Technische Universität Dortmund
Fakultät Erziehungswissenschaft, Psychologie und Bildungsforschung

Studying Educational Inequality: Effects of School Differentiation on Multiple Inequalities and School Segregation

Kumulative Dissertation zur Erlangung des akademischen Grades
Doktor der Philosophie (Dr. phil.)

Vorgelegt von Andrés Ignacio Strello Toledo
geboren am 09.11.1991 in Santiago, Chile

Erstgutachter: PD. Dr. Rolf Strietholt
Zweitgutachterin: Prof. Dr. Isa Steinmann

Juni 2023

TU Dortmund University
Department of Educational Sciences and Psychology

Studying Educational Inequality: Effects of School Differentiation on Multiple Inequalities and School Segregation

Cumulative dissertation for the award of the academic degree of
Doctor of Philosophy (Dr. phil.)

Submitted by Andrés Ignacio Strello Toledo
Born on 09.11.1991 in Santiago, Chile

First reviewer: PD. Dr. Rolf Strietholt
Second reviewer: Prof. Dr. Isa Steinmann

June 2023

Dies ist eine Veröffentlichung als Dissertation an der Fakultät Erziehungswissenschaft, Psychologie und Soziologie an der Technischen Universität Dortmund.
Die Arbeit wurde am 22.12.2022 schriftlich eingereicht und am 02.06.2023 in Dortmund mündlich verteidigt.

Acknowledgments

I would like to thank the following people:

To Valentina, for being by my side all these years. You accepted moving with me to the other side of the world without having a clear idea of what we were getting into or where we were going. Without you, this would not have been possible.

To my parents, Susana and Mauricio, who have provided me with unconditional support throughout my life, no matter what path I have taken. Whether I studied music or moved to another country, they were always there, and I will always be grateful for that.

To my supervisors, Rolf and Isa, for their invaluable guidance. I consider myself fortunate to have had two supervisors who were so committed and dedicated. It has been a pleasure to work with both of you these years, and I am confident that we will continue collaborating for a long time. I would also like to mention here Monica Rosén, who was an excellent host during my secondment in Gothenburg.

To the rest of my team in Dortmund: Laura, Robin, and Hannah. Thanks to all of you, I settled well in Dortmund and into life in Germany; you turned the city into a second home.

To all my fellow doctoral students at OCCAM. To my Latin American friends, Andrés, Ana, Edwin, and Elisa. Staying in constant contact with you made life in Europe much nicer, and I feel fortunate to have gained friends in such distant places. To my office mates during my time in Gothenburg, Erika and Leah. There wasn't much sun in those lands, but we had good laughs. To the ones I'm missing, like Pietro, Silvan, or Jelena, with whom I had the pleasure of sharing in our regular meetings.

Finally, to all my friends and family in Chile whom I haven't mentioned. Even from a distance, your support and love were always with me. My occasional visits to Chile allowed me to recharge my batteries and bring back all your affection and energy.

Agradecimientos

Me gustaría agradecer a las siguientes personas:

A Valentina, por haber estado a mi lado todos estos años. Aceptaste mudarte conmigo al otro lado del mundo sin tener muy claro en qué nos estábamos metiendo ni hacia dónde íbamos. Sin ti, esto no hubiera sido posible.

A mis padres, Susana y Mauricio, que me han dado su apoyo incondicional durante toda mi vida, no importando qué camino haya tomado. Ya sea que estudie música o que me vaya a otro país, siempre estuvieron ahí, y de eso siempre estaré agradecido.

A mis supervisores, Rolf e Isa, por su invaluable guía. Me siento afortunado de haber tenido dos supervisores tan comprometidos y dedicados. Ha sido un agrado haber trabajado con ustedes dos estos años y estoy seguro de que seguiremos colaborando por un largo tiempo. Me gustaría mencionar aquí también a Monica Rosén, que fue una excelente anfitriona en mi intercambio doctoral en Gotemburgo.

Al resto de mi equipo en Dortmund: Laura, Robin y Hannah. Gracias a ustedes, me acomodé muy bien en Dortmund y a la vida en Alemania; hicieron de la ciudad un segundo hogar para mí.

A todos mis colegas estudiantes doctorales de OCCAM. A mis amigos latinoamericanos, Andrés, Ana, Edwin y Elisa. Haber estado en contacto permanente con ustedes permitió que la vida en Europa fuera mucho más amigable, y soy afortunado de haber ganado amigos en lugares tan distantes. A mis compañeras de oficina en mi paso por Gotemburgo, Erika y Leah. No había mucho sol en esas tierras, pero tuvimos buenas risas. Al resto que se me va, como Pietro, Silvan o Jelena, con quienes tuve el agrado de compartir en nuestras reuniones periódicas.

Finalmente, a todos mis amigos y familia en Chile que no he mencionado. Aun a la distancia, su apoyo y amor estuvieron siempre acompañándome. Mis visitas ocasionales a Chile me permitieron recargar las baterías y llevarme de vuelta todo su cariño y energía.

The present dissertation is a part of the research done for the European Training Network on "Outcomes and Causal Inference in International Comparative Assessments" (OCCAM). OCCAM received funding from the European Commission's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 765400.

Summary

This dissertation explores the relevance of the measurement of inequality in education, in terms of the theoretical and methodological decisions involved and the consequences of these decisions for research. Specifically, this dissertation investigates the correlations between several measures of inequality in achievement and social categories indicators, exploring how research on the effects of school differentiation is affected by the particular measurement of inequality. The main thesis is that the term *inequality* hides several conceptualizations, each implying different theoretical and normative frameworks with different sets of metrics, leading to different empirical results.

In this dissertation, the conceptualizations and measurements of inequality are divided into three categories: dispersion inequality, social inequality, and adequacy. I also study social segregation across schools, offering it as an alternative outcome to the common focus on achievement. The concept of social inequality is studied in detail, offering insights into the measurement of socioeconomic inequality, and the relationship between socioeconomic, immigration, and gender inequalities in achievement.

This dissertation is composed of five articles divided into two parts. In each article, I make use of international large-scale assessments and present comparative analyses. The first part of the dissertation explores the concept of social inequality, offering more detailed analysis of its measurement and the relationship between socioeconomic, immigration, and gender inequalities.

Article 1 explores the correlation between different measures of social inequality, based on the social category used to measure the inequality—in this case, achievement gaps between groups. Achievement gaps are compared between high and low SES groups, between native and immigrant students, and between boys and girls. The results indicate that there is no correlation between SES, immigrant, and gender achievement gaps; some countries have high inequalities under one social category but low inequality in other categories. This indicates that there is not an 'umbrella' of social inequality; social inequalities function independently of each other. While SES gaps are present in almost all education systems and almost all follow the same direction, gender gaps and immigration gaps are not present in every context and the gap follows different directions between countries.

Article 2 explores the different ways of measuring socioeconomic inequality; in particular, the correlation between SES measures used to estimate further SES inequality and other indicators of SES inequality. The results reflect a high correlation between most SES indicators, especially when aggregated at the country level, and a high correlation between SES inequalities based on different SES indicators. However, some differences remain, and some SES indicators are shown to work better in some countries than others. In addition, there is only a low correlation between the scores' dispersion and standardized SES gaps.

The second part of the dissertation explores the effects of between-school tracking on several measures of inequality, and on segregation.

Article 3 concerns the effects of between-school tracking on dispersion inequality, SES inequality, and inadequacy, using all available ILSAs data at the time the study was designed. The

findings lead to different conclusions depending on the measurement of inequality studied; between-school tracking had a strong positive effect on SES inequality and was very consistent between replications, it had a positive effect on dispersion inequality which was less consistent, and it had a significant but weak positive effect on inadequacy. The study also investigated the effect on mean performance, finding no effect. Overall, the results constitute very robust evidence concerning the effects on between-school tracking, while also highlighting the importance of the differences between concepts and measures of inequality. Each effect on each concept of inequality is theoretically justified under different terms, and each effect is evaluated under a different normative framework.

In Article 4, I estimate the effects of between-school tracking on socioeconomic segregation across schools. The findings indicate a strong effect of tracking increasing socioeconomic segregation across schools. This research highlights the importance of studying the effects of the school structure on the composition of schools, as achievement is not the only good obtained from education. The study is also the first to perform such analyses with international data.

Article 5 returns to the topic of between-school tracking effects, estimating its effects on gender segregation across schools and on the achievement gap. Tracking has a robust effect on gender segregation; tracking increases the disparity in the gender composition of schools. However, there is no effect from tracking on the achievement gap, except for a weak effect in mathematics.

The dissertation offers further insights regarding challenges and limitations, such as effect identification, doing low-N analyses, and working with international data. Overall, the dissertation illustrates how the measurement of inequality and related concepts affects results. Each concept of inequality and segregation requires different evaluation frameworks, with low correlation across conceptualizations and different results depending on the inequality measurement studied. Researchers should rationalize and explicitly identify the framework underlying their studies of inequality and segregation.

