Educational Assessment, Evaluation and Accountability
https://doi.org/10.1007/s11092-020-09346-4

Early tracking and different types of inequalities in achievement: difference-in-differences evidence from 20 years of large-scale assessments

Andrés Strello [1] · Rolf Strietholt [1,2,3] · Isa Steinmann [1,4] · Charlotte Siepmann [1]

Received: 3 December 2019 / Accepted: 2 December 2020/

Abstract

Research to date on the effects of between-school tracking on inequalities in achievement and on performance has been inconclusive. A possible explanation is that different studies used different data, focused on different domains, and employed different measures of inequality. To address this issue, we used all accumulated data collected in the three largest international assessments—PISA (Programme for International Student Assessment), PIRLS (Progress in International Reading Literacy Study), and TIMSS (Trends in International Mathematics and Science Study)—in the past 20 years in 75 countries and regions. Following the seminal paper by Hanushek and Wößmann (2006), we combined data from a total of 21 cycles of primary and secondary school assessments to estimate difference-in-differences models for different outcome measures. We synthesized the effects using a meta-analytical approach and found strong evidence that tracking increased social achievement gaps, that it had smaller but still significant effects on dispersion inequalities, and that it had rather weak effects on educational inadequacies. In contrast, we did not find evidence that tracking increased performance levels. Besides these substantive findings, our study illustrated that the effect estimates varied considerably across the datasets used because the low number of countries as the units of analysis was a natural limitation. This finding casts doubt on the reproducibility of findings based on single international datasets and suggests that researchers should use different data sources to replicate analyses.

Keywords Ability tracking · Difference-in-differences · Educational inequality · Large-scale assessment · Performance · Stratification

✉ Andrés Strello
andres.strello@tu–dortmund.de

Extended author information available on the last page of the article

1 Introduction

Levels of institutional differentiation are characteristic features of educational systems. In this context, there is a very controversial discussion concerning early between-school ability tracking, i.e., regarding the grade at which students are separated into different ability tracks with different curricula and different access to higher education. For example, Germany tracks students after the fourth grade,[1] while countries like the USA do not track students into ability-grouped schools before higher education.

The arguments in favor of selective schooling center on a perceived trade-off between equity and efficacy (Hanushek and Wößmann 2006). Those who believe in the efficacy of track differentiation argue that it is easier and more efficient to teach more homogeneous student groups. Tracking advocates also argue from a societal perspective that vocational and academic tracks give rise to school leavers with a mix of qualifications, which is beneficial in a heterogeneous job market. However, this does not consider the possible effects of tracking on equity, especially in the case of very early tracking. A possible social bias in the track selection process and differential expectations, motivations, and resources between the different tracks might contribute to increased inequality (Maaz et al. 2008).

Most previous research on tracking compared countries with tracked and comprehensive school systems. The majority of studies, however, were based on simple correlations and failed to account for the possibility that countries with a tracked as opposed to a comprehensive school system might differ in terms of other important institutional features (van de Werfhorst and Mijs 2010). To disentangle the effect of tracking from the effects of other institutional determinants of student achievement, Hanushek and Wößmann (2006) proposed a difference-in-differences approach where they combined primary (before tracking) and secondary (after tracking) school data to identify the causal effect of early between-school tracking on educational outcomes. This approach has also been adopted by other studies since it allows researchers to identify the effect of tracking on achievement. The findings of these studies paint an inconclusive picture. A limitation of international comparative studies is that their effect estimations are based on rather small samples, since the level of analysis is the country level and the number of countries is naturally limited. Furthermore, different studies have focused on different samples of countries, international assessments, assessment cycles, domains, and measures of educational inequality. For this reason, it is difficult to determine whether inconclusive research findings are due to substantive differences in the setup of the different studies or due to imprecisions in the estimations caused by small samples.

The main purpose of the present study was to use the accumulated data of three international large-scale assessments: the Programme for International Student Assessment (PISA), the Progress in International Reading Literacy Study (PIRLS), and the Trends in International Mathematics and Science Study[2] (TIMSS). Combining data from different studies and study cycles increased the sample size and helped us to obtain more precise tracking effect estimates. Furthermore, we used the same data to

[1] Most schools in Germany track students after the fourth grade. There are, however, some exceptions in individual federal states.

[2] In 1995, the TIMSS study was called the Third International Mathematics and Science Study.

systematically replicate the analyses for different outcome measures. Specifically, we focused on the effects of tracking on performance levels and three different types of inequalities in achievement, namely dispersion inequality, social achievement gaps, and educational inadequacy.

This paper is divided into five sections. First, we review the theoretical and empirical research on the effects of tracking on different types of inequalities and on performance. Second, we specify our research question and the aim of this study. Third, we present the analytical approach we use to identify the effect of tracking and our approach to combine the results from different analyses. Fourth, we describe the main results regarding the effects of tracking on inequalities and performance. Fifth, we discuss our findings and provide conclusions for educational policy and future research.

2 Literature review: how does tracking affect educational inequalities?

In the first part of the literature review, we outline a theoretical framework for the effects of tracking on inequalities and performance, preceded by a brief clarification of the distinction between three types of educational inequalities. We focus on achievement as it is an important predictor of, for instance, labor market returns, wellbeing, political engagement, integration, and countries' economic growth (Brighouse et al. 2018; Hanushek 2013; Hanushek et al. 2015). In the second part, we review previous studies on the effects of tracking.

2.1 Different concepts of achievement inequalities

Inequality is a term that has been used in quite different ways by different authors. Van de Werfhorst and Mijs (2010) distinguished between *inequality as dispersion* and *social inequality*. These two conceptualizations have different normative ideas about what is unjust (Strietholt 2014) and we think that identifying the differences between inequality conceptualizations is important for the evaluation of the results. Inequality as dispersion implies that the mere existence of differences in achievement is problematic. Social inequality regards differences *between* social groups as problematic but does not consider the mere existence of variation *within* each group problematic. Strietholt and Borgna (2018) noted that several studies on educational inequality focused on *threshold inequality*, which centers on the lower distribution of scores and refers to the proportion of students who do not reach a minimum performance level. This concept is also referred to as *educational deprivation* (Solga 2014, p. 271), *minimum standard* (UNESCO 2018), or *educational adequacy* (Brighouse and Swift 2008, 2009). The basic idea of threshold inequality is that all students should reach a certain threshold achievement level, while inequalities beyond this threshold are not problematic. Therefore, we evaluate the effects of tracking separately for each conceptualization of inequality, as each concept implies different normative ideas about justice. Different inequalities can furthermore be expected to have different implications for societal and individual development. In addition, there are empirical reasons to study the effects on the three concepts separately, as the measurements of the concepts of inequality are not found to correlate with each other. For instance, the dispersion of scores is not associated with the performance gap between students from lower and higher social classes (Strietholt and Borgna 2018).

2.2 Tracking as transition and the effects on inequality

In theoretical terms, between-school tracking constitutes a type of educational stratification that is external (differentiation between schools) and formal (regulated by law) (Chmielewski 2014; Dollmann 2019; Skopek et al. 2019). While our study focuses on between-school tracking, our findings and arguments may apply to other mechanisms of educational differentiation (e.g., within-school tracking). At least three different mechanisms explain how tracking reinforces inequality in achievement; we introduce these before reviewing studies on tracking effects on different types of inequality. First, we describe how the stigmatization of lower tracks affects students at the lower end of the achievement distribution (*educational inadequacy*). Second, we outline how unequal curricula and resources explain an effect of tracking on the overall achievement distribution (*dispersion inequality*). Third, we depict how social bias in allocating students to different tracks perpetuates social inequalities in achievement (*social achievement gaps*).

Stigmatization of lower tracks One set of arguments against tracking rests on the anticipated disadvantages for students in lower tracks (Slavin 1990). Various researchers observed that students in lower tracks developed negative attitudes toward school; they also expected little future payoff, had lower educational expectations, and had more pronounced feelings of futility than students in higher tracks (Karlson 2015; Lee 2014; van Houtte and Stevens 2015). Such negative attitudes may have consequences for student learning. At the same time, the social composition of schools may have consequences for children's education. More homogeneous groups may inhibit the positive peer effects of heterogeneous classes, where disadvantaged students may benefit from the shared learning environment (Coleman et al. 1966; Sacerdote 2011). In contrast to the idea of *no child left behind*, the existence of lower tracks legitimizes poor performance by some students. Following this line of argumentation, tracking might increase the proportion of students who do not have basic literacy skills, a phenomenon that is essentially related to the concept of educational inadequacy.

Unequal curricula and resources Different tracks lead to different educational pathways that allow students to pursue academic or vocational careers. Such differences are manifested in curricula that are more or less ambitious in lower and higher tracks. In the same vein, the allocation of educational resources—such as teacher quality, infrastructure, and funding—may differ between tracks. Indeed, there is evidence that students in higher tracks benefit from better educational resources (Becker et al. 2012; Guill et al. 2017; Martinková et al. 2020). Such track-specific inequalities in educational opportunity may lead to a higher dispersion of educational outcomes, i.e., dispersion inequality.

Transitions and social bias So far, this paper has not needed to challenge the assumption that students are allocated to different tracks based on their abilities in order to hypothesize that tracking increases different types of inequality in achievement. However, transitions within the educational systems may reinforce social inequality. Boudon (1974) proposed two mechanisms through which transitions may reinforce social inequality: first, privileged children tend to perform better (primary effects), and second, even after controlling for prior achievement, privileged students have greater chances of accessing more ambitious tracks (secondary effects). There is a plethora of evidence

showing that tracking decisions are not solely based on performance (which could have primary segregation effects), but also depend on race or social class after taking previous academic achievement into account (secondary effects) (Batruch et al. 2018; Hallinan 1994; Holm et al. 2013; Horn 2013; Lucas and Berends 2002; Maaz et al. 2008; Pietsch and Stubbe 2007). Additionally, children from privileged backgrounds might receive more support from their parents to reach high tracks (Koerselman 2013). In this respect, the time point at which students are tracked is a critical moment. A recurring hypothesis is that parental background exerts a strong influence on educational transitions, especially when children are younger (Bauer and Riphahn 2006; Chmielewski 2014; Hillmert and Jacob 2010; Lange and von Werder 2017; Schütz et al. 2008). If different tracks lead to a stigmatization of students or provide different educational opportunities for them, social bias in the tracking process will result in higher social achievement gaps. This contradicts the ideal of tracking as a meritocratic process.

