABL = classroom average cognitive abilities; HABL = classroom homogeneity cognitive abilities; NSES = classroom average socio-economic status; HSES = classroom homogeneity socio-economic status; ECEC = early-childhood education and care attendance in months. Grade 5 scores are missing for three students, explaining the lower number of observations in Models 5-8. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

composition (second) or SES composition (third), while the fourth model includes all two-way interactions.

When using the same set of controls as ES, we find some evidence for the predicted moderation effect: both Model 2 and 4 show (marginally) statistically significant positive interactions between strict tracking (T3) and the two measures of classroom ability composition (NABL and HABL) with meaningful effect sizes. However, the second set of models shows that these effects are not fully robust to controlling for grade 5 achievement, which the previous section has shown to be key in controlling for selectivity. The relevant interaction terms, T3*NABL and T3*HABL, remain positive, but their magnitude declines substantially and only one estimate – the T3*NABL interaction in Model 6 – continues to reach statistical significance. Hence, we conclude that the evidence for ES's moderation hypothesis, especially for the key theoretical variable HABL, is suggestive at best – certainly not strong enough to justify the bold wording used by ES, who conclude that "the results correspond almost perfectly to the theoretical model" ("Die Befunde entsprechen nahezu lückenlos dem theoretischen Modell"; Esser & Seuring 2020: 295, authors' translation).

4 Conclusions

The predominant view in the sociology of education and related fields appears to be that early and rigid tracking of students into different schools and educational programs exacerbates social inequalities in educational achievement and attainment, while offering no clear benefits in terms of average student outcomes. In a recent article in this journal, Esser & Seuring (2020) emphatically reject this "standard position" based on a theoretical account and an empirical analysis comparing 13 of the 16 German federal states.

In this article, we have offered a reassessment of Esser and Seuring's analysis and demonstrated that their sweeping claims are unsustainable: 1) In contrast to the previous literature which compares tracked school systems with comprehensive ones, Esser and Seuring compare states that track with different levels of strictness, limiting the scope of their findings: if anything, they are able to show that, given a between-school tracked system, allocation based on objective performance criteria outperforms allocation based on student (and parent) self-selection. 2) Their paper aims to investigate how the alleged effect of strict tracking operates through classroom composition. Yet, the compositional measures used by Esser and Seuring are of dubious quality, most importantly because the number of observed students is too small for many classrooms. 3) Even when we adopt their classroom composition measures, corrections of some key misinterpretations and obvious extensions of their analysis reveal that the NEPS data provides no support for a positive relationship between classroom homogeneity and student achievement – the key channel they claim to mediate the alleged positive effect of strict tracking. 4) What is more, we show that students from more strictly tracking states perform better already at the start of tracking, casting doubt on the existence of a substantial positive effect of strict tracking and leaving selection and/or anticipation effects as more plausible explanations for later performance differences. 5) Finally, the evidence for Esser and Seuring's intriguing and theoretically innovative hypothesis that strict tracking reinforces the alleged positive effects of classroom homogeneity (e. g., by allowing for better tailoring of curricula) is suggestive at best.

In conclusion, we do not see why and how Esser and Seuring's analysis should lead the scholarly community to change their priors on the relationship between tracking, achievement and social background effects. The debate on the ins and outs of tracking and other features of education systems is far from settled. We concur with Esser and Seuring that longitudinal analysis with good selectivity controls has the potential to move the literature forward. We also acknowledge the importance of being explicit about the mechanisms through which the effects of tracking might operate, and of trying to pin them down in empirical analysis. While Esser and Seuring's contribution is an ambitious attempt to address these desiderata, we have shown that their conclusions are not actually backed by the data. This is partly due to erroneous modelling choices and misinterpretations but also due to data limitations that hamper an accurate measurement of classroom composition. On the one hand, we therefore hope that our exchange with Esser and Seuring stimulates further research along the above lines; on the other hand, it highlights the need for greater investment into the appropriate survey and administrative data required for this type of research.

Hinweis zur Replikation: Die Daten dieser Analyse finden sich unter; Code/Syntax: No evidence that strict educational tracking improves student performance through classroom homogeneity: A critical reanalysis of Esser and Seuring (2020)
Doi: https://doi.org/10.7802/2368

Supplemental Material: The online version of this article offers supplementary material (https://doi.org/10.1515/zfsoz-2022-0001).

References

Angrist, J.D., 2014: The Perils of Peer Effects. Labour Economics 30: 98–108.

Bach, M. & M. Fischer, 2020: Understanding the Response to High-Stakes Incentives in Primary Education. ZEW – Leibniz Centre for European Economic Research Discussion Paper (20–066).

Betts, J.R., 2011: The Economics of Tracking in Education. Pp. 341–381 in: E.A. Hanushek, S. Machin & L. Woessmann (Eds.), Handbook of the Economics of Education Vol. 3. Amsterdam/San Diego: Elsevier/North-Holland.

Bol, T. & H.G. Van de Werfhorst, 2013: Educational Systems and the Trade-Off between Labor Market Allocation and Equality of Educational Opportunity. Comparative Education Review 57(2): 285–308.

Brunello, G. & D. Checchi, 2007: Does School Tracking Affect Equality of Opportunity? New International Evidence. Economic Policy 22(52): 782–861.

Elff, M., J.P. Heisig, M. Schaeffer & S. Shikano, 2021: Multilevel Analysis with Few Clusters. Improving Likelihood-Based Methods to Provide Unbiased Estimates and Accurate Inference. British Journal of Political Science 51(1): 412–426.

