Performance in Secondary School in German States –
A Longitudinal Three-Level Approach

Jan Skopek[*] and Jaap Dronkers[§]

* European University Institute
Via delle Fontanelle 10
50014 San Domenico di Fiesole, Italy
Email: jan.skopek@eui.eu

§ Maastricht University, ROA
the Netherlands
Email:
j.dronkers@maastrichtuniversity.nl

(Working Paper, August 2015)

Abstract

This paper contributes to the ongoing debate on how educational systems impact on academic performance in secondary schools by studying the impact of tracking on achievement in secondary education in Germany. We exploit federal heterogeneity in the 16 German states' educational systems by classifying them into three tracking regimes. Using recent, representative, and longitudinal from the NEPS, our study overcomes three severe methodological drawbacks of previous research on the impact of tracking and educational differentiation on level and (social) inequality in achievement: (1) the exclusion of the mediating level of schools, (2) the reliance on cross-sectional data, (3) the failure to account for students' prior ability especially before tracking occurred. Our findings based on a three-level model incorporating states, schools and students highlight the importance of accounting for the mediating role of schools when analyzing effects of educational systems but also the importance of including prior abilities in the study of secondary school performance.

Key words: educational systems; inequality in educational achievement; educational differentiation; ability tracking; school effects; multilevel modeling

Acknowledgements

We thank Hartmut Esser, Walther Müller, and Hans-Peter Blossfeld for intensive discussions and helpful comments on earlier versions of the paper. In addition, we are grateful for valuable comments from a group of researchers at the European University Institute. Jan Skopek acknowledges financial support from the European Research Council (eduLIFE project).

Introduction and background

Several studies on equity and efficiency of educational systems concluded that early stratification and sorting of students to different tracks of secondary schools tend to increase inequalities in academic achievement among students while not improving – maybe even reducing – overall achievement levels (Hanushek & Woessman, 2006). Evidence was found that the negative effects of early tracking vary across the performance distribution, particularly harming low performing students. In addition, it was found that tracking or other forms of educational differentiation as compared with comprehensive systems is associated with increased socioeconomic inequalities in academic achievement and educational attainment (Becker & Schubert, 2006; Bol, Witschge, Van de Werfhorst, & Dronkers, 2014; Brunello & Checchi, 2007; Gamoran & Mare, 1989; Marks, 2005; Van de Werfhorst & Mijs, 2010). Recently, this was labeled as the "standard result" on the effects of tracking (Esser & Relikowski, 2015; see also for a comprehensive overview to related studies). Most of that research builds upon a two level approach characterized by comparing countries and students based on international assessment data. Thus, it is addressing only the total effects of tracking without disentangling various mechanisms that contribute to these outcomes and without considering potential heterogeneity within a country's educational system. Furthermore, these studies rely on rather rigid, mostly dichotomous definitions of tracking versus non tracking.

More recent research tries to incorporate more refined measures of tracking and ability sorting and the school level in order to obtain better insights to the question of how system effects actually operate (Bol et al., 2014; Dronkers, Van Der Velden, & Dunne, 2012; Dunne, 2010; Esser & Relikowski, 2015). Importantly, by re-analyzing PISA data Bol et al. (2014) found that in countries where central examinations are prevalent in the secondary school system, the relationship between the level of tracking and the level of inequality by socioeconomic status (SES) is attenuated. Additionally, taking into account school level characteristics (like student-teacher ratio or school's SES composition) revealed that school heterogeneity seem to mediate parts of the impact of tracking on the link between SES and students' achievement. As a main finding of Dunne (2010) – also relying on PISA data – the impact of educational differentiation at the country level is mediated by effects of social composition of schools. Particularly, she found that compositional effects are stronger in countries that promote educational differentiation while lower SES students benefit more from a higher SES composition of school. Dronkers et al. (2012) highlight the importance of considering the track- and school-level as separate units of the analysis. Their findings suggest that the effects of educational system characteristics are flawed when adhering to a two level model of countries and students while ignoring relevant track- and school-level features (like type of school, entrance selectivity, social and ethnic diversity). Moreover, the inclusion of the track-level is necessary to avoid overestimation of school SES composition effects, especially in stratified educational systems. It turns out, that channels through which institutional settings on the country level impact on student achievement are manifold and complex.

Past research investigating the impact of tracking and differentiation on secondary school achievement and (social) inequality in achievement suffer from three methodological drawbacks: (1) the exclusion of the mediating level of schools, (2) the reliance on cross-sectional data (mostly PISA data measuring achievement at a fixed age), and, maybe most importantly, (3) the failure to account for students' prior ability especially before tracking occurred. While as discussed above (1) has been partly addressed, (3) is a black box in previous studies and should be addressed in further research.[1] If we do not know prior abilities of students (at end of primary school), then we will inevitably overestimate the effect of student's socioeconomic background (individual SES) during secondary schooling due to the primary relation between SES and early abilities.[2] As an extreme, differential achievement in secondary school could just resemble differential achievement in primary school (system effects would operate at most indirectly for instance through anticipation mechanisms). Second, it is impossible to assess effects of tracking on achievement without knowledge of prior abilities that usually serve as an important basis for track allocation in differentiated systems (ability sorting). Even in tracked systems, different school compositions might be produced if mechanisms of allocation of students to secondary school are different as a result of institutional sorting rules that promote or constrain families' free choices. Third, without including prior abilities in empirical models, SES effects on the school level as found earlier are likely to be heavily confounded with ability effects on the school level (effects of the intellectual composition of the student body, peer effects). If we are unable to distinguish the latter, we cannot tell whether schools' SES segregation is effectively promoting inequality of educational opportunity or whether it is rather the segregation by abilities (the explicit idea of ability sorting) that produces the inequality. Educational systems with various degrees of

[1] Some studies like Hanushek & Woessman (2006) mimic a country level longitudinal design by comparing two cross-sectional datasets on early and later abilities for each country relying on a Difference-in-Difference approach. Nonetheless, this design is not without methodological problems (for a review see Van de Werfhorst & Mijs, 2010) and particularly is blind for disentangling direct and indirect system effects.

[2] Nonetheless, we one can assess the total and cumulative effect of SES on student achievement up to a certain age.

differentiation will produce different school structures in terms of SES and abilities composition with important consequences on individual learning. Moreover, even in differentiated systems the strictness of ability sorting as basis of allocation to different tracks of school will impact on compositional features of the schools (e.g., stronger homogenization) and, hence, on the resulting effects on individual development.

Some recent studies try to tackle some of these issues. Dronkers (2015) tried to solve the problem of the missing measures of early abilities by analyzing longitudinal data on students entering secondary school in 1989 in the Netherlands, which possesses a highly differentiated and selective system based on prior achievement and teacher recommendations, which is in many aspects similar to that of the German states. After including prior abilities on the individual and school level (average of student's abilities) as well as type of school, he found that both individual and school SES (average of students' SES) are no longer related with language score in third year of secondary school whereas school type has substantial effects. Dronkers concluded that school characteristics like entrance selectivity based on scholastic ability at the end of primary education affect the link between SES and early scholastic ability on the one hand and later educational achievement in secondary schools at the other hand. Moreover, he concluded that school characteristics seem to mediate some of the effects of the educational system. In several works Hartmut Esser (Esser & Relikowski, 2015; Esser, forthcoming-a, forthcoming-b) specified a theoretical model providing a systematic account of the pathways through which systems, families and schools produce equity and efficiency outcomes in educational achievement. Asking whether ability tracking is responsible for inequalities in achievement outcomes, Esser & Relikowski (2015) tested the implications of Esser's 'model of ability sorting' (MoAbiT) using longitudinal data on Hessia and Bavaria, two German federal states that constitute pretty antagonistic cases in terms of ability sorting and differentiation. Following up students from Grade 5 to 7 they found evidence challenging the "standard" approach; once all relevant conditions and processes (in particular ability and SES on the school level) are empirically controlled, there are no direct effects reinforcing inequality in terms of achievement due to a stricter differentiation; all system effects are explained by individual and school level factors. In fact, Esser & Relikowski (2015) report an additional positive effect of homogenization of schools by cognitive abilities that is even higher for Bavaria (stronger differentiation and strict ability sorting) as it is assumed by proponents of ability tracking.