2.3 Empirical evidence of early tracking effects

The previous research on early tracking effects can be divided into three categories: studies that conduct cross-sectional analyses on a between-country level, studies that apply quasi-experimental designs, and in within-country comparative studies (cf. Skopek et al. 2019). Cross-sectional studies with international data showed mixed findings regarding the associations of between-school tracking and dispersion inequality (Huang 2009; Micklewright and Schnepf 2007; van de Werfhorst and Mijs 2010). Such cross-sectional studies also found that between-school tracking is associated with higher levels of social inequality (Dämmrich and Triventi 2018; Dollmann 2019; Duru-Bellat and Suchaut 2005; Gorard and Smith 2004; Horn 2009; Marks 2005; Schlicht et al. 2010; Schütz et al. 2008; Skopek et al. 2019; van de Werfhorst and Mijs 2010). However, cross-sectional studies only use information from one point in time and do not allow researchers to draw causal conclusions.

Few studies have used robust designs that allowed researchers to draw causal inferences on the effects of tracking. Most of these robust studies estimated difference-in-differences models to exploit the fact that no country has a tracked primary school system, while some countries allocate students to different ability tracks at the secondary school level. Therefore, researchers can compare student outcome measures in tracked versus comprehensive school systems at the secondary school level while controlling for the same measures at the primary school level to identify the effects of tracking. Another robust approach for identifying tracking effects is to study variation in tracking status within countries over time. There are, however, only two studies that employed this approach, since such school-system reforms rarely occur.

In the following, we review studies on the effects of tracking on dispersion inequality, educational inadequacy, and social achievement gaps. Furthermore, we review findings on tracking effects on performance levels in order to provide some evidence for a possible trade-off between efficacy and inequality.

Effects on dispersion inequality Hanushek and Wößmann (2006) used PISA, TIMSS, and PIRLS data from several cycles administered between 1995 and 2003 in the domains of mathematics, reading, and science. They combined eight pairs of primary

and secondary school studies (e.g., PIRLS 2001 and PISA 2000) and estimated a series of difference-in-differences models for each pair. While they found substantial variation in the effect estimates for different pairs of studies, the pooled estimate indicated that early tracking increased the dispersion of test scores. The variation in the effect estimates might have been due to the fact that each pair of studies only looked at 18 to 26 countries. The findings provided little evidence for domain-specific differences in the effect estimates. Jakubowski (2010) replicated Hanushek and Wößmann's (2006) study of PIRLS 2001 and PISA 2000 data and found no significant effect on dispersion inequality. Hanushek and Wößmann found no effect for this particular pair of studies either. However, Jakubowski (2010) also analyzed another combination of TIMSS 2003 and PISA 2003 data and again found no effect. Further studies replicated Hanushek and Wößmann's approach using international data but focused on other educational outcomes and not dispersion inequality (see below).

To our knowledge, only one study has exploited national educational reforms to examine the effects of tracking on dispersion inequality. Piopiunik (2014) combined German data from the PISA 2003 and 2006 cycles and found that lowering the age of tracking increased dispersion inequality significantly. This study focused on a policy change in the federal state of Bavaria, where the tracking age was lowered from sixth to fourth grade. The study provided no evidence that the effects differed for mathematics, reading, and science.

Effects on educational inadequacy Some studies have estimated the effects of early tracking on different quantiles of the achievement distribution. The percentiles at the lower end of the international achievement distribution can be perceived as thresholds defining educational adequacy. The evidence suggests that tracking increases the number of students who do not achieve basic literacy. Hanushek and Wößmann (2006) found that tracking had a negative effect on the performance of students in the lower quantile of the achievement distribution. Similar analyses of more recent study cycles of PIRLS, TIMSS, and PISA replicated the finding that early tracking had a negative effect on performance at the lower end of the achievement distribution (Lavrijsen and Nicaise 2016). The effects were most pronounced in reading. The aforementioned study by Piopiunik (2014) provided further evidence for a negative effect of early tracking on educational adequacy. Lowering the tracking age in the German state of Bavaria increased the share of low performers in mathematics, reading, and science.

Effects on social achievement gaps Findings from the research on effects of tracking on social inequality have been inconclusive. While some studies provided evidence that tracking perpetuated social inequality, most observed no tracking effect on social achievement gaps. Ammermüller (2005) estimated a difference-in-differences model based on PISA 2000 and PIRLS 2001 data from 14 countries and found that the effect of social background on reading achievement was more pronounced in countries with more differentiated school tracks. Other studies used the tracking age instead of the number of school tracks as the main explanatory variable. Waldinger (2007) found no effect of the tracking age on the social gap in reading achievement using PIRLS 2001 and PISA 2003 data from a similar but not identical set of 14 countries. Jakubowski (2010) studied the effects of early tracking on social gaps in reading and mathematics. The analyses of PIRLS 2001 and PISA 2003 reading data from 23 countries revealed

no significant effects. The analyses of TIMSS 2003 and PISA 2003 mathematics data from 15 countries, however, provided some evidence that early tracking significantly increased social gaps in mathematics achievement. A study using more recent data from PIRLS 2006 and PISA 2012 ($N = 33$ countries) observed that an earlier tracking age increased social gaps in reading achievement (Lavrijsen and Nicaise 2015).

A general limitation of the previously presented research was that each study was based on a small set of countries. To address this issue, Ruhose and Schwerdt (2016) combined data from five PISA cycles (2000–2012), five TIMSS cycles (1995–2011), and two PIRLS cycles (2001–2006). In total, they analyzed data from 45 countries. Many of these countries were observed in different studies and at multiple time points. The study provided no evidence that tracking increased the achievement gap between native and immigrant students.

Van de Werfhorst (2018) combined secondary school data from the First International Mathematics Study (FIMS) from 1964, the Second International Mathematics Study (SIMS) from 1980 to 1982, and the Third International Mathematics and Science Study (TIMSS) from 1995. The study showed that social achievement inequality was lower in countries that had transformed their school system from tracked to comprehensive than in countries where tracking was retained. A limitation of this study was that it was only based on nine countries that participated in all three international assessments and that only four of these had reformed their school systems.

Effects on performance levels Studies on the effects of tracking on performance levels revealed mixed findings. Hanushek and Wößmann (2006) and Lavrijsen and Nicaise (2016) replicated analyses on the effects of tracking on performance levels for eight combinations of primary and secondary school assessments. Both reported a tendency for early tracking to reduce performance levels. However, more than half of the single estimates were neutral and one was even significantly positive. Jakubowski (2010) analyzed two study pairs and found one neutral and one negative effect on performance levels.

In the same vain, two single country studies in Germany and Northern Ireland reported contradictory findings. Piopiunik (2014) found a negative effect of tracking on performance levels in Bavaria in Germany. Guyon et al. (2012) found evidence for an improvement of results when increasing the number of students attending the higher track in Northern Ireland.

Summary of the review The review of research revealed inconsistent findings, which makes it impossible to draw robust inferences on the effects of tracking on student outcomes. We propose two possible explanations for the variation in the effect estimates related to conceptual differences in the outcome measures and to the small sample sizes at the country level.

The conceptual distinction between different educational outcomes seems to explain some of the variation in the results of different studies. At the same time, it is difficult to draw strong conclusions about conceptually different outcomes because the number of studies was limited for each outcome. While several studies focused on social achievement gaps as outcomes, only two investigated the effects of tracking on dispersion inequality. Furthermore, the different studies were based on different datasets and focused on different achievement domains, which makes it even more difficult to distinguish between substantive differences and sampling error.

The low sample size at the country level is another serious issue. Typically, studies only used data from around 20 countries when combining primary and secondary school assessments. Studies that replicated the analyses based on different combinations of primary and secondary school datasets revealed a remarkably high variability in the effect estimates. This illustrates that findings based on single combinations of datasets are unreliable. In this regard, the study by Ruhose and Schwerdt (2016) is an exception because it combined data from several cycles of PIRLS, PISA, and TIMSS in 45 study pairs to increase the sample size and to achieve more reliable estimates. However, that study focused on the achievement gap between native and immigrant students, which is conceptually related to but different from *social* gaps in achievement.

3 Research questions

The aim of this paper was to use international data to estimate the effects of early tracking on three different types of inequalities in achievement—dispersion inequality, social achievement gaps, and educational inadequacy—and on performance levels. Following Hanushek and Wößmann (2006), we combined primary and secondary school assessments to identify the effect of tracking by applying difference-in-differences analyses. Previous research used different datasets to study different outcomes and mostly drew on rather small samples of countries. Following Ruhose and Schwerdt (2016), we attempted to overcome these limitations by using all available cycles of PISA, TIMSS, and PIRLS administered between 1995 and 2016. The combined data increased the analytical sample and allowed us to study different outcomes.

4 Methodology

4.1 Data sources: combining primary and secondary school information

To identify tracking effects, we exploited the fact that some countries track their students after primary school, while others employ a comprehensive secondary school system. For this purpose, we combined primary and secondary school data from all available cycles of three international large-scale assessments—PIRLS, PISA, and TIMSS—administered between 1995 and 2016.