Esser, H., 2016: The Model of Ability Tracking – Theoretical Expectations and Empirical Findings on How Educational Systems Impact on Educational Success and Inequality. Pp. 25–42 in: H.-P. Blossfeld, S. Buchholz, J. Skopek & M. Triventi (Eds.), Models of Secondary Education and Social Inequality. An International Comparison. 2nd Edition. Cheltenham: Edward Elgar.

Esser, H. & K. Hoenig, 2018: Leistungsgerechtigkeit und Bildungsungleichheit. KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie 70(3): 419–447.

Esser, H. & J. Seuring, 2020: Kognitive Homogenisierung, schulische Leistungen und soziale Bildungsungleichheit. Zeitschrift für Soziologie 49(5–6): 277–301.

Guyon, N., E. Maurin & S. McNally, 2012: The Effect of Tracking Students by Ability into Different Schools. A Natural Experiment. Journal of Human Resources 47(3): 684–721.

Hanushek, E.A. & L. Wößmann, 2006: Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence across Countries. The Economic Journal 116(510): C63–76.

Heisig, J.P., B. Elbers & H. Solga, 2020: Cross-National Differences in Social Background Effects on Educational Attainment and Achievement: Absolute vs. Relative Inequalities and the Role of Education Systems. Compare: A Journal of Comparative and International Education 50(2): 165–184.

Heisig, J.P. & M. Schaeffer, 2019: Why You Should Always Include a Random Slope for the Lower-Level Variable Involved in a Cross-Level Interaction. European Sociological Review 35(2): 258–279.

Horn, D., 2009: Age of Selection Counts: A Cross-Country Analysis of Educational Institutions. Educational Research and Evaluation 15(4): 343–366.

Kam, C. & R.J. Franzese, 2007: Modeling and Interpreting Interactive Hypotheses in Regression Analysis. Ann Arbor: University of Michigan Press.

Kerr, S.P., T. Pekkarinen & R. Uusitalo, 2013: School Tracking and Development of Cognitive Skills. Journal of Labor Economics 31(3): 577–602.

Marks, G.N., 2005: Cross-National Differences and Accounting for Social Class Inequalities in Education. International Sociology 20(4): 483–505.

Matthewes, S.H., 2021: Better Together? Heterogeneous Effects of Tracking on Student Achievement. The Economic Journal 131(635): 1269–1307.

Pfeffer, F.T., 2015: Equality and Quality in Education. A Comparative Study of 19 Countries. Social Science Research 51: 350–368.

Roller, M. & D. Steinberg, 2020: The Distributional Effects of Early School Stratification-Non-Parametric Evidence from Germany. European Economic Review 125:103422.

Ruhose, J. & G. Schwerdt, 2016: Does Early Educational Tracking Increase Migrant-Native Achievement Gaps? Differences-in-Differences Evidence across Countries. Economics of Education Review 52: 134–154.

Schütz, G., H.W. Ursprung & L. Wößmann, 2008: Education Policy and Equality of Opportunity. Kyklos 61(2): 279–308.

Werfhorst, H.G. Van de, 2019: Early Tracking and Social Inequality in Educational Attainment: Educational Reforms in 21 European Countries. American Journal of Education 126(1): 65–99.

Authors

Jan Paul Heisig
Wissenschaftszentrum Berlin für Sozialforschung
Reichpietschufer 50
10785 Berlin
E-Mail: Jan.Heisig@wzb.eu

Jan Paul Heisig, geb. 1980 in Bremen. Studium der Soziologie, Philosophie und Volkswirtschaftslehre in Berlin, München und Stanford. Promotion an der Freien Universität Berlin (2013). Von 2007–2018 wissenschaftlicher Mitarbeiter und seit 2019 Leiter der Forschungsgruppe "Gesundheit und soziale Ungleichheit" am Wissenschaftszentrum für Sozialforschung Berlin. Seit 2021 zudem Professor für Soziologie an der Freien Universität Berlin. Forschungsschwerpunkte: Gesundheit, soziale Ungleichheit, Bildungs- und Arbeitsmarktsoziologie, quantitative Methoden. Wichtige Publikationen: Late-career Risks in Changing Welfare States, Amsterdam, 2015. The Costs of Simplicity: Why Multilevel Models May Benefit from Accounting for Cross-Cluster Differences in the Effects of Controls (mit M. Schaeffer & J. Giesecke), American Sociological Review 82, 2017: 796–827.

Sönke Hendrik Matthewes
Universität Potsdam
Prof. f. VWL, insb. Wirtschaftspolitik
August-Bebel-Str. 89
14482 Potsdam
E-Mail: Soenke.matthewes@wzb.eu

Sönke Hendrik Matthewes, geb. 1991 in Hamburg. Studium der Volkswirtschaftslehre und Politikwissenschaft in Amsterdam und Barcelona. Promotionsstudium im Berlin Doctoral Program of

Economics and Management Science seit 2015. Von 2017–2021 wissenschaftlicher Mitarbeiter am Wissenschaftszentrum für Sozialforschung Berlin (WZB). Seit 2020 affiliiert mit dem Centre for Vocational Education Research an der London School of Economics and Political Science. Seit 2021 wissenschaftlicher Mitarbeiter an der Universität Potsdam und zudem Gastforscher am WZB.

Forschungsschwerpunkte: Bildungsökonomik, Arbeitsmarktökonomik, Angewandte Ökonometrie
Wichtigste Publikation: Better Together? Heterogeneous Effects of Ability Tracking, The Economic Journal 131(635), 2021: 1269–1307.