Table of Contents

Acknowledgments i
Summary iv
I. Introduction 1
Inequality or Inequalities? Different Concepts, Different Values 1
Additional Discussion on Social Inequality and Segregation 5
Empirical Consequences of Educational Inequalities Measures 6
About this Dissertation 9
References 12
II. Contributions 15
Article 1. Mind The Gap… But Which Gap? The Distinctions Between Social Inequalities in Student Achievement 15
1. Introduction 15
2. Social Inequalities in Achievement on International Analyses 16
3. The present study 20
4. Methods 21
5. Results 26
6. Discussion 32
7. References 35
6 Appendix 39
Article 2. Socioeconomic Inequality in Achievement: Conceptual Foundations and Empirical Measurement 43
Conceptual foundations 43
Measurement Issues 47
Empirical Analyses 51
Concluding Remarks 56
References 59
Article 3. Early Tracking and Different Types of Inequalities in Achievement: Difference-in-Differences Evidence from 20 Years of Large-scale Assessments 61
Literature Review: How does Tracking affect Educational Inequalities? 63
Research Questions 68
Methodology 68
Results 75
Discussion and Conclusion 81
References 85
Annex 89

Article 4. Does Tracking Increase Segregation? International Evidence on the Effects of Between-school Tracking on Social Segregation Across Schools 91

1. Introduction 91
2. Literature Review 93
3. Research Question and Hypothesis 95
4. Methodology 96
5. Results 102
6. Discussion and Conclusion 108
7. References 111
8. Appendices 114

Article 5. The Effects of Early Between-School Tracking on Gender Segregation and Gender Gaps in Achievement: A Differences-in-Differences Study 119

Literature Review 120
Materials and Methods 123
Results 129
Discussion 134
References 137
III. Overall Discussion 141
Summary of Findings 141
Outreach and Limitations 142
Final Remarks 144
References 146
IV. Appendix 147

Publication Status of the Individual Contributions 147
Author Contributions to Articles 148
Eidesstattliche Erklärung 151

I. Introduction

Educational inequality is one of the most prominent topics in educational and sociological research. An important part of educational research focuses on identifying inequalities within education systems and how educational practices and policies may configure those inequalities. Nevertheless, the sole notion of *inequality* hides different conceptualizations of *what* inequality is, with different normative assumptions of justice and distributive values. The metrics on which educational inequalities are measured are determined by these differences in implicit values. The main thesis of this dissertation is that there is not one inequality but several *inequalities*; the conceptualization of inequality and—by consequence—its operationalization, affects the empirical results of educational research.

To further explore the idea of inequalities, this dissertation includes a study of the effects on inequality of education systems' institutional features; specifically, the effect of between-school tracking on different measures of inequality. Between-school tracking refers to the streaming of students into different tracks (different types of schools with different curriculums and further educational paths). I also analyze effects on school segregation, understanding the composition of the school as an alternative dimension to inequality. This dissertation also includes descriptive analyses on measures of inequalities with international data.

In this chapter, I introduce the main concepts studied in the dissertation. In the first section, I discuss the concept of educational *inequalities* and outline the theoretical framework of the thesis. I mention some of the empirical issues associated with the measurement of inequality, followed by a brief discussion on the role of school differentiation in inequality. Finally, I introduce the research questions and describe the articles that comprise this dissertation.

Inequality or Inequalities? Different Concepts, Different Values

Brighouse et al. (2018) assert that educational decisions revolve around three aspects: the educational goods that education provides, distributional values distributing those goods, and independent values which are external to education but interact with it. Educational goods are aspects which the education system aims to deliver to children, such as knowledge, skills, dispositions, and attitudes. These goods are distributed among the student population, and the principles governing this distribution are distributive values. Brighouse et al. (2015, 2018) maintain that distributive values have two components: a distributive rule, and an object of distribution to which that rule applies. In this thesis, the object of distribution is the academic achievement of the students (except in Article 2); this is an important predictor of, for example, labor market returns, wellbeing, political engagement, integration, and countries' economic growth (Brighouse et al., 2018; Hanushek, 2013; Hanushek et al., 2015). The distributive rules and how they affect empirical research are the core topic of this dissertation.

The distributive rules behind each metric of inequality follow different principles. Research on between-school tracking is a good illustration. Van de Werfhorst and Mijs (2010) reviewed the

evidence of the effects of school system stratification on student achievement, categorizing the results into two inequality outcomes: inequality as dispersion and inequality of opportunities. The former is univariate and focuses on the distribution of achievement, while the latter is bivariate and observes the association of achievement with social background characteristics (in this case, socioeconomic status [SES] and immigration). Van de Werfhorst and Mijs recognized the differences between both outcomes and distinguished them as different results in their review; an education system can have a high dispersion of scores without it implying high differences between low and high SES groups, and vice versa. However, not all authors have made this distinction between dispersion and inequality of opportunities. Jakubowski (2010) and Waldinger (2005) replicated the results of Hanushek and Woessmann (2006) and reached different conclusions. Neither Jakubowski nor Waldinger noted that they were observing totally different measurements of inequality. While Hanushek and Woessmnan studied the effects on inequality as dispersion, Jakubowski and Waldinger examined the association of achievement with socioeconomic status, i.e., inequality of opportunities. Besides some methodological differences, the dissimilar conclusions may have resulted from the different inequality measurement used. This difference is important, both empirically and theoretically.

Therefore, in this thesis I consider *concepts* of educational inequality. Although many researchers ignore this, educational inequality should not be addressed as a single unitary concept. Behind each distributive value there are also different visions of justice and fairness. I outline two distinctions: first, within egalitarian approaches to inequalities, the distinction between equality and equity; second, the distinction between egalitarian approaches and adequacy.

Equality vs. Equity

Within egalitarian approaches, it is possible to distinguish between *equality* and *equity* (UNESCO, 2018). The term *equality* refers to "the state of being equal in terms of quantity, rank, status, value, or degree" (Jacob & Holsinger, 2009, p. 4). The term *equity*, in contrast, refers to "the social justice ramifications of education in relation to the fairness, justness and impartiality of its distribution at all levels or educational sub-sectors" (Jacob & Holsinger, 2009, p. 4). Equality is a broad term that aims to reduce the differences between children, while equity concerns how just those differences are. A common way of assessing this fairness is the concept of *inequality of opportunities,* referring to the idea that "everyone should have the same opportunity to thrive, regardless of variations in the circumstances into which they are born" (UNESCO, 2018, p. 17). The distinction between both is not a short discussion. When we discuss inequity, as with other similar terms such as *social inequality* or the aforementioned *inequality of opportunities*, we are making a normative statement: some differences are fair, and some are not.

This dissertation does not go into detail on the idea of fairness or justice, but it is important to briefly mention the discussion regarding *meritocracy*. Most people agree that differentials due to the characteristics of children that are beyond their control are unfair; these include the family into

which they were born and all that is associated with it, such as access to resources, migration background, and religion, or bio-cultural aspects, such as gender or ethnicity. There is not complete agreement on whether differences due to students' skill, ability, or talent is fair; i.e., whether we subscribe to meritocracy. Under the idea of meritocracy, students have some innate talent, intelligence, or ability that justifies the differences between them. A similar alternative is to consider students' effort instead of trying to assess their real ability, though the logic is the same. This concept of meritocracy is disputed for several reasons: i) the existence of innate ability is a strong assumption; ii) if it is unfair for factors beyond the control of children to affect their education, then the idea of an innate ability that they did not decide or control (since they are born with it) affecting their education should also be considered unfair; iii) considering the high correlations between children's backgrounds and their performance and attitudes, it is difficult to disentangle real innate ability from the advantaged or disadvantaged background—this argument also applies when rejecting the idea of meritocracy based on effort (Harel Ben-Shahar, 2016).

In practice, the discourse on the inequality of education is oriented towards an *equity* approach, focusing on the fairness of differences. Every education system is stratified and actively seeks differentiation between students; education systems differ in how the access and distribution of educational goods is configured. Regarding school tracking, all countries have some sort of streaming system, but vary on the number of tracks, how tracking is done, and at which age[1]. Considering this differentiation at the system level, *inequality* is not only expected, but even desirable (Montt, 2011). However, there is a discussion about the *social* inequalities widened by tracking: in the USA, how black and Latin American students are overrepresented in lower tracks within their schools (Hallinan, 1994); in Germany, how Turkish and Arab families' descendants are overrepresented in vocational tracks (Hillmert & Jacob, 2010).

In contrast, a less discussed topic is the collision of both values with independent values external to education (Brighouse et al., 2018). Every decision comes with a compromise; a policy that aims to equalize the field between advantaged and disadvantaged children will probably come at the expense of advantaged children, raising opposition. Such policies include redistribution of resources, or limitations to parental choice. These tensions are even higher when taking an egalitarian perspective that rejects meritocracy (Harel Ben-Shahar, 2015). Within this discussion is where the concept of adequacy acquires relevance.

Adequacy vs. Egalitarian Views

The concept of *adequacy* is raised as an alternative approach to egalitarian perspectives (either *equity* or *equality*). While it does not have a straightforward definition, adequacy generally relates to the idea that education should aim to bring a minimum set of skills and knowledge to the whole population; differences above this threshold are not considered problematic (Brighouse &

[1] For example, in Germany, students are streamed into different schools with different curriculums after grade 4; in the USA, students are not physically separated into different schools until after secondary school, but within schools there are different tracks.

Swift, 2008; Harel Ben-Shahar, 2015; Satz, 2007; Wise, 1983). Alternatively, authors such as Solga (2014) refer to *educational deprivation*, defined as the level of education that is insufficient for post-education labor markets and social life.

Adequacy is defended as a pragmatic approach, less conflictive with other independent values, and a better solution for educational policy discussion (Anderson, 2007; Satz, 2007). Satz (2007) defends adequacy, arguing that not all inequalities have the same importance. The difference in resources between an upper middle-class school and an upper-class school is not as relevant as the lack of basic skills in a group of the population; a focus on the lower part of the distribution of students would make a bigger contribution to their inclusion in society than focusing on the gap between low and high performers. Anderson (2007) argues that the equality perspective only considers education as an individual good, limiting the development of more advantaged groups. Defendants of the equality perspective admit that adequacy is more compatible with external values (Brighouse & Swift, 2009; Harel Ben-Shahar, 2015). However, both Brighouse and Swift (2009) and Harel Ben-Shahar (2015) argue that equity and equality perspectives are compatible with these other external values as long as there is a balance.