Our paper contributes to the ongoing debate on how educational systems impact on academic performance in secondary education by studying the German case using recent, representative, and longitudinal data on a cohort of five graders from the German National Educational Panel Study (NEPS). Contrary to most prior studies on secondary achievement treating Germany as a unified system, we will present a much more refined model on studying systems of secondary education and their impact on the development of individual achievement. Exceeding the horizons of previous studies, we are (a) employing a larger national sample on Germany, (b) exploiting variation across a larger number of educational systems (16 German states), (c) adding prior abilities to the equation, and (d) applying a three level model, deciphering various influences on individual achievement on the level of states, schools, and students. The federal setup of education in Germany is very advantageous for the purpose of our study; while in general there is a common principle of early differentiation of students to a (mostly) tripartite system, there is a vast variety in types of schools, the availability of comprehensive types of schools, the entrance selectivity of schools, the tension between achievement based track selection versus freedom of choice, and also variation in the age of tracking. Consistent with prior research on Germany (Dollmann, 2011; Gresch, Baumert, & Maaz, 2010; Neugebauer, 2010; Roth & Siegert, 2015; von Below, 2006), we will distinguish German states by their strictness of tracking. Nonetheless, such rough classifications might mask important heterogeneity in the admission practice of schools between and within states. Legal frames of educational systems will be manifest in certain common standards among schools within a system, of course. Though, German schools always have autonomy to some degree and will differ in their everyday practice and routine of admitting students. Hence, contrary to previous studies conceiving ability tracking as a macro level feature, we will incorporate additionally measures for entrance selectivity on the level of schools.

Educational systems of German states

A peculiar and sometimes complicating feature of the German educational system is that general education is not centrally organized but historically a federal state affair (for a comprehensive overview, see Maaz, Baumert, Gresch, & McElvany, 2010). In a more stylized fashion, one could even argue that there are 16 educational systems (one for each federal state) rather than one unified system. Table 1 provides a classification of German states by their tracking regime, which we will elaborate in the following.

=== TABLE 1 here ===

Our discussion will be informed by a useful typology of German educational systems as suggested by von Below (2006). She differentiates German educational systems by the degree of regulation regarding *structure* (organization of the school system; e.g. number of tracks, provision of comprehensive schools) and *content* (curricula content and its observance by tests and examinations). Regulations might be rather *tight or strict* (systems provide clear-cut and obligatory norms and rules) or rather *loose* (system provides frames and suggestions, room for interpretation at the individual level). In her terminology, von Below called a loose regulations of structure *reformed* and a tight regulation of structure *traditional*; loose regulations of content she calls *liberal* and tight regulations of content *conservative*. Based on that, von Below develops four sub-types of educational systems in Germany: *Traditional-conservative* (structures and content are strictly regulated, e.g. tripartite track system, early selection based on objective achievement, track mobility difficult and rare, centralized curriculum standard), *reformed-conservative* (strict regulation of content and importance of achievement; structures are loose, e.g. a considerable number of comprehensive schools, track switching easier), *traditional-liberal* (strict structures, but contents not obligatory regulated and oriented more on the individual personality, lower importance of objective achievement), and finally *reformed-liberal* (free school choice, larger provision of comprehensive schools, content determined and designed on the individual level). Prominent examples for traditional-conservative are the southern German states (see Table 1) while for instance the pattern reformed-liberal can by typically found for western states or city states.

In practice, German education is characterized by a remarkable regional heterogeneity not only in the provision of certain types of secondary schools[3] but also in the criteria of access to higher track schools. Nonetheless, there are also important communalities. Across all states track states in to different types of schools usually after Grade 4 at age of 10, although in some states later. Berlin, Brandenburg and Mecklenburg Western Pomerania as they lift the age at tracking to around 12 years by keeping their students longer in comprehensive schooling. In Berlin and Brandenburg primary school lasts six years and students enter secondary education in Grade 7. Mecklenburg Western Pomerania has four

[3] An example is the availability, coverage and implementation of comprehensive types of schools which are not uniformly present in Germany. Similarly, the Hauptschule as separate school form is also not present in all German states.

years of primary school but keeps students two additional years in fully comprehensive schools ('orientation stage'). In total, those three states provide the shortest length of tracking and, thus, can be described by a lower degree of differentiation as compared to other states.

After tracking all states channel students into up to three different educational tracks, which is Hauptschule (vocational), Realschule (technical), or Gymnasium (academic). In the more traditional regimes these three tracks coincide with school types. Nonetheless, some states combined the lower tracks (the more 'reformed' ones). Yet, comprehensive types of schools are expanding since their introduction in West-Germany since 1970ies (socialist East-Germany had a rather comprehensive schooling system). Nowadays, in all German states more inclusive forms of education by maintaining comprehensive schools combining several tracks in mixed or separated classrooms with manifold labels are available at least to some extent.

A common element across German states is the primary school recommendation (Gresch et al., 2010; Neugebauer, 2010). At end of primary school teachers provide a formal recommendation for students indicating the secondary school track being most apt given their abilities (usually marks in Math and German), behavior, and talent. Specifically, the recommendation indicates the eligibility for the academic track. The regulating power of the recommendation and thereby its consequences, however, is different across states.[4] Almost all of the 'conservative' states (regulating content and its observance strictly) have legally binding recommendations restricting families' voice in track choice by knotting access to the higher tracks with prior achievement.[5] In states with loose content regulation ('liberals'), the recommendation is nothing more as a suggestion, without any binding element; parents have freedom of choice.

Summarizing our discussion we arrive at three types of tracking regimes that are relevant for our study of early secondary school achievement (see last column in Table 1). First, there are states that sort students early after Grade 4 to tracks based on prior performance via a binding recommendation (early ability tracking states 'EAT'). To this group belong the larger southern states Baden-Wuerttemberg and Bavaria, Saarland, and the new *Laender* Saxony, Saxony-Anhalt, and Thuringia. A second group of states sort students

[4] The binding character of the teacher recommendation is also a politically very disputed subject. Hence, in some states amendments of school laws introduced and withdrew the bindingness depending on the political color of the current government.

[5] Circumventing a binding teacher recommendation implies obstacles and burdens. In case of a conflict between parents and school, parents can choose a legal process or, as in some states possible, can make use of trial periods or entry examinations of the targeted secondary school.

early after Grade 4, but without a binding recommendation (early tracking states 'ET'). Finally, there are three states that post-pone tracking to Grade 7 by providing comprehensive schooling up to Grade 6 (prolonged comprehensive 'PC'). Except for Mecklenburg-Western Pomerania which has been labeled as being rather conservative, those states share a liberal orientation.

Hypotheses

In accordance with Esser's model of ability sorting (Esser & Relikowski, 2015; Esser, n.d.-a, n.d.-b), one can expect secondary school performance to be an complex outcome of various mechanisms. We will elaborate a set of hypotheses to test those being most relevant for overall achievement and inequality.

First of all, abilities shape the intellectual capacity of knowledge acquisition. Consequently, and very straight-forward, we expect a strong effect of earlier achievement on later achievement in secondary school (H1). Families differ in their resources (like quality of home learning environments) providing differential constrains and opportunities for children's learning process. Hence, over and above prior achievement we expect socio-economic status (SES) of parents to be positively related with secondary school achievement (H2).