PIRLS was conducted in 2001, 2006, 2011, and 2016 and assessed reading achievement in fourth grade, at the end of primary school. PISA was administered in 2000, 2003, 2006, 2009, 2012, and 2015 and tested the reading, mathematics, and science performance of 15-year-old secondary school students. TIMSS was conducted in 1995, 1999, 2003, 2007, 2011, and 2015. TIMSS measured student achievement in mathematics and science in both fourth grade (population A) and eighth grade (population B). TIMSS 1999 only tested eighth graders. All studies contained survey weights to generalize from the representative samples to the underlying student populations in the respective countries or regions.

In order to determine changes between primary and secondary school, we matched primary school data from PIRLS or TIMSS population A with secondary school data

from the same countries from PISA or TIMSS population B. For this purpose, we applied two matching approaches: first, matching roughly the same years (e.g., PIRLS 2001 with PISA 2000), and second, matching roughly the same cohorts (e.g., PIRLS 2001 with PISA 2006). We applied both approaches because combinations from the same years are subject to period effects, while combinations from the same cohorts are subject to cohort effects (e.g., Blanchard et al. 1977). Figure 1 illustrates the 45 study pairs that formed the basis for our analyses. Nine study pairs matched PIRLS with PISA data, 18 matched TIMSS population A with PISA data, and 18 matched TIMSS population A with TIMSS population B data. We counted the study pairs for TIMSS population A and PISA data and the pairs for TIMSS population A and TIMSS population B data twice since we ran all analyses for mathematics and science separately. In sum, our paired analysis dataset contained information from 75 countries or regions and more than 2 million students. Each country was observed at least two times. The overall number of single observations underlying the study pairs in Fig. 1 by study, cycle, domain, and country (study-by-cycle-by-domain-by-country observations) amounted to 1177.

4.2 Variables

Test scores To compare educational outcomes in primary and secondary school, we used plausible values of test scores for reading, mathematics, and science achievement. In each study, the scores were linked across assessment cycles so that they had the same metric over time. The scores were standardized to an international mean of 500 with a

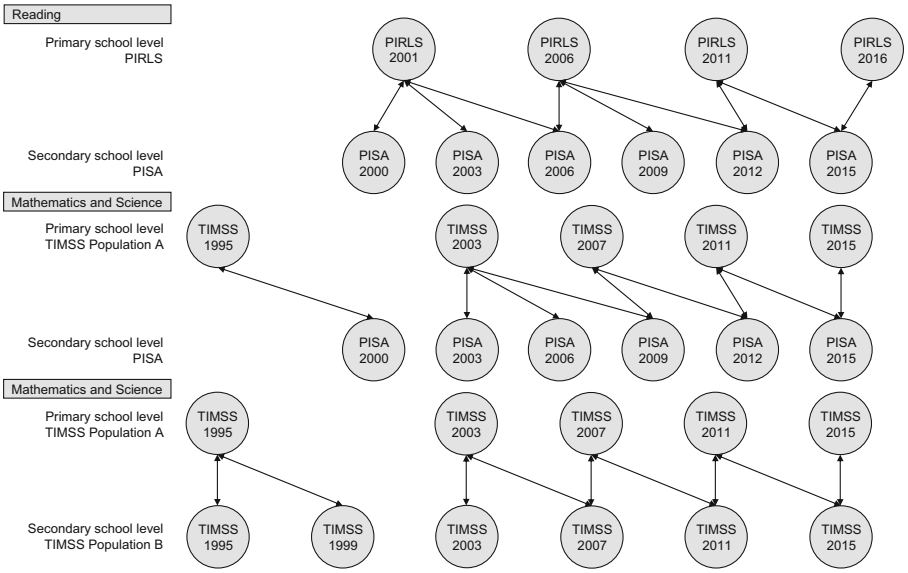

Fig. 1 Study pairs of large-scale assessments at primary and secondary school level. Every arrow reflects a study pair of datasets at primary and secondary school level. The study pairs contain data from all countries that participated in both assessments. The studies were combined so that they roughly matched the same years or the same cohorts. The study pairs of TIMSS Population A and PISA data as well as TIMSS Population A and TIMSS Population B data entered the analyses twice, since mathematics and science were treated separately in the analyses

standard deviation of 100 (Martin et al. 2016, 2017; OECD 2017). We used the test scores to compute three country-level measures of educational inequality and the mean performance level. The plausible values contained no missing data. To ensure that we could measure and compare different conceptualizations of inequality, we aggregated all variables at the country level.

Dispersion inequality We computed the weighted standard deviation of the test scores as our main measure of dispersion inequality for each of the 1177 study-by-cycle-by-domain-by-country observations. Table 1 shows the distribution of the variable in primary and secondary school. Interestingly, dispersion inequality in primary school was higher in late tracking countries but lower in secondary school.

In further robustness checks, we also computed alternative measures of dispersion inequality, namely the range between the 95th and 5th percentile and the range between the 75th and 25th percentile (interquartile range).

Social achievement gaps The social achievement gap was measured as the weighted mean difference in achievement scores between children from households with less than 100 and at least 100 books. We used the student-reported number of books variable in the main analyses since it was the only measure of socioeconomic status that was available in all international assessments of interest. This type of mean score difference is also referred to as a measure of absolute differences. Another frequently

Table 1 Descriptive statistics of the three inequality measures and the performance measure at Primary and secondary school level in the overall country sample and divided by tracking status

	Overall sample		Late tracking		Early tracking	
	M	*SD*	*M*	*SD*	*M*	*SD*
Dispersion inequality						
Primary school level	79.988	14.427	81.243	14.996	75.791	11.394
Secondary school level	89.016	11.509	87.754	11.694	93.238	9.764
Social achievement gap						
Primary school level	30.259	14.787	27.530	14.434	39.323	12.107
Secondary school level	51.999	15.826	48.491	14.716	63.701	13.642
Educational inadequacy						
Primary school level	13.537	17.829	15.947	19.437	5.479	5.785
Secondary school level	12.178	13.172	13.693	14.271	7.116	6.301
Performance level						
Primary school level	507.454	60.696	498.070	63.939	538.826	32.549
Secondary school level	491.128	48.851	486.765	50.213	505.716	40.810

The dispersion inequality is measured as the standard deviation of test scores, the social achievement gap as the mean difference in test scores between students with up to 100 versus at least 100 books at home, the educational inadequacy as the percentage of students not reaching PISA proficiency level 1b or the low benchmarks in PIRLS and TIMSS, and the performance level as the mean test score within countries. Early tracking means that tracking took place before grade eight (TIMSS population B) or in a grade where most students are younger than 15 years old (PISA). All means and standard deviations were estimated based on a total of 1177 study-by-cycle-by-domain-by-country observations

used measure is the relative gap, which considers the overall dispersion of test scores by dividing the absolute differences by the within-country standard deviations. The basic idea is that social groups are more meaningful if the overall dispersion of scores is small. We computed relative social achievement gaps for the number of books variable.

In further analyses, we also used parental education as an alternative measure of social background. Information on parental education was obtained from parents in the primary school studies PIRLS and TIMSS population A and from students in the secondary school studies PISA and TIMSS population B. We computed the absolute achievement gap between children of parents with and without tertiary education. However, information on parental education was not available for TIMSS population A cycles administered before 2011. Therefore, applying this measure reduced the analysis sample.

Missing data ranged from 3% for the books at home variable to 30% for parental education (based on the samples where this item was administered). To account for missing data, we created an imputed dataset using predictive mean matching (e.g., Rubin 1987) in the R package *mice* (van Buuren and Groothuis-Oudshoorn 2011). The imputation model used information on age, gender, parental education, number of books, country of birth of parents, language at home, and achievement scores.

Educational inadequacy To measure educational inadequacy, we computed the shares of students who did *not* meet certain thresholds of the achievement scales for each study-by-cycle-by-domain-by-country observation. We defined the thresholds based on the so-called PISA proficiency level 1b and the low PIRLS and TIMSS international benchmarks. Table 1 shows that, on average, 14% of primary and 12% of the secondary school level students did not reach these levels of adequate achievement in the present sample.

In further analyses, we used more inclusive thresholds and replicated the analyses. Specifically, we used the proficiency level 2 for PISA and the intermediate benchmark for PIRLS and TIMSS. On average, about 30% of the students did not reach these more inclusive adequacy cutoffs.

Performance level We used the weighted mean achievement as a performance level measure in all study-by-cycle-by-domain-by-country observations. As Table 1 shows, the average performance levels were higher in early tracking countries than in late tracking countries at both primary and secondary school level.

Early tracking Educational systems track their students into different ability tracks at different ages and grades. To determine the grade and age at which the countries of interest tracked their students, we reviewed reports by UNESCO (UNESCO-IBE 2007, 2012), Eurydice (2005, 2011, 2013a, b, 2014), and OECD (2004, 2006, 2008, 2010). We crosschecked the results with studies by Hanushek and Wößmann (2006), Brunello and Checchi (2007), Waldinger (2007), and Ruhose and Schwerdt (2016). There were few discrepancies regarding the grade and age at which students are tracked between previous studies and between previous studies and our own review. Where deviations arose, we followed our own criteria, which were mainly based on the country reports in UNESCO-IBE (2007, 2012).

Based on the information on the tracking grade and age, we constructed two different variables to determine whether students were tracked at the time of testing

in the secondary school assessments (early tracking) or whether they were still in compulsory schooling (late tracking). In the analyses with TIMSS population B data, we used information on whether students were tracked in eighth grade. For analyses with PISA, we used the grade with most 15-year-old students (ninth or tenth grade in most countries). Due to this classification, 17 countries were classified as early tracking countries in analyses using PISA and 13 countries in analyses using TIMSS. Table 2 depicts the number of overall, early, and late tracking countries in each study pair. On average, each study pair contained 26 countries. About one-fourth of these were early tracking countries. Appendix 1 shows the tracking status for all countries in our sample.