The evaluation of school tracking is altered by an adequacy approach. From an equality approach, tracking is problematic due to the conscious distancing between children. From an equity approach, tracking is problematic as long as the differentials between tracks are associated with the background of the student. In contrast, from an adequacy approach, tracking may even be desirable, as long as it is not detrimental for disadvantaged children to reach minimum levels of skills.

Concepts Used in This Dissertation

In summary, we discuss *concepts of inequalities* and not just one single inequality. Each corresponds to different values and aims for education. When evaluating educational systems (e.g., policies such as between-school tracking), the equality framework used in the evaluation will change depending on the approach taken. Another issue is the inconsistent naming of these concepts between authors; for instance, Brighouse tends to define *equality* as other authors use *equity*, and others *equality of opportunities*. For clarity, I use the following nomenclatures in the rest of the discussion and in the dissertation contributions:

i) Dispersion inequality corresponds to the concept of broad equality discussed above; focusing on the full distribution of students, i.e., problematizing the differentials between low and higher performers.

ii) Social inequality corresponds to the concept of equity and equality of opportunities discussed above; the association between students' backgrounds (e.g., socioeconomic status) and their performance.

iii) Inadequacy corresponds to the adequacy or deprivation discussed above; focusing on the lower part of the distribution, what percentage of students reach a minimum threshold of skills.

Additional Discussion on Social Inequality and Segregation

Social Categories

In this dissertation, besides the discussion on dispersion inequality vs. social inequality vs. inadequacy, I also explore some issues regarding the measurement of social inequality. Social inequality of achievement is defined as the degree to which the performance of students is associated with their origin, independent of their skills and effort. Social inequality is unusual in that it is bidimensional, as is the intersection between children's origins and their performance; meanwhile, inequality and inadequacy are focused on the distribution of performance (or another outcome). In articles 3 to 5, I explore further the social categories on which performance differentials are compared. *Social inequality* makes explicit how social categories affect the educational trajectories and outcomes of children. I identify two categories of social categories: family background, and biocultural characteristics.

Family background refers to characteristics transmitted historically and inter-generationally. Some examples of this correspond to characteristics such as: i) socioeconomic status (SES), making reference to a broad umbrella of categories regarding families' access to social and cultural resources due to their economic conditions[2]; ii) immigration status, as students from immigrant backgrounds are often disadvantaged compared to students with no immigrant background; iii) ethnic group and religion, recognizing that these are related to immigration, and it can be difficult to distinguish between these three aspects; historically, students from different ethnic groups and religions also have marked differentials in access and outcomes in education; iv) urbanicity; the access, valorization, and expectations of children in education is also affected by the area where they live (level of urbanization, size of the city, etc.). In the case of SES inequality, it is also possible to theorize further on which is the better indicator of socioeconomic differences in an educational context.

Biocultural characteristics refer to the features of students that have a biological dimension, though their relationship with educational outcomes is due to the cultural attitudes and expectations related to these biological features. With this I mainly refer to: i) gender; although biological sex differences do not have any important effect on children's educational performance (see review by Rosén et al., 2021), gender has been historically associated with great differentials in access, educational paths, and outcomes. The word *gender* itself does not correspond to a biological distinction but a cultural one, especially because identifying as a *boy* or *girl* comes with a different set of expectations and attitudes; ii) ethnicity; although there is a discussion in both social and natural sciences of the degree to which ethnicity is a genetic aspect vs. just a social construct (Frank, 2015), it is clear that being a person of color in the USA does not have the same consequences and difficulties as in South America, implying that it has a very strong social factor. It is difficult to draw

[2] For this dissertation, I will avoid discussions regarding stratification based in sociology, such as occupational status vs. social class

a distinction between ethnicity and ethnic inequalities—as is the case for the African-American population in the USA; for this reason, ethnoracial may be a more preferable term.

As I explain in further detail in Article 3, the origins of the inequalities based on these social groups differ between social categories. This brings an added complexity to the study of social inequalities: should these be studied separately? Or is there an umbrella feature of *social inequality*? This even excludes some discussions, such as the intersection between different social characteristics.

School social segregation

Social segregation is not strictly part of the discussion on *inequality*, as theoretically schools could be segregated but still have equal results. However, in this dissertation I also study the topic of school social segregation. School social segregation refers to how children with different social characteristics (e.g., from rich and poor backgrounds, or boys and girls) attend different schools. This dissertation includes a discussion on segregation due to the importance of the social context and peers for students' cognitive learning and socialization. The evidence indicates that a higher proportion of students from advantaged backgrounds in the same school has positive effects on learning for less advantaged students (Sacerdote, 2011). Comparative studies have also revealed a correlation between the degree of social segregation and the social achievement gap (Burger, 2019; Hindriks et al., 2010).

Moreover, from a democratic perspective it is desirable that children have contact with peers from different backgrounds. While academic outcomes are important in evaluating educational systems, they are not the sole objective of education. Following the framework of Brighouse et al. (2018), educational goods include capacities for personal autonomy, democratic competence, healthy personal relationships, and treating others as equal; it is difficult for these values and skills to be easily achievable if children learn in a segregated system. Social segregation across schools can be considered as an important dimension to evaluate the fairness of educational systems, parallel to social inequality.

Empirical Consequences of Educational Inequalities Measures

The choice of concept of educational inequality has both theoretical and empirical implications, as different metrics of inequality correspond to different conceptualizations of inequality. UNESCO (2018) listed some of the most common indicators of educational inequalities, including the range between the highest and lowest parts of the distribution, variance or standard deviation, coefficient of variation, and concentration measures such as the Gini coefficient. They list social inequality indicators, such as achievement gaps, ratios between groups, correlation coefficients, effect slops, and the group Gini coefficient. There has been less methodological development in the measurement of educational inadequacy, though international large-scale assessments such as PISA, TIMSS, or PIRLS have developed thresholds of performance; PISA offers six levels of proficiency levels, while TIMSS and PIRLS offer four international benchmarks for comparison. Which threshold is considered as 'adequate' is open to discussion.

Strietholt and Borgna (2018) explored how correlated the different concepts related to equality are, following approximately the same concepts listed in the previous discussion: dispersion inequality, social inequality (based on socioeconomic status, measured via parental education), and inadequacy. They also included three different operationalizations within each concept. Their results (see Figure 1) indicated that the three indicators of each concept were not correlated with the other concepts' indicators; within each concept, the three measures chosen were highly correlated. This provides evidence for the empirical independence of the three concepts of inequality and validates the measurement of each concept, as within them there is a consistence within the available indicators. Similar analyses have not been replicated between the different social inequalities, i.e., across the different social categories.

The measurement of educational inequalities may have direct consequences for empirical research on education. If the different indicators mean different things and are not correlated, then it can be expected that the results may change. This is in addition to more substantive issues, as the theory supporting a hypothesis (i.e., variable X increases Y inequality) also changes between the different concepts, along with the framework of evaluation (why use this measurement and not another?). The consequences of the measurement of educational inequality are studied in the context of school differentiation.