Beyond these individual level factors, achievement will vary by different environments in schools and classes. There are several reasons to expect why schools should matter. First, in tracked system like Germany, different types of school embody different curricula with different learning plans and, thus, different opportunities for acquiring specific knowledge in form of learning milieus (Neumann et al., 2007). In addition, higher track schools – on average – might attract more motivated and higher qualified teachers increasing effectivity and efficiency of students' learning. Therefore, controlled for other factors, we expect to find a significant influence of track, particularly a premium for Gymnasium, on achievement in secondary school (H3).

Above track specific curricular effects, contextual features of schools might directly impact on learning of children basically via quality of instruction, resources, organization, and composition of the student body. As reported, several previous work based on PISA data has found evidence for effects of school's SES composition on individual performance. Rumberger & Palardy (2005) found SES composition effects to be almost as large sometimes even larger than individual SES effects on achievement growth. Consequently, SES segregation across schools could be an important driving force of social inequalities. A higher

SES composition – over and above track curriculum – might contribute to more favorable conditions (e.g., better teachers, more expensive equipment, or stronger involvement of parents at the school level, a better academic climate) at the school level and, hence, facilitate beneficial learning environments. Hence, we expect a positive effect of SES composition of school on individual achievement (H4). Furthermore, the intellectual quality of peers should matter for individual achievement. If classrooms are visited by more able students, teaching may be more efficient. Reversely, higher ability schools and classes may attract more competent teachers. Moreover, exposure to more able peers might not only provide higher motivation and incentives to the individual learner but also support in school-based friendship networks. For instance, Kerckhoff (1986) analyzing effects of ability grouping in British secondary schools found support for this 'divergence' hypothesis that is that (accounting for prior abilities) students in higher ability classes gain more. Therefore, school's intellectual composition should positively affect individual achievement (H5).

In all educational systems, residential social segregation will inevitably lead to a minimum level of social and – due to the primary relation of abilities and parental background (Boudon, 1974) – to ability segregation of schools. Notwithstanding, in a differentiated system like Germany that explicitly allocates students to different track schools, schools segregation in terms of SES and ability is likely to be much more pronounced. Thus, we will expect that the (within school) effects of SES and ability will be smaller (compared to the total effects) once accounting for the school level, particularly school composition effects resulting from the interplay of ability and preferences at the transition to secondary education (H6).

The interplay of conditions of the school level and individual characteristics is likely to impinge on achievement. Theoretically, such feedback mechanisms can occur on both levels SES and ability. Regarding SES, once could argue that lower SES students will do harder in high SES school environments resulting from a lack of cultural and behavioral adaption to the class and school standard. That should become visible in a positive interaction between a schools SES composition and individual SES background (H7).

With regard to ability, interaction effects may exist if peer effects work in a non-linear fashion (i.e., depend on own ability). For instance, the aforementioned benefit of having better peers in the class may be larger for low performing students whereas peer returns become marginal for students having high ability as a result of ceiling effects; a basic tenet underlying negative views on ability tracking favoring comprehensive and more inclusive schooling (cf.

Gamoran & Mare, 1989). If that is the case, a negative interaction term between individual-level and school-level ability should show up (H8a). However, one could also expect the opposite (positive interaction), that better peers are less beneficial for lower performing students in terms of achievement (H8b). This could happen for various reasons. For instance, more able students on average allow the teacher to speed up the pace of instruction and progress faster in the lectures, which might leave less able students behind. Moreover, everyday exposure to more able fellows could promote learning frustration and eventually isolation in the school context cutting important support networks. As a third hypothetical case, peer ability effects might be just linear, that is, working for less and more able students in the same way resulting in no interaction (H8c). Yet, that does not postulate that there are no feedback mechanisms at work. We might find linear effects also if mechanisms behind H8a and H8b mix and cancel each other.

After having hypothesized factors of the core model – school and individual level factors – of secondary school performance, we will next turn to effects on the system level in the context of German states. Previous cross-national comparative research found, that standardization (for instance central examinations, standardized curricula but also a more objective sorting to school types based on abilities) reduces the magnifying effects of tracking on inequality in educational outcomes (Ayalon & Gamoran, 2000; Van de Werfhorst & Mijs, 2010). Yet, looking at German states it is not a prior clear, how more or less strict ability tracking impacts on overall achievement and inequality as various other institutional features might intervene.

Importantly, one has to consider first the ability influx to different states' secondary schools. If track placement is done based on achievement, parents as well as schools have to invest more in preparation of children anticipating that achievement crucially matters for the transition to secondary education. If that is true, prior abilities should be higher on average in such states where the teacher's voice is binding. Consequently, these higher prior abilities should translate into higher achievement in secondary school on the state level (H9a). That should be mirrored at the school level: Students visiting schools that select their intake stricter based on prior performance should perform better as a result of their higher abilities before (H9b).

Differentiated systems (like Austria, Switzerland, Germany or the Netherlands) that explicitly track students to different school types create a stronger homogenization of schools in terms of SES as compared to more comprehensive systems (Maaz, Trautwein, Lüdtke, &

Baumert, 2008). Nonetheless, even within differentiated systems one can expect different school compositions according to institutional rules of tracking. The more stringent sorting of students to schools is based on prior achievement, the stronger schools will be stratified by performance resulting in a higher between-school heterogeneity and a higher within-school homogeneity of abilities. If there are no peer effects, homogenization would be inconsequential and nothing would be different (ceteris paribus). Nonetheless, if peer effects work and peers effects should be stronger if peers are more similar as a result of ability homogenization within schools. Hence, we would expect to see stronger effects of ability composition in states with a binding recommendation compared to states where track choice is up to the parents (H10). Consequences of the homogenization could be quite different depending on the nature of peer effects; linear peer effects (according to H8c) will facilitate overall dispersion in achievement between schools and tracks. That will be further exaggerated, if peer effects are non-linear and stronger for more able students (see H8b). Nonetheless, compensation could happen if peer effects are stronger for less able students (see H8a).

Eventually, one can speculate that differences in institutional sorting among German states may have an impact on social inequalities in secondary school achievement outcomes. Stronger achievement orientation might provide better structures facilitating opportunities and incentives for learning particularly for less performing students. In effect, social disparity in secondary school achievement could be lower within German states that restrict free choice and allocate students to secondary schools based on abilities. On the other hand, one could also expect the opposite: there might be higher social disparities in school performance in states that track by ability, if the link between SES and ability is sufficiently strong and the more able students and higher SES students are beneficially segregated from the less able and lower SES students. Yet, the various mechanisms discussed are likely to intervene. That makes any a priori predictions on system effects in total very challenging. A superficial inspection of overall differences across states might mask important heterogeneity in the way how different systems work. Hence, in the following our task shall be to test our three-level model empirically.

Data

Our analysis draws on recent data from Starting Cohort 3 of the National Educational Panel Study (Blossfeld, Roßbach, & von Maurice, 2011).[6] This cohort comprises a multi-stage and stratified school-student sample of 6.112 students of the population of students attending the 5th grade secondary school in Germany in autumn 2010 (Aßmann, Walter, & Zinn, 2012; Skopek, Pink, & Bela, 2012). The sample is followed up annually and to the time of conducting this study, two waves were already available. In the first wave in autumn 2010, students' competences have been assessed within the schools under supervision.[7] Hence, assessments took place relatively early in secondary school, 1 to 4 months after first day of school year.[8] Moreover, school principals and teachers were surveyed as well as parents (the latter by CATI). On year later in Grade 6, a second wave assessment took place.[9] Results from the second test represent our dependent variable. Even if our study's scope is obviously limited to achievement up to one and half year after secondary school enrolment, the NEPS data provides important benefits for our purpose.[10]

We applied some few but necessary sample exclusions. We dropped students from an additional sample of special needs schools.[11] Moreover, some few students who participated neither in wave 1 nor 2 tests as well as students with missing data on state of school, date of birth and gender were dropped, too. Finally, a number of 5,444 students attending 230 schools in 16 German states remain for the analysis.