4.3 Analyses

No country had a tracked primary school system, but some countries had tracked secondary school systems. This enabled us to compare educational measures of countries with and without early between-school tracking at the secondary school level while using the same educational measures at the primary school level as a baseline.

Identification strategy Simple comparisons of early and late tracking countries may be biased because the observed differences may have existed before the students were tracked. In such cases, differences between early and late tracking countries would not reflect the effect of tracking but rather of other features of the educational system or differences in the social structure. Indeed, Table 1 shows that early and late tracking countries had different baseline inequalities at the primary school level. On average, early tracking countries showed higher performance levels, lower levels of dispersion inequality and educational inadequacy, and higher social achievement gaps in comparison to late tracking countries.

Following Hanushek and Wößmann (2006), we estimated difference-in-differences models to control for any disparities between early and late tracking countries that existed prior to tracking. The basic idea was to relate differences in educational outcomes—for instance, dispersion inequality at the primary and secondary school levels—to differences in the tracking status at the primary and secondary school levels. For this purpose, we estimated models in which we regressed educational outcomes Y in secondary school s, in country j (Y_{sj}) on a dummy variable that indicated whether the country had a tracked secondary school system (Z_{sj}) while controlling for educational outcomes at the primary school level (Y_{pj}):

$$Y_{sj} = \alpha + \beta_1 Y_{pj} + \gamma Z_{sj} + e_j \quad (1)$$

The key parameter of interest in Eq. (1) was γ, since it estimates the effect of early tracking on the educational outcome. The equation does not include the tracking status at the primary school level because no country in our sample had a tracked primary school system.

We estimated separate models for the four educational outcome measures—dispersion inequality, social achievement gaps, educational inadequacy, and the performance level. The total number of replications for each outcome was 45 including nine replications for reading, 18 for mathematics, and 18 for science (cf. Fig. 1 and Table 2).

Table 2 Number of countries in the overall country sample and divided by the tracking status in the 45 study pairs in the three achievement domains

Primary school level data				Secondary school level data		Overall sample	Early tracking	Late tracking
						N	*N*	*N*
Reading								
1	PIRLS	2001	↔	PISA	2000	21	7	14
2	PIRLS	2001	↔	PISA	2003	18	7	11
3	PIRLS	2001	↔	PISA	2006	23	8	15
4	PIRLS	2006	↔	PISA	2006	24	8	16
5	PIRLS	2006	↔	PISA	2009	29	10	19
6	PIRLS	2006	↔	PISA	2012	26	9	17
7	PIRLS	2011	↔	PISA	2012	32	10	22
8	PIRLS	2011	↔	PISA	2015	35	11	24
9	PIRLS	2016	↔	PISA	2015	33	11	22
Mathematics								
10	TIMSS Pop. A	1995	↔	PISA	2000	19	6	13
11	TIMSS Pop. A	2003	↔	PISA	2003	12	3	9
12	TIMSS Pop. A	2003	↔	PISA	2006	14	3	11
13	TIMSS Pop. A	2003	↔	PISA	2009	16	4	12
14	TIMSS Pop. A	2007	↔	PISA	2009	25	8	17
15	TIMSS Pop. A	2007	↔	PISA	2012	24	8	16
16	TIMSS Pop. A	2011	↔	PISA	2012	34	11	23
17	TIMSS Pop. A	2011	↔	PISA	2015	34	11	23
18	TIMSS Pop. A	2015	↔	PISA	2015	33	11	22
19	TIMSS Pop. A	1995	↔	TIMSS Pop. B	1995	26	6	20
20	TIMSS Pop. A	1995	↔	TIMSS Pop. B	1999	18	4	14
21	TIMSS Pop. A	2003	↔	TIMSS Pop. B	2003	27	4	23
22	TIMSS Pop. A	2003	↔	TIMSS Pop. B	2007	21	2	19
23	TIMSS Pop. A	2007	↔	TIMSS Pop. B	2007	32	3	29
24	TIMSS Pop. A	2007	↔	TIMSS Pop. B	2011	27	2	25
25	TIMSS Pop. A	2011	↔	TIMSS Pop. B	2011	37	2	35
26	TIMSS Pop. A	2011	↔	TIMSS Pop. B	2015	34	3	31
27	TIMSS Pop. A	2015	↔	TIMSS Pop. B	2015	35	4	31
Science								
28	TIMSS Pop. A	1995	↔	PISA	2000	19	6	13
29	TIMSS Pop. A	2003	↔	PISA	2003	12	3	9
30	TIMSS Pop. A	2003	↔	PISA	2006	14	3	11
31	TIMSS Pop. A	2003	↔	PISA	2009	16	4	12
32	TIMSS Pop. A	2007	↔	PISA	2009	25	8	17
33	TIMSS Pop. A	2007	↔	PISA	2012	24	8	16
34	TIMSS Pop. A	2011	↔	PISA	2012	34	11	23
35	TIMSS Pop. A	2011	↔	PISA	2015	34	11	23
36	TIMSS Pop. A	2015	↔	PISA	2015	33	11	22
37	TIMSS Pop. A	1995	↔	TIMSS Pop. B	1995	26	6	20
38	TIMSS Pop. A	1995	↔	TIMSS Pop. B	1999	18	4	14

Table 2 (continued)

	Primary school level data			Secondary school level data		Overall sample *N*	Early tracking *N*	Late tracking *N*
39	TIMSS Pop. A	2003	↔	TIMSS Pop. B	2003	27	4	23
40	TIMSS Pop. A	2003	↔	TIMSS Pop. B	2007	21	2	19
41	TIMSS Pop. A	2007	↔	TIMSS Pop. B	2007	32	3	29
42	TIMSS Pop. A	2007	↔	TIMSS Pop. B	2011	27	2	25
43	TIMSS Pop. A	2011	↔	TIMSS Pop. B	2011	37	2	35
44	TIMSS Pop. A	2011	↔	TIMSS Pop. B	2015	34	3	31
45	TIMSS Pop. A	2015	↔	TIMSS Pop. B	2015	35	4	31

For every study pair in the rows, the number of countries in the overall sample, in the sample of early tracking, and in the sample of late tracking countries are depicted. Populations A and B are abbreviated as Pop. A and Pop. B

Synthesis of effects We computed weighted mean effect sizes to summarize the $i = 45$ estimations per dependent variable. For this purpose, we used the formulas that Card (2012) developed for use in meta-analyses. The basic idea is that some effect estimates are more reliable than others (e.g., due to differences in the sample size), which is reflected in different standard errors. For this reason, the inverse value of the squared standard error (SE_i^2) serves as a weight (w_i) for the corresponding effect estimate. This means that datasets with less efficient results will have a lower weight in the synthesized results:

$$w_i = \frac{1}{SE_i^2} \tag{2}$$

We estimated a weighted mean of the single effects, consisting of the sum of the effect sizes (ES_i) multiplied by their weights (w_i), divided by the total sum of weights:

$$\overline{ES} = \frac{\sum(w_i * ES_i)}{\sum w_i} \tag{3}$$

The weights can be used to compute a standard error for the mean effect size (SE_{ES}). For this purpose, we used the square root of the inverse value of the sum of the weights:

$$SE_{\overline{ES}} = \sqrt{\frac{1}{\sum w_i}} \tag{4}$$

The ratio of the mean effect size and its standard error follows a normal distribution, which can be used to test if the mean effect differs significantly from zero (Card 2012).

5 Results

The results for the different study pairs and the four outcome variables—dispersion, inequality, social achievement gaps, educational inadequacy, and performance level—

are depicted in Fig. 2. Panel a shows, for example, the regression coefficients of the effects of early tracking on dispersion inequality along with the 95% confidence intervals for each of the 45 combinations of primary and secondary school data. Since each estimate was based on a rather small sample of countries, the confidence intervals were large and only few estimates differed significantly from zero. Correspondingly, we also observed large confidence intervals for the results of the other outcomes in panels b, c, and d. In panel b, the estimates were only statistically significantly different from zero in seven out of 45 analyses due to the small sample size of countries.

The low precision of the estimation of the difference-in-differences models made it difficult to draw robust conclusions based on a single pair of primary and secondary school data. However, the replications were based on 45 different combinations and the findings revealed some interesting patterns. For dispersion inequality and social achievement gaps, the large majority of the parameters were positive. For educational inadequacy and performance levels, we observed no overall tendency since roughly half of the estimates were positive and the other half negative.

5.1 Mean effects of early tracking on inequalities

We applied a meta-analytical strategy to combine the effect estimations of different study pairs for each of the four outcomes of interest. Table 3 (column 1) shows the synthesized mean effect across all achievement domains, which was based on all 45 study pairs. The results showed that early tracking increased the three educational inequality measures. The effects were particularly pronounced for the social achievement gap, followed by dispersion inequality. The effect of early tracking on educational inadequacy was small but statistically significant. In contrast to the consistent findings that tracking increased inequality, our study provided no evidence that tracking affected the performance level.

In detail, our analyses showed that early tracking significantly increased dispersion inequality by 2.91 score points ($p < .001$). While there was a general trend of dispersion inequality increasing from the primary to secondary school level, the increase was significantly larger in early tracking countries in comparison to late tracking countries. The outcome measure of dispersion inequality—the standard deviation of test scores at the secondary school level—had an international mean of 89.02 with a SD of 11.51 (see Table 1). We used this information to compute the standardized effect size measure Cohen's d. The standardized effect of tracking on dispersion inequality was $d = 0.25$.

We also found strong evidence that tracking increased the social achievement gap. Tracking increased the gap between students from families with few and with many books by 6.90 score points ($p < .001$), which corresponds to an effect size of $d = 0.44$. Therefore, the social achievement gaps widened more between primary and secondary school in early tracking countries than in late tracking ones.