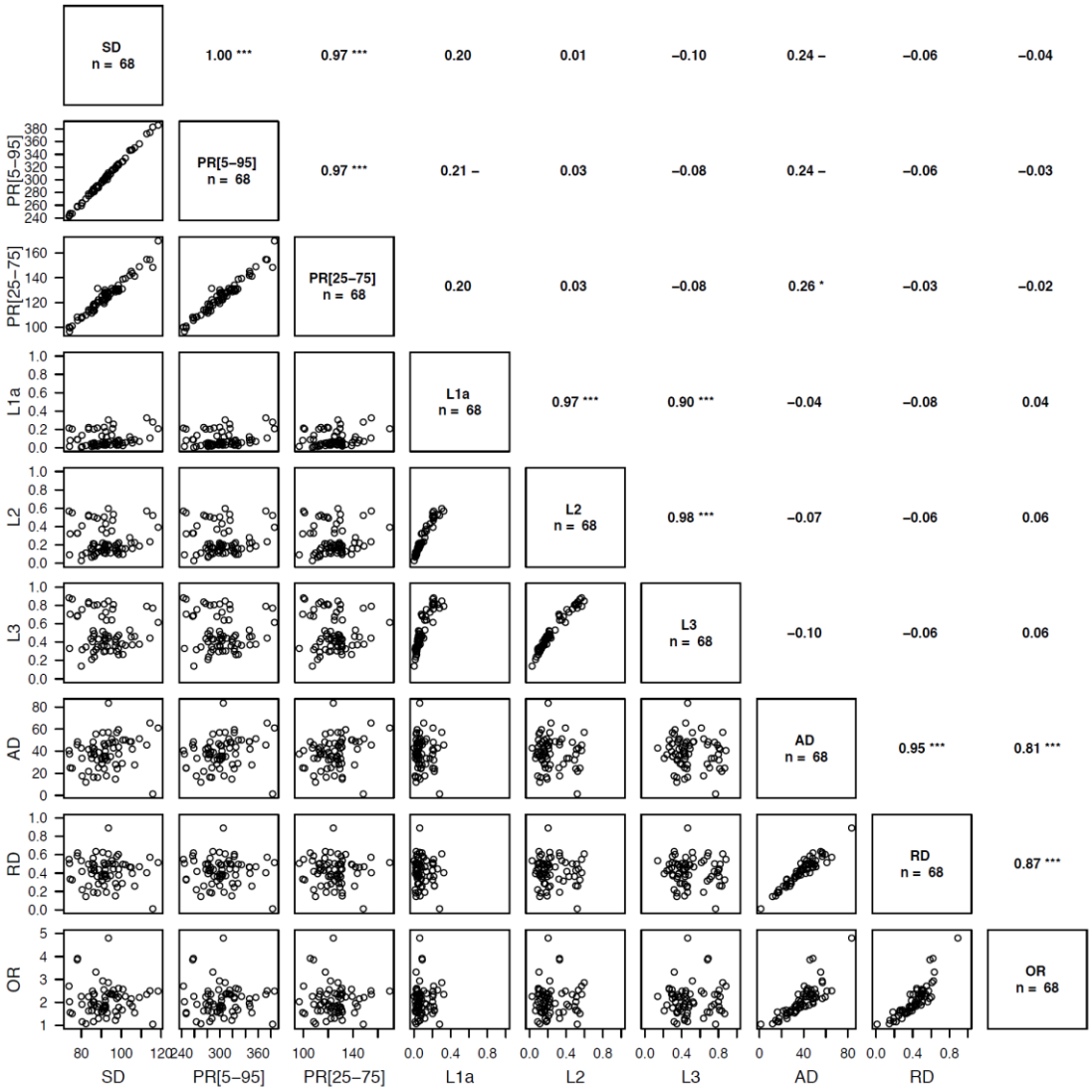


Figure 1. Correlations between inequality measures in reading in PISA 2012. Source: Strietholt and Borgna (2018). *Note.* Pearson's correlations, significance levels: *** p < .001; ** p < .01: * p < .05: - p < .10 (two-tailed). Measures of inequality: scores SD, range percentile 5 – 95, range percentile 25-75; measures of inadequacy: percentage not reaching proficiency level 1a, level 2, and level 3; measures of social inequality: absolute achievement gap, relative achievement gap, odds ratio of reaching proficiency level 2.

School Differentiation, Tracking, and Inequalities

In the context of inequalities, a common topic to study is the effect of the education system's features on outcomes (e.g., achievement or attainment). One of the most common dimensions studied is the level of school differentiation, or how stratified educational systems are. All educational systems provide differentiated ways of schooling, but vary in how this differentiation occurs. Variation in differentiation can be categorized within two dimensions: externalization and formalization (Skopek et al., 2019). The former corresponds to the differentiation between schools vs. differentiation within schools, while the latter corresponds to how formally or informally schools are differentiated. In this dissertation, we specifically study the effects of between-school tracking, an external and formal type of school differentiation (Chmielewski, 2014; Dollmann, 2019; Skopek

et al., 2019). Between-school tracking is discussed in terms of a tradeoff between efficiency and equity (Hanushek & Wößmann, 2006), as more homogeneous schools and classrooms supposedly enable more efficient teaching; resource allocation and teacher methods can be more easily adapted, enabling an unequal allocation of resources leading to different educational paths.

Previous research on the effects of inequality on between-school tracking has found differing results between inequality outcomes: for dispersion inequality, previous evidence showed mixed results; for social inequality, the effect of tracking tends to be positive (i.e., it increases inequality); for inadequacy, there is some evidence indicating a detriment to the lower part of the distribution of scores (van de Werfhorst & Mijs, 2010). Although there has been no systematic study comparing the effect of tracking on different inequality measures (besides the aforementioned study by van de Werfhorst & Mijs), previous studies have shown inconsistent results between measures. This is expected, as the hypotheses are sustained using different arguments, as explained in the previous section: tracking is expected to increase dispersion inequality, as it is an explicit differentiation of curriculums and expected educational outcomes; regarding social inequality, if the transition between primary and secondary school is not unbiased (as is often the case), then it is expected that tracking will increase the distances between social groups; regarding inadequacy, tracking may stigmatize lower tracks, hindering the development of children in the lower part of the distribution (see literature review in Article 1).

About this Dissertation

Research Question

This dissertation aims to explore the relevance of the measurement of inequality in education in theoretical and methodological terms, and the consequences for research of these decisions; specifically, how research on the effects of school differentiation is affected by the measurement of inequality. This dissertation addresses the following research questions:

I. How correlated are different measures of educational inequality?

The aim is to explore whether there is one umbrella conception of inequality in which the various conceptualizations of inequality correspond to roughly the same idea, or rather the opposite; to what degree the different conceptions of inequality are totally separated across them. It is also possible to ask the same question of social inequality: do all social inequalities correspond to the same phenomenon, or are they parallel?

II. How do the conclusions of school differentiation research change depending on the measure of educational inequality used?

The aim is to explore how the measurement of inequality also influences the interpretations, and on what levels—with the specific example of the effects of school differentiation (between-

school tracking). The different metrics have different normative assumptions behind them, but do the conclusions made change between different metrics?

Methodology

Data Sources

All contributions to this dissertation make use of publicly available international large-scale assessment (ILSAs) datasets. Starting in the mid-20th century with the first studies comparing achievement levels in mathematics in different countries, there are currently three major ILSAs ongoing: the OECD's Programme for International Student Assessment (PISA), the International Association for the Evaluation of Educational Achievement's (IEA) Trends in International Mathematics and Science Study (TIMSS), and Progress in International Reading Literacy Study (PIRLS). The PISA has been undertaken every 3 years since 2000, measuring the reading, mathematics, and science proficiency of 15-year old students; TIMSS has been undertaken every 4 years since 1995, measuring mathematics and science knowledge in students of mainly grade 4 and grade 8; PIRLS has occurred every 5 years since 2001, measuring the reading proficiency of grade 4 children. The use of these three studies enables access to information from more than 75 education systems from every continent.

In this dissertation I treat educational inequalities as a system-level measure. Therefore, international data must be used, as there is a need for variation on the main interest variable. Current ILSAs present a very heterogeneous sample of countries, especially in the case of PISA, enabling the comparison of different contexts and observation of variability of both inequalities and system-level features.

Analysis Strategy

This dissertation uses a mix of descriptive analyses and causal effects identification strategies. The descriptive analyses correspond to system-level correlations and are used to answer research question 1 (on the correlations between the different inequality measures). To answer research question 2 (how the inequality measurement affects the conclusions), I estimate difference-in-differences methods that exploit the variability within ILSAs and the availability of system-level information at both primary and secondary school.

Contributions

This dissertation includes five contributions. These can be divided into two parts. The first part concerns the correlation between variations of educational social inequalities.

Part One

Article 1. In the first article, *Mind The Gap… But Which Gap? The Distinctions Between Social Inequalities in Student Achievement*, we explore the correlation between socioeconomic inequality, immigration inequality, and gender inequality. The article explores the degree to which it is possible to talk of a broad umbrella of *social inequality,* or rather three independent concepts.

Article 2. In the second article, *Socioeconomic Inequality in Achievement: Conceptual Foundations and Empirical Measurement*, we explore the different indicators of socioeconomic inequality and how to measure SES itself. The article compares measures of inequality using different SES indicators and illustrates some important issues regarding research on socioeconomic inequality.

Part Two

The second part consists of research on school differentiation, assessing the effect of between-school tracking into different types of educational inequality and school segregation.

Article 3. In the third article, *Early Tracking and Different Types of Inequalities in Achievement: Difference-in-Differences Evidence from 20 Years of Large-scale Assessments*, we assess the effect of between-school tracking on dispersion inequality, social inequality, and inadequacy. No previous study has systematically compared the effects of tracking under different inequality measurements. In addition, the article makes further methodological contributions, mainly on the importance of replicating results, especially in the case of small N studies such as system-level analyses.

Article 4. The fourth article, *Does Tracking Increase Segregation? International Evidence on the Effects of Between-school Tracking on Social Segregation Across Schools*, follows a similar approach to Article 3 and observes the effects of tracking on socioeconomic segregation across schools. The article is the first to quantify the effect of tracking on segregation with an international perspective, reflecting how little attention has been given to such an important issue for democratic societies and children's development.

Article 5. In the fifth article, *The Effects of Early Between-School Tracking on Gender Segregation and Gender Gaps in Achievement: A Differences-in-Differences Study*, we study the effect of tracking on both gender school segregation and the gender achievement gap. The chapter incorporates some advances from Article 3 and Article 4 into the discussion on gender inequality in education. In this article, we show how the discussion and results on gender segregation and inequality differ from the discussion on socioeconomic inequalities, reflecting how both are important to fully understand the effects of policies on inequalities.

References

Anderson, E. (2007). Fair opportunity in education: A democratic equality perspective. *Ethics*, *117*(4), 595–622. https://doi.org/10.1086/518806

Brighouse, H., Ladd, H. F., Loeb, S., & Swift, A. (2015). Educational goods and values: A framework for decision makers. *Theory and Research in Education*, *14*(1), 3–25. https://doi.org/10.1177/1477878515620887

Brighouse, H., Ladd, H. F., Loeb, S., & Swift, A. (2018). *Educational Goods: Values, Evidence, and Decision-Making*. The University of Chicago Press.

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

Burger, K. (2019). The socio-spatial dimension of educational inequality: A comparative European analysis. *Studies in Educational Evaluation*, *62*(May), 171–186. https://doi.org/10.1016/j.stueduc.2019.03.009

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

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

Frank, R. (2015). Back to the Future? The Emergence of a Geneticized Conceptualization of Race in Sociology. *Annals of the American Academy of Political and Social Science*, *661*(1), 51–64. https://doi.org/10.1177/0002716215590775

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., Schwerdt, G., Wiederhold, S., & Woessmann, 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

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

Harel Ben-Shahar, T. (2015). Distributive Justice in Education and Conflicting Interests: Not (Remotely) as Bad as you Think. *Journal of Philosophy of Education*, *49*(4), 491–509. https://doi.org/10.1111/1467-9752.12122

Harel Ben-Shahar, T. (2016). Equality in Education – Why We Must Go All the Way. *Ethical Theory and Moral Practice*, *19*(1), 83–100. https://doi.org/10.1007/s10677-015-9587-3

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

Hindriks, J., Verschelde, M., Rayp, G., & Schoors, K. (2010). School tracking, social segregation and educational opportunity: evidence from Belgium. In *CORE Discussion Paper 2010/81*. https://uclouvain.be/fr/node/26415#Alfresco

Jacob, W. J., & Holsinger, D. B. (2009). Inequality in Education: A Critical Analysis. In *Inequality in Education* (pp. 1–33). Springer Netherlands. https://doi.org/10.1007/978-90-481-2652-1_1

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). Springer. https://doi.org/10.1007/978-90-481-3993-4_3

Montt, G. (2011). Cross-national differences in educational achievement inequality. *Sociology of Education*, *84*(1), 49–68. https://doi.org/10.1177/0038040710392717

Rosén, M., Steinmann, I., & Wernersson, I. (2021). *Gender differences in achievement*.