Variables

Individual level

[6] This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 3 – 5th grade (From Lower to Upper Secondary School), doi:10.5157/NEPS:SC3:2.0.0. The NEPS data collection is part of the Framework Programme for the Promotion of Empirical Educational Research, funded by the German Federal Ministry of Education and Research and supported by the Federal States.

[7] Field time took place during October 2010 to January 2011. About 50% of the field work was done beginning of December and about 98% end of December.

[8] The effective start of the school year varies slightly across federal states. However, in 2010, in most states summer holidays ended between mid of August and mid of September.

[9] The time span between first and second tests counted on average 12 months with a minimum of 12 and a maximum of 14 months.

[10] Our analyses will be replicated once further wave updates are released by the NEPS.

[11] Competence data is not available for these students.

Dependent variable of our analysis is achievement in second year of secondary school (Grade 6). We measure that by test scores in Science literacy.[12] The framework of the NEPS test on science embraces both knowledge of basic scientific concepts and facts ('Knowledge of Science') as well as the understanding of scientific processes ('Knowledge about Science') and implements relevant features of the *PISA 2006 framework of scientific literacy*, the *Benchmarks of Science Literacy of the American Association for the Advancement of Science*, and the *German National Educational Standards for graduation after intermediate secondary schooling* (Hahn et al., 2013).

The data provided rich variables to measure our independent concepts on several levels of explanation. *Socio-economic background (SES)* of student's was measured by the average among parents' education classified on a 4 point scale ranging from low (lower secondary or less) to high education (tertiary training).[13] For that, we took reports from the responding parent in the phone interview that took place in the first wave.[14]As proxy for *prior abilities (ABIL)* we took standardized WLE scores in Mathematics which has been assessed in the first wave (Duchhardt & Gerdes, 2012).[15] Even if Math was assessed within secondary school (as a constraint of the panel study's design), it was taken relatively early in Grade 5 were differences due to track exposure can be assumed to be still very small. Taking Math as proxy for prior abilities is particularly favorable in the German context as Math competence constitutes a crucial element of the primary school recommendation and higher track entrance hurdles (usually marks in Math and German are important) shaping students' sorting. Moreover, Math score turned out to correlate high with Science literacy (about 0.66). As control variables we included dummy indicators for *migration background (Migrant)*[16] and student's gender (*Female*). All missing data on the individual level have been imputed by

[12] Second wave assessment also included tests on ICT literacy and listening comprehension on the word level (vocabulary). We decided for science because of its particular importance for educational pathways and because it matches with major reports based on PISA data.

[13] Alternatively to parental education we tried also highest ISEI-08 among parents as measure for SES (which many other studies pursue). Results were very close. Nevertheless, our study requires measures as precise as possible in order to effectively disentangle influences of prior abilities and social background. Eventually, we opted for average of parental education instead of ISEI as the former performed clearly better in predicting test scores.

[14] If data on parental education was missing in the first interview, we filled up with values from the second wave interview.

[15] Alternatively, we took the standardized Reading score in Grade 5 (Pohl, Haberkorn, Hardt, & Wiegand, 2012). We also tried a composite measure of Math and Reading. Results were largely similar (available upon request). From a substantial point of view, for keeping uncertainty induced by missing values lowest, and lastly for simplicity of presentation, we decided for Math.

[16] We define migration background by a variable on generational status provided by the NEPS (Olczyk, Will, & Kristen, 2014). Our definition includes status 1-2.25, thus ranging students no born in Germany to students with one parent born in Germany (but with grandparents born abroad).

multiple imputation generating 20 imputation datasets.[17] Afterwards, we standardized competence scores and the SES variable for our sample of students to have mean 0 and standard deviation 1within each imputation dataset.

School level

On the school level, we included *social composition (S-SES)* and *ability composition (S-ABIL)* by averaging individual level variables within schools.[18] In addition, we include dummies to identify track-types of secondary school. One should note that Gymnasium is the only school type that is uniformly provided in all German states. For that reason, we took a binary indicator for *Gymnasium (GYM)* versus other types.[19]

For measuring entrance *selectivity of schools*, as a measure of state level features on the intermediate level of schools, we exploit items from the school principal questionnaire of the first wave. School principals have been asked on the relevance of several factors in admission of students to their schools.[20] Among others, factors relating to prior achievement were: (1) previous students' school achievement or grades, (2) entrance examination, (3) trial lesson, (4) recommendation of students' previously attended school. For each, principals respond on a four-point scale ranging from *not considered at all*, *low relevance*, *high relevance*, and *necessary precondition*.

For reducing the dimensionality of these items, we used principal component analysis (PCA) to extract components relating to *selection on achievement.* In doing so, we first pooled schools with responding principals in Wave 1 of Starting Cohort 3 with Wave 1 of Starting Cohort 4 (a bigger sample of ninth graders based on the same sampling frame). As a result we were able to exploit information from 445 different schools in total (due to the common sampling frame 122 of these schools were included in both cohorts whereas their principals were surveyed only once). Second, we applied multiple imputation procedures to impute missing value in these items (about 12 percent missingness) using an array of school

---

[17] We used a chained equation approach. The MI model included a large variety of individual level factors and accounted for school factors by fixed effects for schools.

[18] This has been done within imputation datasets. Modelling composition effects makes clear the advantage of multiple imputation since it may effectively reduce the risk of a two-level bias. Nonetheless, running our analyses on complete data yielded substantially very similar results.

[19] Nonetheless, we tried alternative specifications using more precise dummies for Hauptschule (HS), Realschule (RS), comprehensive school types (CS), and Gymnasium (GYM). Results were virtually the same.

[20] The precise question was "How do you weigh the following factors when admitting students to your school?".

characteristics.[21] Third, from a set of five generated imputation data sets we took the mean imputed value for missing data in the admission standard variables. Fourth, we run the PCA on the covariance matrix of the 4 items (see Table A1 for results). We extracted one component having an Eigen value greater than one (1.98). Overall the extracted component represents about 62 percent of the four items' covariance.[22]

Our constructed measure *selection on achievement (SOA)* represents the z-standardized score (weighted on school level) of the achievement component from the PCA analyses. Yet, this score is only available for schools with principals who were responding to the NEPS questionnaires. In the analytical sample 74 out of 230 schools had non-responding principals. Dropping those schools did not substantially change the coefficients but resulted in higher standard errors due to severely lowered case numbers on the student level (3,795 instead of 5,444 students). Hence, we opted for a missing coding in the school's score by defaulting to 0 (average over schools) in case of a missing value while controlling for non-response of the school's principal in the multivariate models.

State level – tracking regime

On the state level we constructed a categorical variable for distinguishing tracking regimes EAT, ET, and PC states according to our classification in Table 1. On substantial grounds, we decided to code the German *Land* on the basis of the location of the school. Finally, out of 16 German *Laender* 6 are classified EAT, 7 as ET, and 3 as PC.

Analytical Strategy

Will test the theoretical arguments discussed by including the school level. For that, we relied on linear mixed models to statistically specify our three-level approach. Random intercepts for schools take into account unobserved heterogeneity on the school level. To capture variance on the level of states we experimented with several specifications and finally decided for a fixed effects approach including dummies on track level and, as a robustness check,

[21] We used chained equation models for iteratively imputing values for the admission practice items. The model additionally included information on school size (number of students, teachers, and classes), school type, social composition of school (fraction of migrants and social strata), size of community, teachers' attitudes and moral, and private/public funding body.