The mean effect of tracking on educational inadequacy was 0.88 points ($p < .01$). This suggests that early tracking increased the share of students who did not reach basic literacy cutoffs by roughly 1%. In comparison to the other concepts of inequality, the standardized effect $d = 0.07$ is rather small.

In contrast to the results for the three inequality measures, our analyses provided no evidence for an effect of early tracking on the performance level. The mean effect was − 1.00 ($d = 0.02$) and did not differ significantly from zero ($p > .05$).

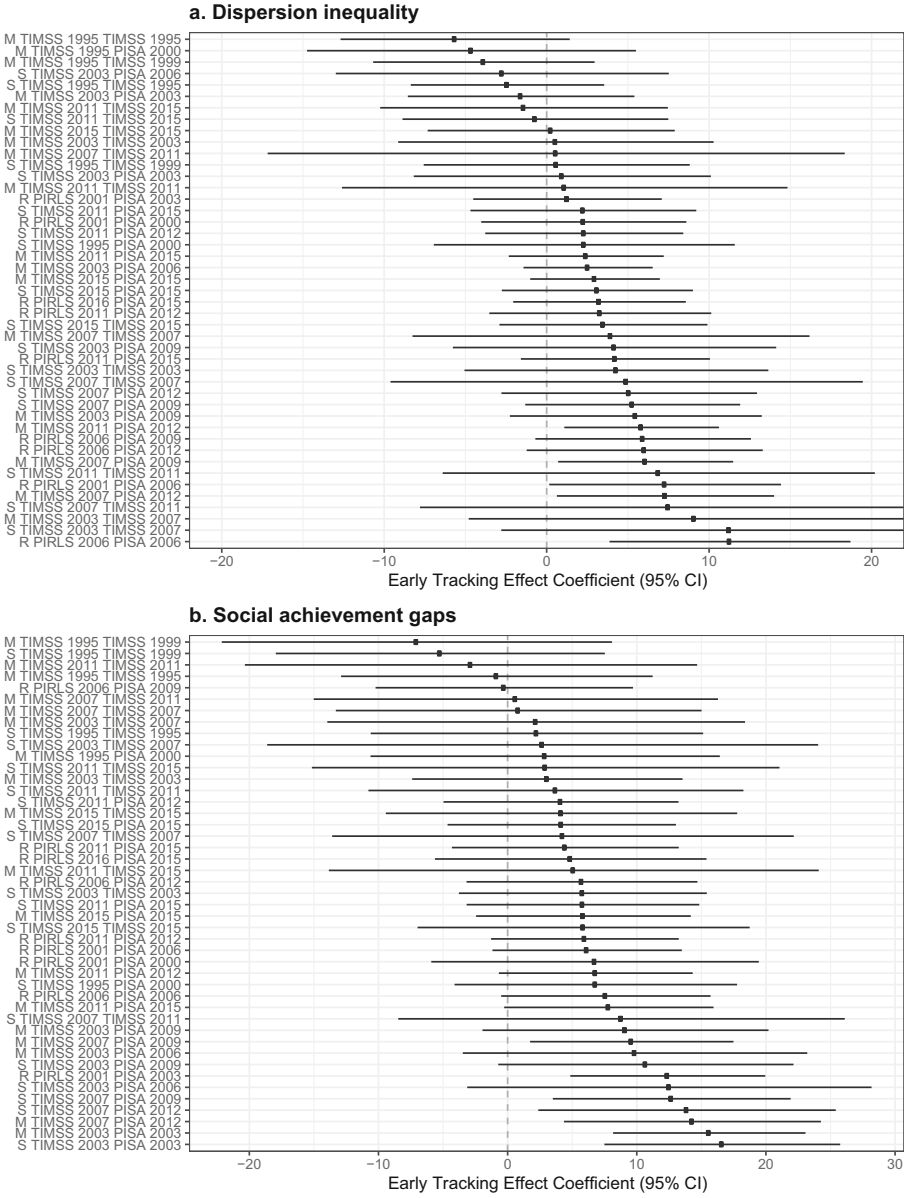

Fig. 2 Individual estimates of the effects of early tracking on the four dependent variables in the 45 study pairs. The single estimates of the early tracking effect on the four outcome variables are depicted for 45 study pairs per outcome. CI is short for confidence interval. In the 45 study pair abbreviations, R stands for reading, M for mathematics, and S for science. In the pair labels, the primary school level dataset is followed by the corresponding secondary school dataset

In the main analyses, we used the inverse standard error to weight each study pair by the precision of its estimate. An alternative approach is to weight each study pair equally. To test the sensitivity of our analyses, we replicated all analyses with equal weights (see Appendix 2). The results remained qualitatively the same.

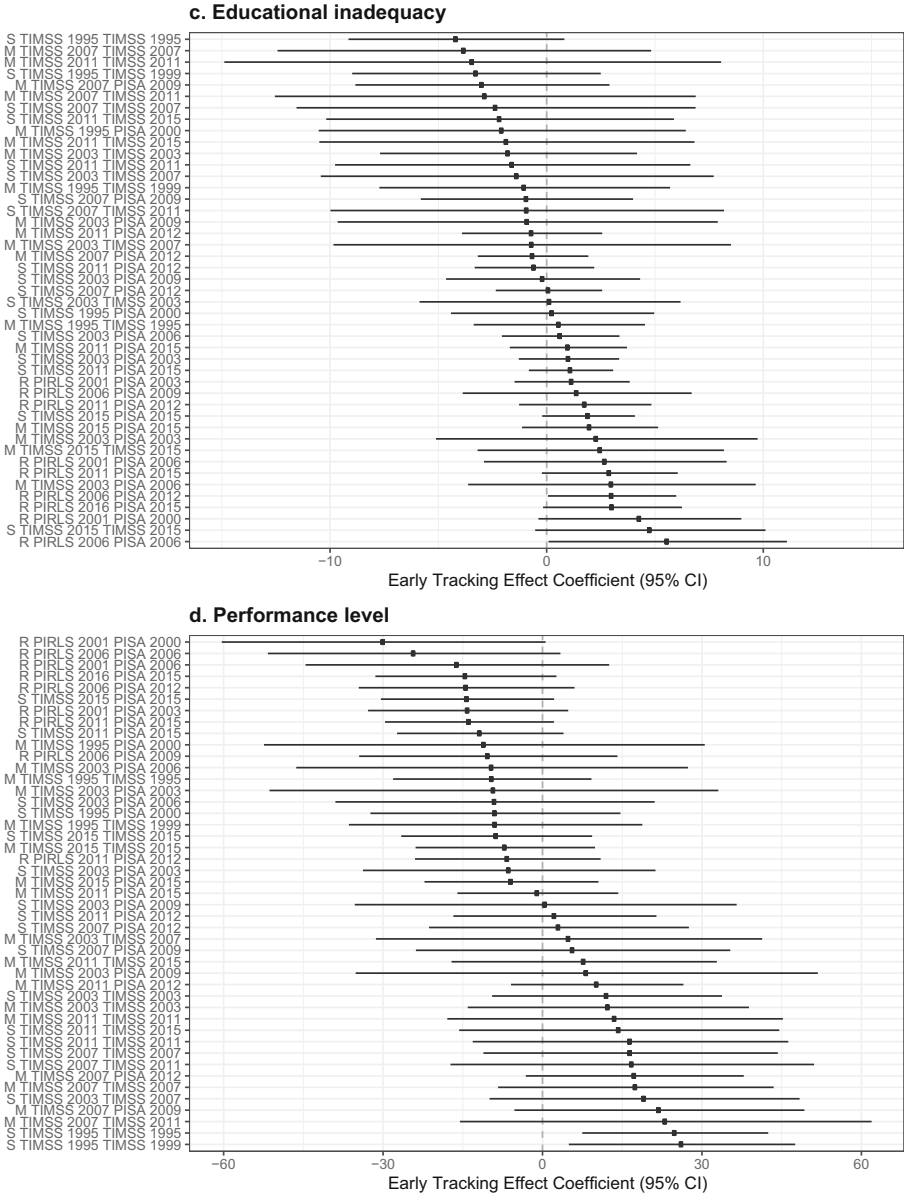

Fig. 2 (continued)

5.2 Further analyses

The review of previous research revealed rather inconsistent findings. We assumed that the small number of countries in each study might be an explanation for the variation in the previously reported findings. An alternative explanation pertains to substantive differences. Our attempt to address this controversy entailed replicating the analyses for different educational outcomes

Table 3 Synthesis of the Effects of Early Tracking ($\overline{\gamma}$) on the Four Dependent Variables for All Domains and Divided by Domain

	(1) All domains		(2) Reading		(3) Mathematics		(4) Science	
	$\overline{\gamma}$	*SE*	$\overline{\gamma}$	*SE*	$\overline{\gamma}$	*SE*	$\overline{\gamma}$	*SE*
Dispersion inequality	2.908***	0.534	4.551***	1.086	2.328**	0.787	2.473*	0.978
Social achievement gap	6.904***	0.796	6.399***	1.466	6.982***	1.293	7.271***	1.393
Educational inadequacy	0.881**	0.298	2.559***	0.606	0.084	0.575	0.490	0.425
Performance level	− 1.002	1.714	− 14.147***	3.500	2.948	2.822	3.330	2.739
N countries	75		45		71		71	
N early tracking countries	17		14		14		14	
N study pairs	45		9		18		18	

The unstandardized parameter γ $\overline{\gamma}$ reflects the synthesized mean effect of early tracking. Significance levels indicated by * $p < .05$, ** $p < .01$, and *** $p < .001$

using the same data. Additionally, we conducted a series of alternative specifications to test the robustness of our main analyses.