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

Satz, D. (2007). Equality, adequacy, and education for citizenship. *Ethics*, *117*(4), 623–648. https://doi.org/10.1086/518805

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). https://doi.org/10.4337/9781788110426.00022

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., & Borgna, C. (2018). *Inequality in Educational Achievement. Different Measures, Different Conclusions* [Unpublished manuscript].

UNESCO. (2018). *Handbook on Measuring Equity in Education*. https://doi.org/10.1016/S0733-8619(03)00096-3

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

Waldinger, F. (2005). *Does tracking affect the importance of family background on students' test scores?* https://www.fabianwaldinger.com/research

Wise, A. E. (1983). Educational Adequacy: A Concept in Search of Meaning. *Journal of Education Finance*, *8*(3), 300–315. http://www.jstor.org/stable/40703367

II. Contributions

Article 1. Mind The Gap… But Which Gap? The Distinctions Between Social Inequalities in Student Achievement

Andrés Strello, Rolf Strietholt, Isa Steinmann

The version of the manuscript printed below is the preprint of the article accepted for publication in *Social Indicators Research* (in print).

Abstract

International large-scale assessments have revealed social inequalities in achievement in almost all countries, reporting achievement gaps between socioeconomic status (SES) groups, by immigration background and by gender. However, there has been little research on whether individual countries show smaller or larger gaps across all three different social categories, or whether the gaps corresponding to these categories are independent of each other. This article explores the degree to which social inequality can be understood as one umbrella concept, or whether different categories of social inequality are substantially different concepts. Using the OECD's Programme for International Student Assessment (PISA) 2018 results in Mathematics in 76 countries, the study observes the correlation between the three achievement gaps across countries, and compares how each achievement gap is associated with some typical country-level covariates. Several results are highlighted. First, the size and direction of the immigration and gender gaps vary across countries; most countries present achievement gaps in favor of boys and native students, but this direction is reversed in several countries. Second, there is hardly any correlation between the three achievement gaps. One education system may be egalitarian in one category, but profoundly unequal in another. Third, this lack of correlation is also related to how we study these inequalities, as the results show that each achievement gap is associated with a different set of institutional features. To properly assess how unequal or egalitarian education systems are, researchers and interested parties need to consider and address different indicators of social inequality.

1. Introduction

International large-scale assessments (ILSAs) such as PISA or TIMSS reveal considerable variation in both the mean performance levels and the extent of social inequality that exists within participating countries. Regarding social inequality, the most commonly studied social categories are socioeconomic status (SES), immigration status, and student gender (e.g., Andon et al., 2014; Rosén et al., 2022; Jerrim et al., 2019). The answers to our research questions could have multiple

implications for educational monitoring, as well as for research on social inequality in student achievement. If achievement gaps between different social categories are highly correlated, then examining them separately adds little value for educational monitoring, and their reporting should be reframed. In this scenario, it also seems plausible that research findings on the institutional determinants of social inequality would be consistent across different social categories. But if different social gaps are largely uncorrelated, there is a need for a differentiated perspective in education policy and research.

This study empirically examines whether there is a single broad social inequality, or whether there is a need to distinguish between different forms of social inequalities corresponding to the categories of SES, immigration background, and gender. Are there countries which systematically show social inequality in student performance across different categories, or are countries characterized by a higher degree of inequality in one social category and lower in another? To address this question, we first use data from an international large-scale assessment to compute social inequalities in SES, immigration status, and gender. We use this data to review the variability in social inequality for each social category across countries, before evaluating the correlation between the three measures at country-level. To further validate these analyses, we conduct a comparative study and investigate the association between various institutional features and the three forms of social inequalities. Specifically, are institutional features of school systems consistently associated with all forms of social inequality or only to specific ones? This study is explorative and aims to contribute to the discussion on how researchers can evaluate education systems.

2. Social Inequalities in Achievement on International Analyses

2.1 The concept of social inequality

Inequality in education can be conceptualized using different terms, with different normative ideas about injustice and the evaluation of education systems (Strietholt, 2014; Brighouse & Swift, 2008). The concept of social inequality—similar to 'inequity' and 'inequality of opportunities' (c.f. UNESCO, 2018)—problematizes achievement differences that originate from the social origin of the student, rather than from effort or ability. In educational research, the most commonly used social categories are SES, immigration background, and gender. While these are the three categories studied in this article, we acknowledge that there are other characteristics determined by social origin that are also related to inequalities within education, such as religion, sexual orientation, ethnicity, and place of residence (e.g., urbanicity).

Understanding the categories of social inequality presents a dilemma. Different categories of social inequalities have some common aspects. First, since there are great differentials in the outcomes and trajectories of students, we can expect that there are groups that are able to take more advantage than others in a systematic way. This is especially true in contexts of high general inequality, i.e., high dispersion in outcomes, where differentials between students are bigger and there is more variance that can be unevenly distributed. Second, we selected these three categories

of social inequality because they are present in most education systems in the world and have been an ongoing topic in educational research for decades. The global relevance of these three categories enable us to hypothesize that there is one broad 'umbrella' social inequality, in which these social distinctions (SES, immigration, gender) are associated with the distribution of outcomes simultaneously, i.e., highly associated between the three of them. However, different categories of social inequality within an education system emerge for different reasons. The association between each social category with performance outcomes could run on parallel paths, implying null correlations between them.

Researchers have been able to measure the different categories of social inequality in achievement on an international perspective since the mid-20th century. Currently, the three largest ongoing international large-scale assessments measuring achievement in school students are the OECD´s Program for International Student Assessment (PISA), the International Association for the Evaluation of Educational Achievement's (IEA) Trends in International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRLS). Below, we present a short review of the current international evidence and the theories explaining each category of social inequality. We will explore prominent theories regarding the emergence of achievement gaps related to SES, gender, and immigration. Our aim is to demonstrate that the underlying mechanisms behind these gaps are fundamentally distinct from one another. Subsequently, we will examine previous research that investigates the correlation between institutional features of educational systems and the three distinct types of inequality. In addition, it is worth noting that previous research in this area has been somewhat fragmented. There has been a lack of systematic evaluations where the same data were utilized to study the relationships between institutional features and the various forms of inequality. Moreover, the few studies that have attempted this approach have yielded inconsistent findings.

2.2 SES Inequality in Achievement

The association between family SES and student achievement can be explained by the tendency of children from families with a lower socioeconomic background to receive fewer resources for their education. This difference in resources accumulates along the children's developmental trajectory and generates disparities in achievement between children from different families. This is further exacerbated by the inheritability of resources between generations that increases the resource gap between families. According to Bourdieu's theory, these resources are manifested first as economic resources (e.g., families with higher incomes can send their students to private schools or afford private tutoring) and later manifest in cultural and social capital (Bourdieu, 1986; Broer et al., 2019; Coleman, 1988, 1990).

The association between a student's SES background and performance has been a common finding across studies, cycles, and subjects, though with differences between countries in the association's magnitude (Hopfenbeck et al., 2018). PISA 2018 (OECD, 2019a) shows a positive

association between SES and achievement in all countries and in all three subjects (reading, mathematics, and science), with SES explaining between 2% and 24% of the variance in performance, depending on the country and subject. TIMSS (Mullis et al., 2020) and PIRLS (Mullis et al., 2017) presented similar patterns in their latest editions in 2019 and 2016, respectively. While these are recent results, SES inequality in performance is not new and has even increased in some countries (Broer et al., 2019; Chmielewski, 2019).

2.3 Immigration Inequality in Achievement

The association between immigration background and student achievement can be explained by two groups of mechanisms: structural and cultural (Nauck, 2019). The structural mechanism is the inherent disadvantage experienced by immigrant groups due to their economic reality. Families with immigrant backgrounds show lower academic performance or take different educational choices due to their poorer access to resources (both economic and social). Cultural explanations ultimately focus on why certain groups of immigrants or ethnic groups perform better than others; the disadvantage is explained in terms of different mindsets. Studies have shown that the gap between immigrant and native students is not only due to immigrant families' lower SES, but also due to speaking a different language at home, sociocultural factors, system-wide factors of the origin and destination countries (such as political stability, economic development, and religion), and the destination countries' policies (Buchmann & Parrado, 2006; Dronkers & Levels, 2007; Jackson, 2012; Levels et al., 2008; Schmid, 2001; Strand, 2011, 2014).

Most international research in English on the association between immigration and achievement has focused on Western European countries and the USA. In European countries, students who speak a different language at home perform worse in PISA, especially at reading (Lenkeit et al., 2017). Moreover, most studies using ILSAs data have each only investigated a limited set of countries, focusing on the differences between immigrant groups within a country, e.g., the disadvantage of Turkish communities within Germany (Söhn & Özcan, 2006), or how immigrants are disadvantaged in the USA depending on their origin country (Worrell, 2014). As most research is centered in European and North American contexts, some other contexts are excluded. In Qatar and United Arab Emirates, immigrant children perform better than natives, supposedly because these countries attract high-skilled immigrants and their education systems are tailored to this (Bouhlila, 2017). Overall, the achievement gap varies greatly across the assessed countries, contents, and cycles (Andon et al., 2014).