[22] For robustness checks we did the PCA only for on the subset of 125 schools from the 5th graders cohort (Starting Cohort 3). Compared to the full sample of schools component loadings very similar and standardized scores were almost perfectly correlated (.998). Hence, incorporating all information available we decided to construct the component scores based on the full sample of schools.

dummies for states.[23] To check if our results are robust against non-normality of residual school effects we estimated models with fixed effects for schools (pure within school models that account for all time-constant school heterogeneity, not reported), too. Nonetheless, fixing states data on schools become partly very sparse as a result of the limited sample size; there are some states with less than a handful schools. Yet, our conclusions did not change by the state fixed effects models, rather, estimates were partly even more pronounced. All our estimations were done using design weights. For the linear mixed models we specified design weights for school (school level weights) and student weights accounting for unequal selection probabilities conditional on school (student level weights).

Descriptive results

Mean performance

Our empirical investigation starts with a descriptive overview. Table 2 reports means, standard deviations and fractions of central variables grouped by tracking regime. At first, we observe a clear gradient in Science sores in Grade 6 between EAT and ET. On average students in EAT states score about 26 percent of a standard deviation higher ($p < .01$) and PC students about 18 percent higher ($p <.01$) compared to students in ET states (difference between EAT and PC about 8 percent, $p=.183$). This pattern is pretty good resembled by the gradient of tracking regimes in prior Math abilities (mean differences statistically significant on $p<.01$ with exception of the contrast between ET and PC). While overall performance differs remarkably, EAT and ET states hardly differ by SES ($p=.125$). Nonetheless, students in PC countries appear to have more advantageous SES family backgrounds on average (statistically significant only compared to ET, $p<.05$), but have also the lowest fraction of migrants.

Schools

On the school level, we see that schools are most selective when it comes to the practice of admitting of students in EAT states and least selective in PC states whereas average SOA scores of schools in ET states range in between. A disaggregation of the score based on school types (see Table A2) reveals that the differences between ET and EAT are

[23] To have a more parsimonious approach, we also tried random intercepts for states. Albeit this might violate assumption of normality since there are only 14 units on the state level, comparisons with the non-parametric fixed effects approach showed only minor differences in estimations. Though, for some models and imputation datasets, convergence problems emerged due to the considerably high complexity of the resulting model and the marginal variance on the state level. Hence, we decided for the specification with fixed state effects.

particularly strong among school types Gymnasium and Realschule. This is full in line with what one could expect with the state level rules of sorting.

School level averages in prior abilities and SES differ from the individual averages as a result of a non-random distribution of students on schools. Yet, a closer look reveals that EAT school average in SES are closest to individual SES averages and most distant in individual and school ability, which might point to a lower degree of SES segregation and a higher degree of ability segregation across schools in the EAT states. Very consistently, in PC states there are throughout comprehensive schools in the sample. Fractions of school types are rather similar across EAT and ET states, with a slightly higher variability of types in the ET states.

=== TABLE 2 here ==

Inequality

Do tracking types differ in the degree of dispersion and social inequality in secondary school achievement and early abilities? Table 3 allows some tentative conclusions. Using the standard deviation and percentile differences as measures of dispersion, it appears that individual inequalities are slightly larger in EAT states. Nonetheless, differences are minor. Social inequality (here measured by the fraction of variance in achievement explained by SES) seems to the lowest in the EAT states, particularly on the level school composition variables. Very similar patterns can be found for Math abilities early in Grade 5.

=== TABLE 3 here ===

Multivariate results

Table 4 presents estimates of the linear mixed models. Model 1-4 include in a stepwise fashion tracking regime, schools' entrance selectivity, school type, as well as SES measured by parental education on the individual and the school level. As a baseline, Model 1 includes only tracking regime and random effects on the school level. Patterns in differences between tracking regimes conditional on school effects are the same as found in the descriptive results. Nonetheless, the overall contrast between PC and ET does not reach statistical significance. The variance estimates tell us that about 31 percent of the total variance in Science is located on the school level.

Model 2 includes schools' entrance selectivity (SOA) revealing a pretty strong association with Science performance. Higher selectivity of schools explain partly the edge of students in EAT states as compared to ET states, whereas this seems not to be true for the comparison PC versus ET which even increases; that is not too surprising, considering that school selectivity is extremely low in the PC states while performance is comparably high. Schools' selectivity accounts for about 14 percent of the variance between schools as estimated by Model 1.

Model 3 introduces both SES variables with interaction term. Students with more advantageous family backgrounds have higher performance in Science. In addition, students in higher SES school do also have higher performance, which represents an indirect SES effect. The model shows an interaction with positive sign, but it is rather small and far from statistical significance. Furthermore, SES differences across students and school compositions account for a good part of the PC and ET contrast and the differences by selectivity of school (visible in the smaller coefficients for system dummies SOA).

Model 4 additionally accounts for school type Gymnasium; students enrolled in the academic track perform better. Gymnasium explains also part of the school SES differences in Science score (not the individual social gradient in Science score) and part of the remaining differences by SOA. The contrast between tracking regimes ET and EAT is not affect by controlling for Gymnasium, but so the contrast between PC and ET; compared to Model 3 it is larger and statistically significant. Again this is a result of system differences; in PC there are (almost) no Gymnasium students, so the contrast effectively relates to students in non-Gymnasium schools in ET states and students in PC states (conditional on the other variables). Including the additional variables led to a sharp reduction of explained variance across schools, but only a minor explanation of the individual variance.

=== TABLE 4 here ===

To sum up, Models 1 to 4 suggest that students in states that track early without ability tracking and students in less selective schools perform weakest in Grade 6 on average. There are remarkable differences by individual SES and SES school composition. More selective schools and Gymnasiums have more favorable SES compositions associated with higher performance, but beyond that entrance selectivity of schools and particularly Gymnasium show positive effects.

The next set of models (5a to 8) adds prior abilities of students and schools ability composition to the equation. We present Model 5 in a stepwise fashion, first including individual abilities only (Model 5a), which is best compared with Model 2. Prior ability in Math is very strong predictor of Science score in Grade 6 (1 standard deviation increase in Math is associated with .59 standard deviation increase in Science). Abilities explain much of conditional differences tracking regimes and also a large part of the advantage of more selective schools. Model 5b includes ability compositions of schools and the interaction, thus, is equivalent in structure to Model 3. Peer ability effects are sizable: students with an average ability gain (loose) about .3 standard deviations in Science score, if average school ability increases (decreases) by one standard deviation of ability. Moreover, the statistically significant interaction term indicates non-linear and reinforcing peer ability effects: the higher (lower) the individual ability the larger (smaller) the difference by school ability composition. Once we account for individual and peer ability, the coefficient for school selectivity is shrinking towards zero as a result of a selection effect. Also, the contrasts between tracking regimes are rendered marginal. Obviously, differences between states and schools in Grade 6 secondary school achievement are brought out by differences in prior abilities, schools' ability compositions and peer effects.

Model 6 includes both, SES and ability variables. We see that differences by SES found in Model 3 are confounded by differences in ability. This is most true for SES composition; the coefficient reduced by almost 82 percent and is statistically significant only if we accept a 10 percent probability of committing type I error. Nevertheless, there remains an important individual-level SES gradient in Grade 6 performance.

Accounting for school type Gymnasium in Model 7 does not add much to the model; the whole effect as found in Model 4 (.37) is explained by prior abilities of students and peer ability. Model 8 is adding the controls for migration background and gender; students having a migration background perform worse and there is not much differences by gender.