Effect heterogeneity across achievement domains In order to test whether tracking affected outcomes in reading, mathematics, and science differently, we replicated the analyses for the three domains separately. The results are depicted in Table 3 (columns 2–4). The findings largely confirmed those of the main specification. Tracking increased the dispersion inequality and the social achievement gap consistently and significantly in all three domains. Furthermore, the analyses on reading suggested that tracking reinforced educational inadequacy and decreased the performance level. We observed no significant effects for educational inadequacy and performance level in mathematics and science. However, only nine study pairs were available to investigate effects in the reading domain, while 18 pairs were available for the estimation of the effects in mathematics and science. For this reason, we suggest that the findings for reading should not be over interpreted.

Alternative inequality measures Different measures of dispersion inequality, social achievement gaps, and educational inadequacy were used in previous research. In our main analyses, we focused on one measure for each educational outcome. To check the robustness of our analyses, we used alternative measures of educational inequality and replicated the analyses for the same 45 study pairs of primary and secondary school data. In Table 4, each row contains the result of an alternative specification.

We used the within-country standard deviation of the test scores as the measure of dispersion inequality in the main analyses. In additional robustness checks, we used the range between the 95th and 5th percentile and between the 75th and 25th percentile of the achievement distribution as alternative measures of dispersion inequality. We observed that early tracking also increased the dispersion inequality in these alternative specifications (rows 1–2 in Table 4).

The social achievement gap was operationalized as the mean score difference between children from households with less than 100 and at least 100 books (absolute

Table 4 Robustness checks of the synthesis of the effects of early tracking ($\overline{\gamma}$) for all domains in the 45 study pairs

	$\overline{\gamma}$	*SE*
Alternative inequality measures		
Dispersion inequality		
1 Range between 95th and 5th percentile	8.943***	1.734
2 Range between 75th and 25th percentile	5.435***	0.810
Social achievement gap		
3 Relative gap depending on the number of books	0.054***	0.008
4 Absolute gap depending on parental education	5.099***	1.086
Educational inadequacy		
5 Intermediate benchmark resp. level 2 thresholds	1.475**	0.563
Tracking as nondichotomous		
6 Dispersion inequality	0.765***	0.149
7 Social achievement gap	2.372***	0.213
8 Educational inadequacy	0.112	0.081
9 Performance level	0.230	0.482

The unstandardized parameter γ $\bar{\gamma}$ reflects the synthesized mean effect of early tracking. Eight of the nine analyses were based on 75 countries including 17 early tracking countries and overall 45 study pairs. The analysis in row 4 on the absolute social achievement gap using parental education as an indicator was based on 67 countries including 17 early tracking countries and a total of 21 study pairs. Significance levels indicated by $^{*}p < .05$, $^{**}p < .01$, and $^{***}p < .001$

difference). In further analyses, we standardized this difference by the respective within-country standard deviation (relative difference) and used this variable as an alternative outcome. The scale of the effect changed due to the standardization but it remained significant ($p < .001$) (row 3 in Table 4). The number-of-books-at-home variable is probably the most commonly used measure of the socioeconomic status in comparative research. It is, however, often criticized, for example because certain student groups tend to systematically underestimate the number of books at home (e.g., Engzell 2019). For this reason, we replicated the main analyses with another frequently used measure of socioeconomic background, namely parental education. The additional analysis replicated the finding that tracking significantly increased the absolute gap between children of parents with and without tertiary education (row 4 in Table 4).

The threshold that defines educational inadequacy can be a more or less inclusive cutoff. In our main analyses, we identified a little more than 10% of the students as having an inadequate level of achievement. In further analyses, we used the intermediate benchmark in PIRLS and TIMSS and proficiency level 2 in PISA instead. This lead us to identify about 30% of the students as failing to attain an adequate level of achievement. The replicated analyses showed that early tracking increased the proportion of students not reaching the TIMSS intermediate benchmark or level 2 in PISA by about 1.5% ($p < .01$; row 5 in Table 4).

Tracking as nondichotomous Just as in most previous research, we used a binary tracking indicator in our main analyses. In further analyses, we replaced this binary

indicator with a continuous variable for the tracking grade to exploit the variation in how many years students were exposed to a tracked school system (see Appendix 1). A value of zero means that a country had a comprehensive secondary school system at the secondary school level when testing occurred, and values between 1 and 5 imply that students were allocated to different ability tracks one to five grades before the secondary school assessment was administered. However, a drawback of this approach was the limited number of countries tracking students at different times. We replicated the main analyses for all four outcomes using the nondichotomous tracking indicator.

The analyses for the three types of inequalities and the performance levels are presented in rows 6–9 in Table 4. The effects of the tracking grade on the dispersion inequality and social achievement gaps were positive and significant ($p < .001$). One extra year of exposure to a tracked system increased the countries' standard deviations of achievement scores by about 0.77 points and the social achievement gap by 2.37 points. These findings imply that postponing tracking by 5 years—for example, from tracking after fourth to tracking after eighth grade—reduced the dispersion inequality by 3.85 points and social achievement gaps by 11.95 points. Consistent with the main results, the effect on educational inadequacy was smaller and, in this case, nonsignificant. Just as in the main analyses, we observed no significant effects for the performance level.

6 Discussion and conclusion

For a long time, the controversy around between-school ability tracking was mainly ideological. Robust empirical evidence on the effects of tracking on student outcomes was rare. However, in the past 15 years, a number of studies with robust designs have been conducted with the aim of contributing empirical evidence to the discussion about the effects of tracking on student learning. Most of the new studies used international data to compare student achievement in countries with tracked versus comprehensive school systems while controlling for prior achievement differences (e.g., Hanushek and Wößmann 2006). While these studies applied sound strategies to identify the effects of tracking on achievement, most suffered from the limitation that international analyses are based on relatively small samples of countries. Furthermore, it was difficult to synthesize previous research because different studies focused on different educational outcomes. Against this backdrop, the main aim of the present study was to use the data accumulated in international assessments to systematically investigate the effects of tracking on educational inequalities and performance levels. Previous research used different data to investigate the effects of tracking on different outcomes. We used the same data to study multiple outcomes.

6.1 Summary of key findings

The literature frequently refers to a perceived trade-off between equity and efficacy in the field of between-school tracking. While previous research was inconclusive, we found strong evidence that tracking increased dispersion inequality and social

achievement gaps. Tracking was also associated with educational inadequacy, but the evidence was less robust. In contrast, we found no evidence that tracking boosted performance levels. These main findings were very consistent across different model specifications. We replicated the analyses using different tracking indicators and outcome measures, and the general results confirmed our main findings.

6.2 Conceptual clarity: different outcomes, different findings

We found that the effects of early tracking on educational inequality varied according to the *theoretical concept* behind the inequality measures; this was confirmed by the series of further analyses on the robustness of our findings. It is worth remembering that our results varied between *different concepts of inequality* but they were very similar for the *same concepts of inequality*. The clearest effect was on social achievement gaps, where the effect of tracking seemed to be the most pronounced across all domains and for different measures of student background. This is of particular relevance since it contradicts the argument that tracking is meritocratic; if it were meritocratic, the inequality determined by social characteristics would not vary. This point is reinforced when looking at the effects of tracking on dispersion inequality: Early tracking increased the dispersion of achievement scores, but compared to the effects on social achievement gaps, this was not as relevant to the overall existing dispersion. Finally, looking at the educational inadequacy, we found more inconclusive evidence. We observed significant effects for tracking on educational inadequacy in reading but not in the two other domains. On the other hand, the overall effects and the alternative specification with a more ambitious threshold revealed significant effects of an increase of the proportion of students not reaching minimum levels of achievement. This means that tracking did not help the most disadvantaged students. At worst, these would perform better without tracking, while, at best, tracking does not have discernible effects.

We contrasted the analysis on educational inequalities with analyses on the effect of tracking on performance levels. In line with previous studies, we found no evidence that tracking increased performance levels. If anything, there was some evidence for tracking *decreasing* performance levels in reading.

6.3 The reproducibility of findings: the issue of a small sample size at the country level

Our study illustrates that the reproducibility of research findings based on international data is limited. We observed a remarkable variability in results between different combinations of primary and secondary school assessments. For this reason, it comes as no surprise that previous research was inconsistent and sometimes even contradictory. International assessments collect information from millions of students, but, at the country level, the number of units of analysis is small. Small samples are generally associated with large standard errors, which means that research findings based on a single international assessment are unreliable. Our findings should encourage researchers to replicate analyses based on data from different international assessments or to combine different assessments to reduce publication bias and establish reliable evidence.

6.4 Limitations of the study

The first limitation is related to the need to simplify the tracking variable itself. As Hillmert and Jacob (2010) noted, studies on transitions in educational careers follow an ideal-typical sequence of transitions and phases in education, while students can and do follow more complex paths in reality. In line with previous research, we used a binary tracking indicator, but we are well aware that between-school tracking can take different forms simultaneously. Following this, our analyses are well suited to detecting the effects of between-school tracking, which is our research question, but they do not account for every form of selection. Studying different types of within-school tracking (both whole-class differentiation and on a course-by-course basis) is beyond the scope of our study (see Chmielewski 2014; Chmielewski et al. 2013). On the other hand, we suspect that if we had been able to measure within-school tracking, the effects would have been even more pronounced. Let us assume, for the moment, that there is a continuum along the distinction between comprehensive, within-school tracking, and between-school tracking systems. In the present study, we regarded within-school tracking countries as comprehensive school systems. This means that our estimates are rather conservative and that the effects would have been even larger if we had considered countries that applied within-school tracking as a separate category.