2.4 Gender Inequality in Achievement

There are different and longstanding theories on why gender gaps in student achievement tests occur, and they can be divided into two broad explanations: nature and nurture (see overviews by Halpern, 2012; Hyde, 2014). The nature category includes theories that assume innate, stable differences between boys and girls that affect learning processes. The comprehensive literature on cognitive gender gaps suggests, however, that boys and girls mostly score equally on cognitive ability

tests (cf. Gender Similarity Hypothesis; (Hyde, 2014; Zell et al., 2015). In contrast, the nurture category includes theories about environmental influences differing between boys and girls. Nurture-related theoretical perspectives all suggest that societal gender norms and existing gender differences in education transmit to students, perpetuating educational gender inequalities. For instance, stereotypical beliefs about science, technology, engineering, and mathematics (STEM) subjects being male domains and a higher representation of men in STEM majors at school and university level or in the STEM labor market can lead to girls underestimating their abilities in these subjects, potentially impacting their achievement (Eccles et al., 1990; Halpern, 2012; Neuville & Croizet, 2007).

International comparative studies document pronounced gender gap differences between countries and academic achievement domains. Girls outperform boys in reading in most countries at both the primary and secondary school level. Gender gaps in the participating countries range between non-existent reading gender gaps to large advantages for girls (Mullis et al., 2017; OECD, 2019a). Gender gaps are more varied in mathematics, with medium advantages for boys in some countries, some countries without gender gaps, and even some countries with medium advantages for girls (Mullis et al., 2020; OECD, 2019a). Interestingly, gender gaps in reading and mathematics appear to correlate; countries with pronounced reading advantages for girls also tend to show mathematics advantages for girls, and countries without reading advantages for girls tend to show mathematics advantages for boys (Guiso et al., 2008; Stoet & Geary, 2013). Furthermore, gender gaps in academic achievement appear to be quite stable over time (Rosén et al., 2022; Steinmann et al., 2023; Meinck & Brese, 2019).

2.5 Covariates of Social Inequalities

Within each category of social inequality, the associations between social origin with performance vary between countries. This suggests that institutional features of education systems generate variations in social inequality (Jerrim et al., 2019). We next review some studies that have identified institutional features related to social inequality in achievement. We explore whether previous studies suggest that institutional covariates are associated in the same way with different forms of social inequality.

2.5.1 Education-system Factors

One important feature of education systems is the level of differentiation, seen in policies such as between-school tracking, in which students are sorted into different types of schools. If transitions and school choice are affected by social characteristics, either by the achievement differential between social groups or by different decisions taken after considering children's skills, differentiation in the education system should lead to larger social achievement gaps. Previous international studies have found that educational differentiation (specifically between-school tracking) increases SES inequality in achievement (Strello et al., 2021; Lavrijsen & Nicaise, 2016; van de Werfhorst, 2018; van de Werfhorst & Mijs, 2010). There is less research on the effect of

tracking on immigration inequality in achievement and the findings are inconsistent; some studies suggest a positive effect while others do not (Bodovski & Munoz, 2020; Ruhose & Schwerdt, 2016; Teltemann & Schunck, 2016). Between-school tracking has mixed effects on gender inequality, with studies consistently showing that later tracking increases the gender gap in reading (in favor of girls), but heterogeneous results regarding the effect on mathematics and science (Bodovski & Munoz, 2020; Hermann & Kopasz, 2019; Scheeren & Bol, 2022). Similar results have been found in studies on general education-system differentiation indexes (Ayalon & Livneh, 2013; van Hek et al., 2019; van Langen et al., 2006).

2.5.2 External Factors

A common factor in comparative research is the level of economic development of a country or education system. In general, previous studies have found mixed evidence on its effect on social inequality. Measures such as GDP (gross domestic product) per capita are inconsistently associated with SES achievement inequality (Bodovski & Munoz, 2020; Chmielewski, 2019; Ferreira & Gignoux, 2014; Schütz et al., 2008). Previous studies have also found mixed results regarding the association between SES inequality and public expenditure on education, although the association seems more markedly negative when considering countries' development levels (Strietholt et al., 2019). Chmielewsky (2019) found that income inequality (measured as Gini) has a positive association with SES inequality in mid and low-income countries. Using TIMSS 2011 data, Bodovski and Munoz (2020) found an inverse association between GDP per capita and the immigrant achievement gap (in particular, richer countries have a lower gap between immigrants and native students), but found no association with the gender achievement gap.

Cultural features may also play a role in gender inequalities in achievement. In more gender-egalitarian countries, the relative performance of girls over boys is higher, especially in reading (see review of Rosén et al., 2022; González de San Román & de La Rica, 2016; Guiso et al., 2008; Marks, 2008; Reilly, 2012). Nosek et al. (2009) found that in societies with more marked stereotypes (e.g., regarding science as a male domain and liberal arts as a female domain), the gap in favor of boys is larger in mathematics.

3. The present study

Previous research on social inequality has identified a number of social categories related to student achievement, with the most prominent categories in international comparative research being SES, immigration, and gender. The theories explaining the emergence of each performance gap differ, and research on the three areas has developed relatively independently.

Only a few studies have explicitly compared the different gaps. Lenkeit et al. (2017) studied the relative importance of the three categories of achievement gaps, though only in four Western European countries (Germany, Sweden, France, and United Kingdom). They estimated multilevel models using data from PISA 2000 to 2012. The authors concluded that each category of social inequality is important for explaining the disparities between students, and that results have remained

stable in those four countries. Bodovski et al. (2020) studied how different country-level predictors may mitigate the three categories of social inequality. The authors used information from TIMSS 2011 with a sample of 45 countries. They found mixed effects between the different social inequality domains, showing that the role of school system features cannot be generalized over the different categories of social inequality. Whether large gaps in one social category are associated with large gaps in another social category has not been the subject of research to date.

In this study, we explore the relationship between the three categories of social inequality on achievement, and the degree to which they are correlated or uncorrelated. We investigate whether countries can be evaluated as more or less socially unequal based on one only category, or how important it is to evaluate the effects of certain policies on different categories of inequality. Specifically, we aim to answer the following research question:

RQ. *How correlated are the three categories of social inequality in achievement (socioeconomic status, immigration background, and gender)?*

A high correlation between the different types of social inequality would suggest that the differentiation between the three types of social inequality has no additional empirical value, whereas low correlations would underpin the importance of a differentiated view of social inequalities.

Furthermore, we proceed to examine the nomological linkages between the three distinct types of social inequality and external variables. This line of inquiry aligns with the principles of construct validity (Cronbach & Meehl, 1955). By assessing how different types of inequality correlate with relevant variables, we can gather evidence supporting their meaningful distinction. More specifically, our investigation focuses on the relationship between education system-level features and social inequality across the various categories of social inequality.

Different patterns in the regression parameters would provide evidence that the three types of inequality need to be differentiated when analyzing social inequality. If there are no differences in the regression estimates, however, differentiating between the types of social inequality would not provide additional empirical value.

4. Methods

4.1 Data sources

To study the correlation between the different categories of social inequalities, we use the OECD's Programme for International Student Assessment (PISA). This study measures 15-year-olds' proficiency in mathematics, reading, and science. Specifically, we use the dataset of PISA 2018 focusing on the mathematics assessment. We remove Korea and Vietnam from the sample as they sampled fewer than 20 students with immigrant backgrounds (as defined in the Variables chapter). The remaining sample of n=76 education systems[3] is heterogeneous and covers all parts of the world. Each country contains a sample between 3,296 to 35,493 students (mean: 7,791), with the number of

[3] Hereafter, we refer to the education systems sampled as "countries", even if some participants did not sample the full country, namely, China and Azerbaijan (Baku).

schools ranging between 44 and 1,089 schools (mean: 279). The total sample contains 592,145 students from 21,264 schools. Table 1 shows the total N of students and schools per country.

PISA draws a stratified two-stage sampling. The first stage samples schools within the country or education system, and the second stage samples 15-year-old students within those schools. The results are representative of the population at both the student-level and the school-level. However, PISA sample only students enrolled within schools, meaning that interpretations of these results must consider that some specific countries/regions have lower proportions of secondary-school attainment and therefore exclude early school leavers (Steinmann & Rutkowski, 2023).

4.2 Analysis

Do all measures of inequality show the same picture, or is it necessary to differentiate between multiple types of social inequality? Are social achievement gaps consistent, or are there countries in which certain social gaps are high and others low? To address these questions empirically, we examine whether different measures of social inequality in student achievement lead to the same or different rankings in international comparisons. All analyses are based on the three types of social inequality in student achievement available for the n=76 participants in PISA: SES, immigrant background, and gender.

Our analysis consists of three steps. First, we identify the three gaps per country (see Variables section below). Second, to answer RQ1, we examine the correlation of these different types of social inequality at the country level. Third, to answer RQ2, we attempt to validate the correlational analyses by regression analyses. We regress the three types of social inequalities on a set of institutional features and compare the regression parameters for the three outcomes. We use cross-sectional data, and the aim of the regression analyses is not to estimate causal effects or bring substantive conclusions, but rather to examine whether different social inequalities are associated differently with various system-level features.

4.3 Variables

4.3.1 Social Achievement Gap on Mathematics

The main variables of interest are measures of three categories of social inequalities in achievement. Achievement scores in PISA are calculated so that they had an international mean of 500 and an international standard deviation of 100 points in the first edition in 2000. The scores are designed to be comparable between countries. In our analysis, we focus on mathematics achievement.