Model 8 (FE) instead of specifying aggregated tracking regime types includes states as fixed effects and, hence, is more precise as the other models in accounting for residual state level heterogeneity in performance.[24] All substantial results remain robust. After all, the last model explains about 27 percent individual level and 94 percent school level variance within tracking regimes (i.e., compared with Model 1).

Referring to our hypotheses we can conclude the following. H1 (positive effect of prior ability) and H2 (positive effect of social background) found strong support. Particularly, early ability is a strong predictor for later achievement. However, when inspecting the full model, we found no support for an additional track effect (H3) – at least on average.[25] Also, we can report only weak evidence for a general SES composition effect (H4). Instead, it is the ability composition that matters (H5) and strongly confounds SES composition effects as found earlier (H6). The hypothesis of a positive interaction between SES composition and individual SES background as a result of a cultural fit or misfit of student and school (H7) was disappointed across all models; in our data, we found no sign for such feedback mechanisms. Rather we found relevant effects of peer ability benefitting all students but better students more. Hypothesis 8a of a negative interaction (low performing students benefit more from better peers) had to be rejected. Nonetheless, the positive interaction is rather weak in size and, thus, not strong enough to support opposite hypothesis 8b (better peers are detrimental for lower performing kids). Eventually, our results are most in line with hypothesis 8c (linear peer effects) but with imperfect linearity.

Our discussion theorized that in states with stricter ability tracking, prior abilities should be higher in the first place, which then translates into advantages during secondary schooling (H9a). Our findings from Model 5a strongly support that very simple argument. Comparing full Model 5b with the Model 1, we saw that the coefficient for the contrast between EAT and ET shrank by almost 87% once conditioning on prior abilities and compositional effects. Hence, by installing achievement orientation states indirectly already improve their student input to secondary schools. This is also true on the level schools (H9b), the coefficient for SOA shrinks tremendously after accounting for abilities, albeit Model 5b

[24] As we are not allowed to compare or single out German states by our data contract with the NEPS, we do not report the coefficients. What we can tell is that the standard deviation of residual state effects is about .09 as compared to .22 in a null model that defines fixed state effects and random school effects only.

[25] It should be noted that this might not be necessarily true for single states. In additional analyses (not shown) separating out states sometimes yielded positive net effects for Gymnasium schools and sometimes negative net effects. So, it is likely that various effects cancel each other out on average. Another explanation for the insignificant effect of Gymnasium might be the limited observation period of our study.

shows that the effect of school selectivity goes fully to zero only after accounting for the ability composition of schools (also the coefficient for contrast EAT versus ET shrinks further). We can conclude that states that ability-track their students to secondary schools have a better input overall which results also in more favorable school contexts. Nonetheless, even after accounting for abilities there remains still a difference of about .1 standard deviations (though statistically significant only on the 10 percent level) between ET and the group of states having prolonged comprehensive schooling. Inspecting state fixed effects models (Model 8 FE) revealed that this is particularly driven by one state.

Hypothesis H10 remains to be tested. Our core argument was that if students are sorted to tracks and schools on the basis of prior abilities, schools compositions are more homogenous in terms of ability, which could foster peer effects. We will test that on the basis of the state fixed effects model (Model 8 FE), adding interaction terms between ability and SES variables and dummies for tracking regime. We will leave out in the school level variable on entrance selectivity because it had no effect in the full model but might reduce the model efficiency due to high correlation with school ability (more selective schools have better ability compositions, as we have found out earlier).

Table 5 reports results from two additional models in a compact way. Model 10 is more parsimonious by testing differences in composition effects between EAT and other regimes; Model 11 is adds more parameters for testing coefficient differences across all regimes.

Both models point to important differences between EAT and the other regimes with regard to school composition effects. The models provide solid evidence that the peer ability effects are significantly stronger in EAT states, supporting H10 (nonetheless, also in non-EAT states there are peer effects due to the interaction between individual and peer ability which does not differ between tracking regimes). Unexpected, but interestingly, we found also a sizable interaction with SES composition: in EAT states they are absent, while strong and significant in the non-EAT states. Contrary, individual SES effects do not differ between tracking regimes neither do individual ability effects or the respective interaction terms (additional models not shown). Entrance selectivity of schools explain part of the found tendencies, but sizable coefficients for interactions ($p<.10$) remain (models not shown).

Net of the prior ability context visiting a higher (lower) SES school brings a sizable additional advantage (disadvantage) for achievement, but not in states that select on ability.

How can we make sense of that finding? With the data at hand, we can only speculate. A possible explanation might be quality factors of schools (resources and teachers) that have effects on students learning, but are unobserved by our model. These factors might be correlated with schools' social composition in states that have a lower achievement orientation. One could make the case that states that ability-track their students on a legal basis at the same time provide a higher standardization of schools and curricula across all tracks due to the higher responsibility and accountability on the system level. This would be consistent with the arguments of von Below discussed above, since most of ET and PC states are rather liberal and reformed ("loose structures") adhering less to general norms while putting emphasize on autonomy and the individual (family).

=== TABLE 5 here ===

Conclusion

This paper contributes to the ongoing debate on how educational systems impact on academic performance in secondary education by studying the German case using recent, representative, and longitudinal data on a cohort of five graders from the German National Educational Panel Study (NEPS). We are (a) employing a large national sample on Germany, (b) exploiting variation across 16 educational systems, (c) adding prior abilities of pupils to the analysis, and (d) applying a three level model, distinguishing various influences on individual achievement on the level of states, schools, and students.

Overall, we found that academic performance in Science of sixth graders differs by tracking regime and is highest in German states that track students early based on prior achievement and lowest in states that track early but on the basis of freedom of choice. While differences between tracking regimes regarding the dispersion in performance are rather minor, social inequality in academic performance seems to be actually lowest in states following ability tracking. Our multivariate results clearly showed the importance of including information on prior ability of pupils for the correct understanding of selection in the different types of German educational systems. Without the inclusion of prior ability, the meaning of other variables like parental background and socio-economic compositions of schools can be easily misunderstood. However, this does not mean that parental socio-economic status has become irrelevant after inclusion of prior ability: it has still an important impact of educational performance in secondary education.

We did not find a significant effect for attending Gymnasium or another secondary track. But this was only true after controlling for school's ability composition. Our observation window (1 to 1.5 years of exposure to secondary education) might be too short to pick-up an additional gymnasium effect after controlling for prior ability and ability composition. As we have seen, the later matters a lot for educational performance and confounds strongly SES composition effects. Neither had we found a significant interaction between SES composition and individual SES. Peer ability seems to benefit all students but better students a bit more.

Extending previous studies on secondary school achievement mainly based on cross-sectional data, our findings contribute to a more in-depth understanding of system level differences. We found pupils' achievement in secondary school to be higher in German states that explicitly track based on abilities. However, prior achievement is already higher in exactly those states. Consequently, earlier advantages translate into persisting advantages during secondary education. Our results corroborated that finding on the level of schools; students in more selective schools perform better, but have already a more able influx of students. Hence, a sizable part of the difference in secondary education is attributable to what happens before students enter secondary education.

Consistent with theoretical expectations on the homogenizing effect of ability tracking, we found that in states with early ability tracking peer ability effects are significantly stronger than in other states. Surprisingly, we found also SES composition effects to vary across tracking regimes: while in early ability tracking states they are absent, they are quite strong and significant in the other states. We provided an ad-hoc explanation, namely that quality features of schools might be less related to SES composition in German Laender with state-law enforced ability tracking. However, this finding deserves more thorough future investigations.