6.5 Policy implications

When discussing the consequences of between-school tracking, it is useful to revisit the debate on what types of inequality are considered acceptable or even desirable and what types are considered problematic and unjust. With respect to the frequently perceived tradeoff between efficiency and equity, it is important to stress that we did not find any evidence supporting the suggestion that early between-school tracking increases average performance levels. Regarding the question of what types of inequality are acceptable, different perspectives have to be considered. In modern societies, it is generally accepted that performance levels vary between students (inequality as dispersion) and that this mix of skillsets is even needed because the labor market demands differently skilled workers. At the same time, it is more difficult to justify social inequalities, i.e., the idea that children get different opportunities based on their social background and not their educational potential. In the same vein, it is difficult to find arguments supporting educational inadequacy, i.e., the notion that a proportion of students would not even reach the basic performance levels that are necessary to participate in the society and in all parts of the labor market. Therefore, we regard social achievement gaps and educational inadequacy as particularly important outcome measures of educational policies.

Hanushek and Wößmann (2006, p. C75) closed their study with the following statement: “From a policy perspective, it seems incumbent on those advocating early tracking in schools to identify the potential gains from this. These preliminary results suggest that countries lose in terms of the distribution of outcomes, and possibly also in levels of outcomes, by pursuing such policies.” More than 10 years later, with a larger amount of evidence, we have come to a similar conclusion. If we had to make a policy recommendation, it would be to reform early between-school tracking systems into comprehensive school systems.

Acknowledgments The authors would like to thank Roisin Cronin for copy editing the manuscript and Robin Grugel for his help on the analyses.

Funding Open Access funding enabled and organized by Projekt DEAL. Andrés Strello, Rolf Strietholt, and Isa Steinmann are part of the European Training Network OCCAM. This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant number 765400.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

Table 5 School Tracking Status According to Age and Grade in All Countries and Regions

Country	Tracking age	Tracking grade	Early tracking country in PISA	Early tracking country in TIMSS
Abu Dhabi, UAE	15	9		
Alberta, Canada	18	12		
Algeria	15.5	9		
Argentina	15	9		
Armenia	15	9		
Australia	16	10		
Austria	10	4	X	X
Bahrain	15	9		
Belgium (Flem. Gem.)	12	6	X	X
British Columbia, Canada	18	12		
Buenos Aires, Argentina	12	6	X	X
Bulgaria	14	7	X	X
Canada	18	12		
Chile	16	10		
Colombia	15	9		
Croatia	15	8		
Cyprus	15	9		
Czech Republic	11	5	X	X
Denmark	16	10		
Dubai, UAE	15	9		
El Salvador	16	9		
England	16	11		
Finland	16	9		
France	15	9		
Georgia	15	9		
Germany	10	4	X	X

Table 5 (continued)

Country	Tracking age	Tracking grade	Early tracking country in PISA	Early tracking country in TIMSS
Greece	15	9		
Hong Kong	16	11		
Hungary	10	4	X	X
Iceland	16	10		
Indonesia	16	9		
Iran	15	9		
Ireland	12	6	X	X
Israel	15	10		
Italy	14	8	X	
Japan	15	9		
Kazakhstan	15	9		
Korea	14	9	X	
Kuwait	18	12		
Latvia	16	9		
Lithuania	15	8		
Luxembourg	12	6	X	X
Macedonia	15	8		
Malta	16	11		
Moldova	15	10		
Mongolia	16	8		
Morocco	15	9		
Netherlands	12	6	X	X
New Zealand	16	11		
Norway	16	10		
Oman	16	10		
Ontario, Canada	18	12		
Philippines	16	10		
Poland	15	9		
Portugal	15	9		
Qatar	15	9		
Quebec, Canada	18	12		
Romania	14	8	X	
Russian Federation	15	9		
Saudi Arabia	15	9		
Scotland	16	11		
Serbia	15	8		
Singapore	12	6	X	X
Slovakia	10	4	X	X
Slovenia	15	9		
Spain	15	9		
Sweden	16	9		

Table 5 (continued)

Country	Tracking age	Tracking grade	Early tracking country in PISA	Early tracking country in TIMSS
Taiwan	15	9		
Thailand	15	9		
Trinidad and Tobago	11	5	X	X
Tunisia	16	10		
Turkey	14	8	X	
Ukraine	15.5	9		
United Arab Emirates	15	9		
United States	18	12		

Tracking age reflects the mode age in the grade when tracking takes place. Tracking age and grade describe the year of the first school differentiation in each country or region. Sources: UNESCO-IBE (2007, 2012), Eurydice (2005, 2011, 2013a, b, 2014), and OECD reports (2004, 2006, 2008, 2010)

Table 6 Synthesis of the effects of early tracking on the four dependent variables for all domains in unweighted analyses

	(1) All domains
Dispersion inequality	2.996
Social achievement gap	5.775
Educational inadequacy	0.168
Performance level	0.736
N countries	75
N early tracking countries	17
N study pairs	45

The unstandardized parameter reflects the synthesized mean effect of early tracking. The analyses were equivalent to the main analyses but incorporated equal weights for all countries and cycles

References

Ammermüller, A. (2005). Educational opportunities and the role of institutions. ftp://ftp.zew.de/pub/zew-docs/dp/dp0544.pdf

Batruch, A., Autin, F., Bataillard, F., & Butera, F. (2018). School selection and the social class divide: How tracking contributes to the reproduction of inequalities. *Personality and Social Psychology Bulletin, 45*(3), 1–14. https://doi.org/10.1177/0146167218791804.

Bauer, P., & Riphahn, R. T. (2006). Timing of school tracking as a determinant of intergenerational transmission of education. *Economics Letters, 91*(1), 90–97. https://doi.org/10.1016/j.econlet.2005.11.003.

Becker, M., Lüdtke, O., Trautwein, U., Köller, O., & Baumert, J. (2012). The differential effects of school tracking on psychometric intelligence: Do academic-track schools make students smarter? *Journal of Educational Psychology, 104*(3), 682–699. https://doi.org/10.1037/a0027608.

Blanchard, R. D., Bunker, J. B., & Wachs, M. (1977). Distinguishing aging, period and cohort effects in longitudinal studies of elderly populations. *Socio-Economic Planning Sciences, 11*(3), 137–146. https://doi.org/10.1016/0038-0121(77)90032-5.

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

Brighouse, H., & Swift, A. (2008). Putting educational equality in its place. *Education Finance and Policy, 3*(4), 444–466. https://doi.org/10.1162/edfp.2008.3.4.444.

Brighouse, H., & Swift, A. (2009). Educational equality versus educational adequacy: A critique of Anderson and Satz. *Journal of Applied Philosophy, 26*(2), 117–128. https://doi.org/10.1111/j.1468-5930.2009.00438.x.

Brighouse, H., Ladd, H. F., Loeb, S., & Swift, A. (2018). *Educational goods*. Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226514208.001.0001.

Card, N. A. (2012). *Applied meta-analysis for social science research*. New York: The Gilford Press.

Chmielewski, A. K. (2014). An international comparison of achievement inequality in within- and between-school tracking systems. *American Journal of Education, 120*(3), 293–324. https://doi.org/10.1086/675529.

Chmielewski, A. K., Dumont, H., & Trautwein, U. (2013). Tracking effects depend on tracking type: An international comparison of students' mathematics self-concept. *American Educational Research Journal, 50*(5), 925–957. https://doi.org/10.3102/0002831213489843.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). Equality of educational opportunity. Washington D.C.

Dämmrich, J., & Triventi, M. (2018). The dynamics of social inequalities in cognitive-related competencies along the early life course—A comparative study. *International Journal of Educational Research, 88*, 73–84. https://doi.org/10.1016/j.ijer.2018.01.006.

Dollmann, J. (2019). Educational institutions and inequalities in educational opportunities. In R. Becker (Ed.), *Research Handbook on the Sociology of Education* (pp. 268–283). doi:https://doi.org/10.4337/9781788110426.00025.

Duru-Bellat, M., & Suchaut, B. (2005). Organisation and context, efficiency and equity of educational systems: What PISA tells us. *European Educational Research Journal, 4*(3), 181–194. https://doi.org/10.2304/eerj.2005.4.3.3.

Engzell, P. (2019). What do books in the home proxy for? A cautionary tale. *Sociological Methods and Research*, 1–28. doi:https://doi.org/10.1177/0049124119826143.

Eurydice. (2005). Key data on education in Europe 2005. *Reproduction.* http://www.indire.it/lucabas/lkmw_file/eurydice/Key_Data_2005_EN.pdf.

Eurydice. (2011). The structure of the European education systems 2011/12: Schematic diagrams. https://publications.europa.eu/en

Eurydice. (2013a). The structure of the European education systems 2012/13: Schematic diagrams. doi:https://doi.org/10.2797/40560

Eurydice. (2013b). The structure of the European education systems 2013/14: Schematic diagrams. doi:https://doi.org/10.2797/206797

Eurydice. (2014). The structure of the European education systems 2014/15: Schematic diagrams. doi:https://doi.org/10.2797/607957

Gorard, S., & Smith, E. (2004). An international comparison of equity in education systems. *Comparative Education, 40*(1), 15–28. https://doi.org/10.1080/0305006042000184863.

Guill, K., Lüdtke, O., & Köller, O. (2017). Academic tracking is related to gains in students' intelligence over four years: Evidence from a propensity score matching study. *Learning and Instruction, 47*, 43–52. https://doi.org/10.1016/j.learninstruc.2016.10.001.

Guyon, N., Maurin, E., & McNally, S. (2012). The effect of tracking students by ability into different schools: A natural experiment. *Journal of Human Resources, 47*(3), 684–721. https://doi.org/10.3368/jhr.47.3.684.

Hallinan, M. T. (1994). Tracking: From theory to practice. *Sociology of Education, 67*(2), 79–84. https://doi.org/10.2307/2112697.

Hanushek, E. A. (2013). Economic growth in developing countries: The role of human capital. *Economics of Education Review, 37*, 204–212. https://doi.org/10.1016/j.econedurev.2013.04.005.

Hanushek, E. A., & Wößmann, L. (2006). Does early tracking affect educational inequality and performance? Differences-in-differences evidence across countries. *Economic Journal, 116*(115), C63–C76. https://doi.org/10.1111/j.1468-0297.2006.01076.x.