The three achievement gaps were calculated for gender, SES, and immigration status, using the simple mean difference between the groups (described in the next section). We divided these gaps by the standard deviation of the mathematics scores observed in the respective country, to account for cross-country variation in the dispersion of the test scores. Therefore, all gaps are measured as Cohen's standardized effect sizes *d*. For example, a gender gap of 1 means that boys perform on average one standard deviation better than girls. Probability weights were used in the

estimation of the achievement gaps and standard error account for the sample design using replications weights. We followed the Balanced Repeated Replication (BRR) method, as indicated by PISA guidelines (OECD, 2019b). All ten plausible values available in the PISA public database were used on the analyses following Rubin's rules (Rubin, 1987). The three achievement gaps, by country, can be found in Appendix (Table 3).

SES Achievement Gap. We used parental education as a measure of socioeconomic status and compared students with parents with university education (ISCED 5A) against parents with less than university education. If the educational attainment of the parents differed, we used the highest educational attainment reached between both parents—i.e., one parent having university level education is enough to be considered in the highest category. We marked parental education as missing if there was no information about both parents. We opted for a single proxy of SES instead of a complex index, such as the ESCS reported in PISA, for the sake of simplicity and consistency with the other categories that also use a single indicator. Table 1 shows the proportion of parents with university education. There is a high heterogeneity between countries on socioeconomic levels. This proportion ranges from 7% in Vietnam to 73% in Denmark.

Immigration Achievement Gap. We operationalize the immigration background by comparing students whose parents were both born abroad with students with one or no parent born abroad. We categorize the first group as "immigrant" and the second as "native", aware that this is a simplified category of a more complex phenomenon. We marked this variable as missing if there was no information about both parents. Immigration background has a high heterogeneity between countries, ranging from slightly over 0% in several countries to as high as 63% in Macao (see Table 1). Six countries (China, Korea, Peru, Poland, Romania, and Vietnam) have fewer than 30 cases with immigrant backgrounds. The efficiency of the estimation of achievement gaps based on immigration is reduced. However, excluding these cases does not affect the results of this study (see Results below).

We calculate the "raw" association between immigration and achievement scores. An alternative would be to estimate the achievement gap, controlling first for student SES. However, we want to highlight how immigration has different connotations between countries, as shown in Figure 1. In addition, the association between immigration background and SES tends to be small (even non-significant) in several countries, and in different directions, as shown in Appendix (Table 3). While many countries (e.g., Western European countries) show a positive association between being a native student and having parents with university education, in many others (e.g., South American and Middle-East countries) the correlation is negative. Moreover, the between-countries correlation of the correlation of University-Native with Native-immigrant achievement gap is only $r=0.21$ ($p < 0.1$). Therefore, we consider it appropriate to study the immigrant achievement gap fully detached from its interaction with the student SES.

Gender Achievement Gap. To measure gender, we use the variable available on the PISA student dataset. The proportion of girls is mostly balanced across countries, ranging from 47% to 53% (see Table 1).

Table 1. Proportion of social groups and subsamples N, N of students, N of schools by country

	University education		Immigrant background		Girl		Total N	
Country	%	N	%	N	%	N	Students	Schools
Albania	20%	1,310	1%	40	49%	3,167	6,359	327
Argentina	34%	4,450	5%	654	51%	6,232	11,975	455
Australia	53%	6,722	28%	3433	49%	7,075	14,273	763
Austria	30%	2,019	23%	1413	49%	3,321	6,802	291
B-S-J-Z (China)	21%	3,790	0%	21	48%	5,775	12,058	361
Baku (AZ)	23%	1,518	5%	340	47%	3,262	6,827	197
Belarus	43%	2,505	4%	228	48%	2,772	5,803	234
Belgium	49%	4,025	18%	1508	50%	4,271	8,475	288
Bosnia and Herzegovina	22%	1,429	3%	182	49%	3,148	6,480	213
Brazil	32%	3,260	1%	60	50%	5,478	10,691	597
Brunei Darussalam	35%	2,379	8%	555	50%	3,383	6,828	55
Bulgaria	49%	2,563	1%	70	47%	2,533	5,294	197
Canada	60%	12,708	35%	5667	50%	11,307	22,653	821
Chile	34%	3,200	4%	258	49%	3,814	7,621	254
Chinese Taipei	26%	1,855	1%	52	50%	3,624	7,243	192
Colombia	24%	1,831	1%	43	51%	3,857	7,522	247
Costa Rica	43%	3,043	10%	725	51%	3,618	7,221	205
Croatia	36%	2,370	9%	601	50%	3,311	6,609	183
Czech Republic	31%	2,501	4%	253	49%	3,518	7,019	333
Denmark	73%	5,205	11%	1578	50%	3,816	7,657	348
Dominican Republic	41%	2,244	3%	155	50%	2,890	5,674	235
Estonia	46%	2,432	10%	543	50%	2,651	5,316	230
Finland	63%	3,503	6%	319	49%	2,772	5,649	214
France	47%	2,843	14%	959	49%	3,078	6,308	252
Georgia	61%	3,437	1%	76	48%	2,682	5,572	321
Germany	38%	1,712	22%	1055	46%	2,525	5,451	223
Greece	47%	3,017	12%	710	49%	3,178	6,403	242
Hong Kong	22%	1,225	38%	2202	49%	2,955	6,037	152
Hungary	48%	2,516	3%	125	50%	2,605	5,132	238
Iceland	69%	2,210	6%	181	50%	1,656	3,296	142
Indonesia	23%	3,173	0%	36	51%	6,240	12,098	397
Ireland	47%	2,574	18%	983	50%	2,777	5,577	157
Israel	57%	3,638	17%	1021	53%	3,544	6,623	174
Italy	36%	4,134	10%	1080	48%	5,680	11,785	542
Japan	48%	2,845	1%	39	51%	3,120	6,109	183
Jordan	37%	3,267	12%	1612	51%	4,619	8,963	313
Kazakhstan	33%	6,667	8%	1434	49%	9,576	19,507	616
Kosovo	40%	1,983	1%	68	50%	2,457	5,058	211
Latvia	45%	2,301	4%	246	51%	2,685	5,303	308
Lebanon	30%	1,641	6%	302	54%	3,079	5,614	313
Lithuania	50%	3,332	2%	150	49%	3,377	6,885	362
Luxembourg	45%	2,201	55%	2828	49%	2,594	5,230	44
Macao	25%	943	63%	2371	49%	1,862	3,775	45
Malaysia	29%	1,771	2%	97	51%	3,131	6,111	191

	University education		Immigrant background		Girl		Total N	
Country	%	N	%	N	%	N	Students	Schools
Malta	40%	1,325	9%	288	48%	1,612	3,363	50
Mexico	27%	1,889	2%	88	52%	3,826	7,299	286
Moldova	37%	2,012	2%	80	49%	2,621	5,367	236
Montenegro	42%	2,811	6%	392	48%	3,240	6,666	61
Morocco	21%	1,365	1%	57	48%	3,262	6,814	179
Netherlands	67%	3,083	14%	695	50%	2,330	4,765	156
New Zealand	45%	2,664	27%	1623	50%	3,154	6,173	192
North Macedonia	42%	2,267	2%	80	48%	2,596	5,569	117
Norway	41%	2,253	12%	696	49%	2,880	5,813	251
Panama	36%	2,210	6%	343	50%	3,173	6,270	253
Peru	32%	1,920	1%	31	49%	3,000	6,086	340
Philippines	28%	2,004	1%	74	53%	3,868	7,233	187
Poland	37%	2,042	1%	27	50%	2,857	5,625	240
Portugal	37%	2,061	7%	344	49%	2,944	5,932	276
Qatar	71%	9,604	57%	7476	49%	6,954	13,828	188
Romania	31%	1,532	1%	30	48%	2,444	5,075	170
Russian Federation	63%	4,824	6%	434	50%	3,861	7,608	263
Saudi Arabia	37%	2,301	12%	698	48%	2,992	6,136	234
Serbia	37%	2,394	9%	612	49%	3,272	6,609	187
Singapore	45%	2,907	25%	1555	49%	3,277	6,676	166
Slovak Republic	40%	2,409	1%	69	50%	3,002	5,965	376
Slovenia	42%	2,388	9%	578	49%	2,993	6,401	345
Spain	46%	17,311	12%	4170	49%	17,956	35,943	1,089
Sweden	57%	3,000	21%	1077	50%	2,763	5,504	223
Switzerland	38%	2,173	34%	1954	47%	2,789	5,822	228
Thailand	26%	2,880	1%	70	53%	4,693	8,633	290
Turkey	23%	1,569	1%	54	50%	3,396	6,890	186
Ukraine	39%	2,353	2%	146	47%	2,857	5,998	250
United Arab Emirates	70%	12,720	56%	9671	51%	9,380	19,277	755
United Kingdom	48%	5,870	20%	1786	51%	6,996	13,818	471
United States	53%	2,453	23%	1011	49%	2,376	4,838	164
Uruguay	19%	1,013	1%	64	52%	2,732	5,263	189

Note. N indicates the N of the subsamples within each category, i.e., N of students with parents with university education, N of students with immigrant background and N of girls. Proportions are weighted.

4.3.2 Country-level Covariates

To validate the empirical differentiability of the three types of inequality, we use some key correlates of student achievement and social inequality in achievement commonly used in previous studies. This section is not intended to bring substantive conclusions, but to complement the previous analyses; if the regression models differ across the different types of social inequality, it brings some evidence on how this discussion have consequences on substantive educational research too. Some information is derived from the PISA school principal questionnaire, which is also representative of each country's school system, while other variables are derived from external sources.