Our paper contributes to a methodological advancement of research on the effects of tracking. Using the example of Germany our paper tried to overcome three major drawbacks of the current research on the impact of tracking and differentiation on secondary school achievement and (social) inequality in achievement: (1) the exclusion of the mediating level of schools, (2) the reliance on cross-sectional data, (3) the failure to account for students' prior ability especially before tracking occurred. Eventually, our analysis calls for a critical reflection of what Esser labeled as the "standard approach". It is basically blind in testing the various mechanisms that operate in certain educational systems and produce educational outcomes like overall levels or socio-economic inequalities in achievement.

References

Aßmann, C., Walter, H., & Zinn, S. (2012). *Weighting the Firth and Ninth Grader Cohort Samples of the National Educational Panel Study, Panel Cohorts. Technical Report. NEPS Research Data Paper*. Bamberg.

Ayalon, H., & Gamoran, A. (2000). Stratification in academic secondary programs and educational inequality in Israel and the United States. *Comparative Education Review*, 54–80.

Becker, R., & Schubert, F. (2006). Soziale Ungleichheit von Lesekompetenzen. *Kölner Zeitschrift Für Soziologie Und Sozialpsychologie*, *58*(2), 253–284. doi:10.1007/s11575-006-0055-0

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

Bol, T., Witschge, J., Van de Werfhorst, H. G., & Dronkers, J. (2014). Curricular Tracking and Central Examinations: Counterbalancing the Impact of Social Background on Student Achievement in 36 Countries. *Social Forces*, (1), 1–48. doi:10.1093/sf/sou003

Boudon, R. (1974). *Education, opportunity, and social inequality: Changing prospects in western society.* New York: Wiley.

Brunello, G., & Checchi, D. (2007). Does school tracking affect equality of opportunity? New international evidence. *Economic Policy*, (October), 781–861.

Dollmann, J. (2011). Verbindliche und unverbindliche Grundschulempfehlungen und soziale Ungleichheiten am ersten Bildungsübergang. *Kolner Zeitschrift Fur Soziologie Und Sozialpsychologie*, *63*, 595–621. doi:10.1007/s11577-011-0148-z

Dronkers, J. (2015). In wiens voordeel werkt selectie aan het begin van het voortgezet onderwijs? [Who benefits from selection early in secondary education? A new approach to an old question]. *Mens En Maatschappij*, *90*(1), 5–24.

Dronkers, J., Van Der Velden, R., & Dunne, A. (2012). Why are migrant students better off in certain types of educational systems or schools than in others? *European Educational Research Journal*, *11*(1), 11–44. doi:10.2304/eerj.2012.11.1.11

Duchhardt, C., & Gerdes, A. (2012). *NEPS Technical Report for Mathematics - Scaling Results of Starting Cohort 3 in Fifth Grade* (No. 19). *NEPS Working Papers*. Bamberg.

Dunne, A. (2010). *Dividing Lines: Examining the relative importance of between- and within-school differentiation during lower secondary education.* European University Institute.

Esser, H. (n.d.-a). Bildungssysteme und ethnische Bildungsungleichheit. In C. Diehl, C. Hunkler, & C. Kristen (Eds.), *Ethnische Ungleichheiten im Bildungsverlauf: Mechanismen, Befunde, Debatten*. Wiesbaden: Springer VS.

Esser, H. (n.d.-b). Sorting and (much) more. Prior ability, school-effects and the impact of ability tracking on educational inequalities in achievement. In A. Hadjar & C. Grosse (Eds.), *Education systems and educational inequalities*. Bristol: Policy Press.

Esser, H., & Relikowski, I. (2015). *Is Ability Tracking (Really) Responsible for Educational Inequalities in Achievement ? A Comparison between the Country States Bavaria and Hesse in Germany* (No. 9082). *IZA Discussion Paper*.

Gamoran, A., & Mare, R. D. (1989). Secondary School Tracking and Educational Inequality: Compensation, Reinforcement, or Neutrality? *American Journal of Sociology*, *94*, 1146. doi:10.1086/229114

Gresch, C., Baumert, J., & Maaz, K. (2010). Empfehlungsstatus, Übergangsempfehlung und der Wechsel in die Sekundarstufe I: Bildungsentscheidungen und soziale Ungleichheit. In *Bildungsentscheidungen* (pp. 230–256).

Hahn, I., Schöps, K., Saß, S., Hansen, S., Martensen, M., Wagner, H., & Funke, L. (2013). *The Assessment of Scienfic Literacy in the National Educational Panel Study (NEPS) including example items for Kindergarten, grade 6, students and adults.* Bamberg.

Hanushek, E. A., & Woessman, L. (2006). Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence Across Countries. *The Economic Journal*, *116*(1984), 63–76.

Kerckhoff, A. C. (1986). Effects of Ability Grouping in British Secondary Schools. *American Sociological Review*, *51*(6), 842. doi:10.2307/2095371

Maaz, K., Baumert, J., Gresch, C., & McElvany, N. (Eds.). (2010). *Der Übergang von der Grundschule in die weiterführende Schule - Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten*. *Der Übergang von der Grundschule in die weiterführende Schule*. Berlin: Bundesministerium für Bildung und Forschung.

Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Educational transitions and differential learning environments: How explicit between-school tracking contributes to social inequality in educational outcomes. *Child Development Perspectives*, *2*, 99–106. doi:10.1111/j.1750-8606.2008.00048.x

Marks, G. N. (2005). Cross-National Differences and Accounting for Social Class Inequalities in Education. *International Sociology*, *20*(December), 483–505. doi:10.1177/0268580905058328

Neugebauer, M. (2010). Bildungsungleichheit und Grundschulempfehlung beim Übergang auf das Gymnasium: Eine Dekomposition primärer und sekundärer Herkunftseffekte. *Zeitschrift Für Soziologie*, *39*(3), 202–214.

Neumann, M., Schnyder, I., Trautwein, U., Niggli, A., Lüdtke, O., & Cathomas, R. (2007). Schulformen als differenzielle Lernmilieus. *Zeitschrift Für Erziehungswissenschaft*, *10*(3), 399–420. doi:10.1007/s11618-007-0043-6

Olczyk, M., Will, G., & Kristen, C. (2014). *Personen mit Zuwanderungshintergrund im NEPS: Zur Bestimmung von Generationen-Status und Herkunftsgruppe* (No. 41b). *NEPS Working Papers*.

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). *NEPS Technical Report for Reading – Scaling Results of Starting Cohort 3 in Fifth Grage* (No. 15). *NEPS Working Papers*. Bamberg.

Roth, T., & Siegert, M. (2015). Freiheit versus Gleichheit? Der Einfluss der Verbindlichkeit der Übergangsempfehlung auf die soziale Ungleichheit in der Sekundarstufe. *Zeitschrift Für Soziologie*, *44*(2), 118–136.

Rumberger, R., & Palardy, G. (2005). Does segregation still matter? The impact of student composition on academic achievement in high school. *The Teachers College Record*, *107*(9), 1999–2045.

Skopek, J., Pink, S., & Bela, D. (2012). *Data Manual. Starting Cohort 3 – From Lower to Upper Secondary School. NEPS SC3 1.0.0. NEPS Research Data Paper*. Bamberg.

Van de Werfhorst, H. G., & Mijs, J. J. B. (2010). Achievement Inequality and the Institutional Structure of Educational Systems: A Comparative Perspective. *Annual Review of Sociology*, *36*(1), 407–428. doi:10.1146/annurev.soc.012809.102538

Von Below, S. (2006). Bildungssysteme und Selektivität. Eine Typologie am Beispiel der neuen Bundesländer. *Die Deutsche Schule*, *98*(2), 230–242.