Hanushek, E. A., Schwerdt, G., Wiederhold, S., & Wößmann, L. (2015). Returns to skills around the world: Evidence from PIAAC. *European Economic Review, 73*, 103–130. https://doi.org/10.1016/j.euroecorev.2014.10.006.

Hillmert, S., & Jacob, M. (2010). Selections and social selectivity on the academic track: A life-course analysis of educational attainment in Germany. *Research in Social Stratification and Mobility, 28*(1), 59–76. https://doi.org/10.1016/j.rssm.2009.12.006.

Holm, A., Jæger, M. M., Karlson, K. B., & Reimer, D. (2013). Incomplete equalization: The effect of tracking in secondary education on educational inequality. *Social Science Research, 42*(6), 1431–1442. https://doi.org/10.1016/j.ssresearch.2013.06.001.

Horn, D. (2009). Age of selection counts: A cross-country analysis of educational institutions. *Educational Research and Evaluation, 15*(4), 343–366. https://doi.org/10.1080/13803610903087011.

Horn, D. (2013). Diverging performances: The detrimental effects of early educational selection on equality of opportunity in Hungary. *Research in Social Stratification and Mobility, 32*(1), 25–43. https://doi.org/10.1016/j.rssm.2013.01.002.

Huang, M. H. (2009). Classroom homogeneity and the distribution of student math performance: A country-level fixed-effects analysis. *Social Science Research, 38*(4), 781–791. https://doi.org/10.1016/j.ssresearch.2009.05.001.

Jakubowski, M. (2010). Institutional tracking and achievement growth: Exploring difference-in-differences approach to PIRLS, TIMSS, and PISA data. In J. Dronkers (Ed.), *Quality and inequality of education: Cross-national perspectives* (pp. 44–81). Dordrecht: Springer. https://doi.org/10.1007/978-90-481-3993-4_3.

Karlson, K. B. (2015). Expectations on track? High school tracking and adolescent educational expectations. *Social Forces, 94*(1), 115–141. https://doi.org/10.1093/sf/sov006.

Koerselman, K. (2013). Incentives from curriculum tracking. *Economics of Education Review, 32*(1), 140–150. https://doi.org/10.1016/j.econedurev.2012.08.003.

Lange, S., & von Werder, M. (2017). Tracking and the intergenerational transmission of education: Evidence from a natural experiment. *Economics of Education Review, 61*, 59–78. https://doi.org/10.1016/j.econedurev.2017.10.002.

Lavrijsen, J., & Nicaise, I. (2015). New empirical evidence on the effect of educational tracking on social inequalities in reading achievement. *European Educational Research Journal, 14*(3–4), 206–221. https://doi.org/10.1177/1474904115589039.

Lavrijsen, J., & Nicaise, I. (2016). Educational tracking, inequality and performance: New evidence from a differences-in-differences technique. *Research in Comparative and International Education, 11*(3), 334–349. https://doi.org/10.1177/1745499916664818.

Lee, B. (2014). The influence of school tracking systems on educational expectations: A comparative study of Austria and Italy. *Comparative Education, 50*(2), 206–228. https://doi.org/10.1080/03050068.2013.807644.

Lucas, S. R., & Berends, M. (2002). Sociodemographic diversity, correlated achievement, and de facto tracking. *American Sociological Association, 75*(4), 328–348. https://doi.org/10.2307/3090282.

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*(2), 99–106. https://doi.org/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*(4), 483–505. https://doi.org/10.1177/0268580905058328.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). TIMSS 2015 Achievement scaling methodology. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA). https://timssandpirls.bc.edu/publications/timss/2015-methods.html

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2017). PIRLS 2016 achievement scaling methodology. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in PIRLS 2016* (pp. 11.1-11.9).

TIMSS & PIRLS international study center, lynch School of Education, Boston College and International Association for the Evaluation of educational achievement (IEA). https://eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED580352

Martinková, P., Hladká, A., & Potužníková, E. (2020). Is academic tracking related to gains in learning competence ? Using propensity score matching and differential item change functioning analysis for better understanding of tracking implications. *Learning and Instruction, 66*(November 2019), 101286. https://doi.org/10.1016/j.learninstruc.2019.101286.

Micklewright, J., & Schnepf, S. V. (2007). Inequality of learning in industrialised countries. In S. P. Jenkins & J. Micklewright (Eds.), *Inequality and poverty re-examined* (pp. 129–145). Oxford: Oxford Univ. Press.

OECD. (2004). Learning for Tomorrow's world: First results from PISA 2003. *OECD.* https://doi.org/10.1787/9789264006416-en.

OECD. (2006). Education at a glance 2006. *OECD.* https://doi.org/10.1787/eag-2006-en.

OECD. (2008). *PISA* 2006. Volume 2: Data. OECD. doi:https://doi.org/10.1787/9789264040151-en.

OECD. (2010). PISA 2009 results: What makes a school successful? *OECD.* https://doi.org/10.1787/9789264091559-en.

OECD. (2017). *PISA 2015 Technical Report.* https://www.oecd.org/pisa/data/2015-technical-report/

Pietsch, M., & Stubbe, T. C. (2007). Inequality in the transition from primary to secondary school: School choices and educational disparities in Germany. *European Educational Research Journal, 6*(4), 424–445. https://doi.org/10.2304/eerj.2007.6.4.424.

Piopiunik, M. (2014). The effects of early tracking on student performance: Evidence from a school reform in Bavaria. *Economics of Education Review, 42*, 12–33. https://doi.org/10.1016/j.econedurev.2014.06.002.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ, USA: John Wiley & Sons, Inc.. https://doi.org/10.1002/9780470316696.

Ruhose, J., & Schwerdt, G. (2016). Does early educational tracking increase migrant-native achievement gaps? Differences-in-differences evidence across countries. *Economics of Education Review, 52*, 134–154. https://doi.org/10.1016/j.econedurev.2016.02.004.

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In E. A. Hanushek, S. Machin, & L. Wößmann (Eds.), *Handbook of the Economics of Education (Vol. 3, pp. 249–277)*. Elsevier. https://doi.org/10.1016/B978-0-444-53429-3.00004-1.

Schlicht, R., Stadelmann-Steffen, I., & Freitag, M. (2010). Educational inequality in the EU. *European Union Politics, 11*(1), 29–59. https://doi.org/10.1177/1465116509346387.

Schütz, G., Ursprung, H. W., & Wößmann, L. (2008). Education policy and equality of oppurtunity. *KYKLOS, 61*(2), 279–308. https://doi.org/10.1111/j.1467-6435.2008.00402.x.

Skopek, J., Triventi, M., & Buchholz, S. (2019). How do educational systems affect social inequality of educational opportunities? The role of tracking in comparative perspective. In R. Becker (Ed.), *Research Handbook on the Sociology of Education* (pp. 214–232). doi:https://doi.org/10.4337/9781788110426.00022.

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research, 60*(3), 471–499. https://doi.org/10.3102/00346543060003471.

Solga, H. (2014). Education, economic inequality and the promises of the social investment state. *Socio-Economic Review, 12*, 269–297. https://doi.org/10.1093/ser/mwu014.

Strietholt, R. (2014). Studying educational inequality: Reintroducing normative notions. In R. Strietholt, W. Bos, J.-E. Gustafsson, & M. Rosén (Eds.), *Educational Policy Evaluation Through International Comparative Assessments* (pp. 51–58). Münster: Waxmann Verlag.

Strietholt, R., & Borgna, C. (2018). Inequality in educational achievement: Different measures, different conclusions. *Manuscript submitted for publication*.

UNESCO. (2018). Pseudotumor cerebri. In *Handbook on measuring equity in education*. https://doi.org/10.1016/S0733-8619(03)00096-3.

UNESCO-IBE. (2007). World Data on Education: Sixth edition 2006–07. http://www.ibe.unesco.org/en/document/world-data-education-sixth-edition-2006-07

UNESCO-IBE. (2012). World Data on Education: Seventh edition 2010–11. http://www.ibe.unesco.org/en/document/world-data-education-seventh-edition-2010-11

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, *45*(3), 1–67. doi:https://doi.org/10.18637/jss.v045.i03.

Van de Werfhorst, H. G. (2018). Early tracking and socioeconomic inequality in academic achievement: Studying reforms in nine countries. *Research in Social Stratification and Mobility, 58*, 22–32. https://doi.org/10.1016/j.rssm.2018.09.002.

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. https://doi.org/10.1146/annurev.soc.012809.102538.

Van Houtte, M., & Stevens, P. A. J. (2015). Tracking and sense of futility: The impact of between-school tracking versus within-school tracking in secondary education in Flanders (Belgium). *British Educational Research Journal, 41*(5), 782–800. https://doi.org/10.1002/berj.3172.

Waldinger, F. (2007). Does tracking affect the importance of family background on students' test scores? https://www.fabianwaldinger.com/research

Affiliations

Andrés Strello [1] · Rolf Strietholt [1,2,3] · Isa Steinmann [1,4] · Charlotte Siepmann [1]

Rolf Strietholt
rolf.strietholt@tu-dortmund.de

Isa Steinmann
isa.steinmann@tu-dortmund.de

Charlotte Siepmann
charlotte.siepmann@fh-dortmund.de

[1] Center for Research on Education and School Development, TU Dortmund University, Vogelpothsweg 78, 44227 Dortmund, Germany

[2] Department of Education and Special Education, University of Gothenburg, Västra Hamngatan 25, Gothenburg 41117, Sweden

[3] International Association for the Evaluation of Educational Achievement (IEA), Überseering 27, 22297 Hamburg, Germany

[4] Centre for Educational Measurement, University of Oslo, Postboks 1161, Blindern, Oslo 0318, Norway