Tables

Table 1 Classification of binding primary school recommendation on state level

Federal state	PSR Binding [a]	Tracking starts in Grade	Typology (von Below) [b]	Tracking regime type [c]
Baden-Wuerttemberg	Yes	5	traditional-conservative	EAT
Bavaria	Yes	5	traditional-conservative	EAT
Berlin	No	7	reformed-liberal	PC
Brandenburg	Yes	7	reformed-liberal	PC
Bremen	No	5	reformed-liberal	ET
Hamburg	No	5	reformed-liberal	ET
Hesse	No	5	reformed-liberal	ET
Mecklenburg-Western Pomerania	No	7	traditional-conservative	PC
Lower Saxony	No	5	reformed-liberal	ET
North Rhine-Westphalia	No	5	reformed-liberal	ET
Rhineland-Palatinate	No	5	traditional-liberal	ET
Saarland	Yes	5	reformed-conservative	EAT
Saxony	Yes	5	reformed-conservative	EAT
Saxony-Anhalt	Yes	5	reformed-conservative	EAT
Schleswig-Holstein	No	5	traditional-liberal	ET
Thuringia	Yes	5	reformed-conservative	EAT

Notes: [a] Binding character of primary school recommendation; valid for end of 2010; classification based on Gresch et al. (2010) and Neugebauer (2010), a review of state reports, and legal texts from state education laws. A border case is North Rhine-Westphalia, which had temporarily introduced a binding recommendation in 2006 which was abandoned in 2010 again. [b] compare von Below (2006). [c] EAT = early ability tracking, ET = early tracking, PC = prolonged comprehensive.

Table 2 Variables on individual and school level by tracking regime

	TOTAL	State level tracking regime		
		EAT	ET	PC
	n=5,444	n=2,046	n=2,914	n=484
Individual level [a]				
Science, Grade 6 (z-std.)	0.00	0.15	-0.11	0.07
	(1.00)	(1.02)	(0.97)	(0.99)
ABIL	0.00	0.17	-0.11	-0.02
	(1.00)	(1.01)	(0.98)	(0.96)
SES	0.00	0.03	-0.03	0.14
	(1.00)	(1.00)	(1.00)	(0.99)
Female	.48	.47	.48	.49
Migrant	.20	.16	.24	.13
School level [b]	n=230	n=87	n=117	n=26
'Gymnasium'	.31	.35	.32	(.05)[d]
SOA	0.00	0.32	-0.14	-0.90
(non-missing, z-std.)	(1.00)	(1.09)	(0.78)	(0.78)
Principal non-response [c]	.32	.28	.38	.19
S-ABIL	-0.14	0.00	-0.24	-0.17
	(0.70)	(0.72)	(0.69)	(0.63)
S-SES	-0.11	-0.07	-0.16	-0.01
	(0.64)	(0.63)	(0.64)	(0.67)

Notes: Estimation based on M=20 imputation datasets; standard deviations averaged across imputation datasets and put in parentheses; [a] using student weights, [b] using school weights, [c] unweighted sample fraction. [d] only one school.

Table 3 Measures of inequality by tracking regime

	State level tracking regime		
	EAT	ET	PC
Science score, Grade 6			
SD	1.02	.97	.99
$95^{th} – 5^{th}$	3.27	3.21	3.01
$75^{th} – 25^{th}$	1.33	1.32	1.23
R^2 (SES)	.13	.18	.16
School level: R^2 (SES)	.51	.74	.83
ABIL, Grade 5			
SD	1.01	0.98	0.96
$95^{th} – 5^{th}$	3.24	3.22	3.15
$75^{th} – 25^{th}$	1.33	1.32	1.28
R^2 (SES)	.14	.18	.19
School level: R^2 (SES)	.57	.77	.82

Notes: Estimation based on M=20 imputation datasets; calculations based on student weights, for school level indicators based on school weights. 95^{th} -5^{th} : difference between 95^{th} and 5^{th} percentile; analogous for 75^{th}-25^{th}; R^2 (SES): fraction of explained variance of a regression on mean parental education (SES).

Table 4 Standardized Science score in Grade 6 (linear mixed models)

	1	2	3	4	5a	5b	6	7	8	8 (FE)
Tracking system										
EAT (ref. ET)	0.240*	0.145	0.162**	0.166**	0.064	0.031	0.058	0.059	0.047	State dummies
PC (ref. ET)	0.194	0.357*	0.138^{+}	0.255***	0.190**	0.120^{+}	0.101^{+}	0.105^{+}	0.105^{+}	
School level										
SOA		0.268***	0.099**	0.069*	0.082***	0.004	0.007	0.007	0.012	0.016
S-SES			0.579***	0.397***			0.107^{+}	0.106^{+}	0.104^{+}	0.086
S-ABIL						0.282***	0.138*	0.131*	0.102	0.124*
GYM				0.372***				0.013	0.057	0.050
Individual level										
SES			0.174***	0.175***			0.103***	0.103***	0.092***	0.093***
SES x S-SES			0.029	0.021			0.023	0.023	0.023	0.016
ABIL					0.586***	0.535***	0.520***	0.520***	0.503***	0.502***
ABIL x S-ABIL						0.068**	0.061**	0.060**	0.062**	0.067**
Controls										
Female									-0.044^{+}	-0.044^{+}
Migrant									-0.230***	-0.231***
Missing SOA		-0.034	0.062	0.030	0.028	0.040	0.048	0.047	0.042	0.036
Constant	-0.195***	-0.158*	-0.124***	-0.256***	-0.056	-0.062^{+}	-0.079**	-0.084*	-0.029	0.000
Variances										
School	0.317	0.272	0.075	0.059	0.051	0.032	0.026	0.026	0.023	0.019
Individual	0.697	0.698	0.676	0.675	0.522	0.520	0.512	0.512	0.506	0.507

Notes: N=5,444 students, nested in N=230 schools. Significance level: $^{+}$ p<.1; * p<.05; ** p<.01; *** p<.001.

Table 5 School composition effects by tracking regime

Coefficients	Model 10			Model 11	
	S-ABIL	S-SES		S-ABIL	S-SES
EAT (1)	0.194**	-0.017	EAT (1)	0.193**	-0.017
$\delta^{1\text{-}2}$	*	+	$\delta^{1\text{-}2}$	+	+
ET & PC (2)	0.005	0.204*	ET (2)	0.008	0.197*
			$\delta^{2\text{-}3}$	*n.s.*	*n.s.*
			PC (3)	-0.014	0.246+
			$\delta^{3\text{-}1}$	*n.s.*	*n.s.*

Notes: Model 10 and 11 are modified versions of Model 8 (FE), including interaction terms and excluding SOA. N=5,444 students, nested in N=230 schools. δ = p-value of difference. Significance level: + p<.1; * p<.05; ** p<.01; *** p<.001.

Appendix

Table A1: Loadings of items on admission criteria of secondary schools regarding achievement on component "achievement" (principal component analysis).

Admission criteria applied by secondary school	Component "*selection on achievement*" (EV 1.98, 62% of variance)
(2) prior school achievement/grades	0.69
(3) entrance examination	0.28
(4) trial lesson	0.27
(5) students' previous school's recommendation	0.61

Notes: Component's Eigenvalue in parentheses. N=445 schools with responding principals. Missing values (12 percent of cases) imputed by multiple imputation.

Table A2: Selection on achievement score by tracking regime and type of school

Type of secondary school	EAT		ET		PC		Total	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Hauptschule	-0.40	0.90	-0.53	0.72	-	-	-0.46	0.82
Realschule	0.90	0.90	0.17	0.78	-	-	0.44	0.88
Gymnasium	1.17	0.75	0.06	0.70	(1.68)[a]	-	0.65	0.91
Comprehensive	-0.40	0.55	-0.42	0.74	-1.10	0.31	-0.70	0.63
Total	0.32	1.09	-0.14	0.78	-0.90	0.78	0.00	1.00

Notes: N=153 schools with non-missing selection on achievement score (= with responding principals). [a] only one school.