GDP per Capita. To indicate a country's economic wealth, we used gross domestic product (GDP) per capita. This information is based on the World Bank database (World Bank, 2022), and we used the latest information for each country or region (up to 2018).

Growth Mindset. To capture cultural differences across countries, we used the variable growth mindset, available on the PISA 2018 student dataset. 'Growth mindset' refers to the belief that someone's ability and intelligence can be developed over time (OECD, 2019c). Within each country, we averaged the percentage of students that strongly disagree or disagree with the statement "Your intelligence is something about you that you can't change very much".

Between-school Tracking Age. This variable indicates at what age (based on the modal age for the corresponding grade) students are placed into different school tracks. Different tracks typically have different curricula, and the transition from a comprehensive to a tracked school system constitutes an important event in students' educational careers. We followed the information indicated in Strello et al. (2021), complemented by our own elaboration based on UNESCO-IBE's World Data on Education (UNESCO-IBE, 2012).

Selectiveness. Besides tracking, we also included two indicators of the degree of selectiveness within educational systems: the importance for school admission of (1) Students' record of academic performance (including placement tests) and (2) Residence in a particular area. For each of these we calculate the percentage per country of school principals that declare they *Always* (vs. *Sometimes* or *Never*) consider these factors in school admissions. A larger percentage of the first item is an indicator of a more selective system, while a larger percentage of the latter item is an indicator of a less selective system.

Grade Repetition. We included the percentage of students that had repeated a grade in their school course, using information on the PISA 2018 student questionnaire aggregated to the country level.

5. Results

5.1 Social Achievement Gaps Across Countries

As a preliminary step, we describe the social achievement gaps by social category. SES achievement gaps consist of the mean difference between high SES and low SES in standardized mathematics achievement scores. Figure 1a shows that the vast majority of countries present a positive and significant SES achievement gap. The only exceptions are the Philippines and Kazakhstan with a negative gap; and Lebanon, Baku (AZ), Brunei Darussalam, and Albania with a non-significant gap. Among those with significant positive achievement gaps, there is a high variability on the magnitude of these gaps; Norway and Indonesia have an SES achievement gap of 0.14 SD, while Belarus and Vietnam have SES achievement gaps of 0.66 SD and 0.74 SD, respectively.

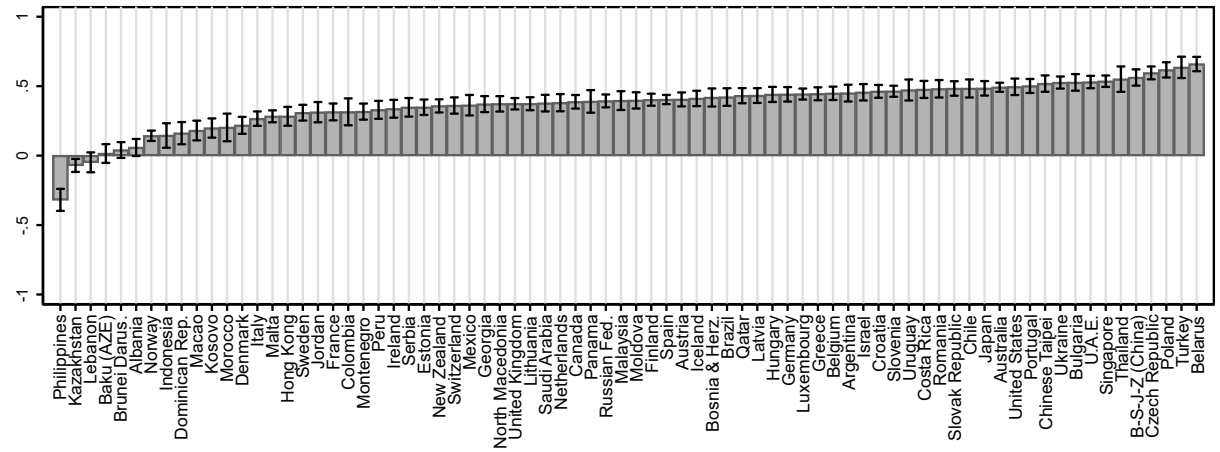
The immigration achievement gap is calculated as the mean score difference between native students (without an immigrant background) and students with an immigrant background.

Immigration achievement gaps are present in most countries on where there is a positive gap (i.e., natives perform better than immigrants), although several countries present a negative gap (i.e., immigrants perform better than natives), and many others where the differences are non-significant (see Figure 1b). The range of the magnitude of the immigration achievement gaps is also larger than for SES achievement gaps, from a negative gap of -0.8 SD in United Arab Emirates to around 1.50 SD in Indonesia. Moreover, the unbalanced shares of natives and immigrant subsamples imply large confidence intervals in some countries, so the estimations of gaps in this category are less efficient than on SES and gender.

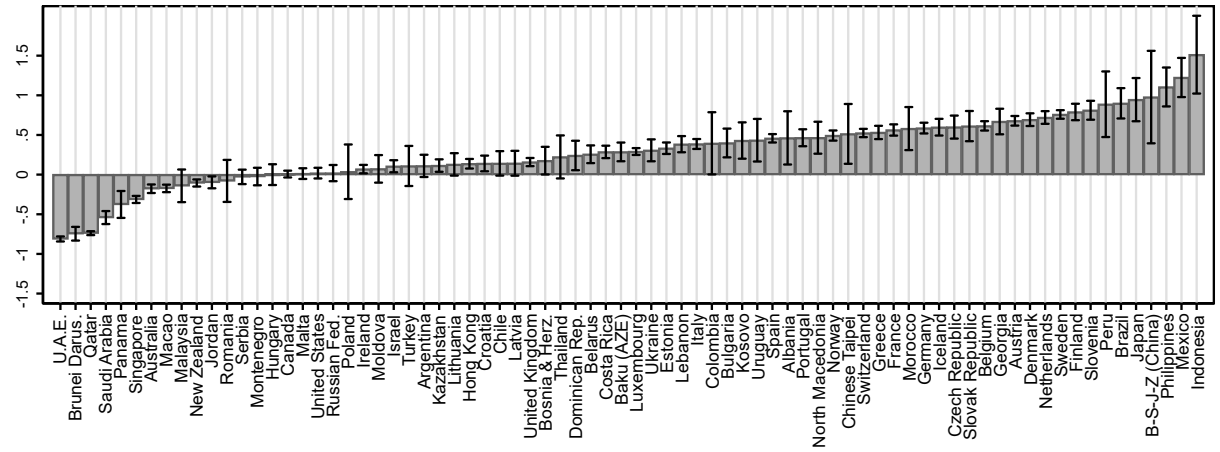
The gender achievement gap is calculated as the mean score difference between boys and girls. A positive gap indicates a higher mean score for boys than girls. One contrast with both previous measures is the smaller range of achievement gaps overall, from a -0.24 SD gap in Qatar to a 0.24 SD gap in Colombia (see Figure 1c). While most countries present positive gaps (boys achieve better mathematics scores than girls), there are negative gaps in many countries, with girls achieving better mathematics scores than boys.

Before we explore the association between the three types of social inequality with the full sample of 76 countries, we take a closer look at individual countries. In Turkey, the SES achievement gap is very high, whereas the immigration and gender gaps are small compared to the other countries. In Italy, the gaps for SES, immigration, and gender are low, medium, and high, respectively. Such patterns suggest that social inequality must be understood multidimensionally, since certain types of inequality are typically higher than others within the same country. Accordingly, a single type of social inequality is insufficient to conclude that social inequality in performance is generally low or high in a country.

a. Socioeconomic achievement gaps (High SES – Low SES)



b. Immigration achievement gaps (Native – Immigrant background)



c. Gender achievement gaps (Boys – Girls)

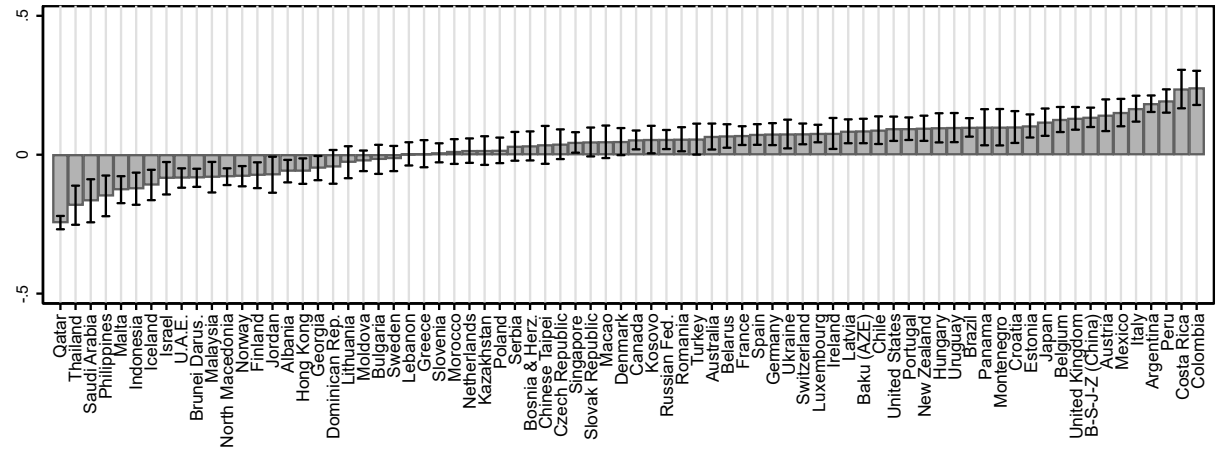


Figure 1. SES, Immigrant, and Gender achievement gaps

Note. Confidence intervals at 95% confidence level. Y axis are on different scale between each plot. Available as table format in Appendix (Table 3).

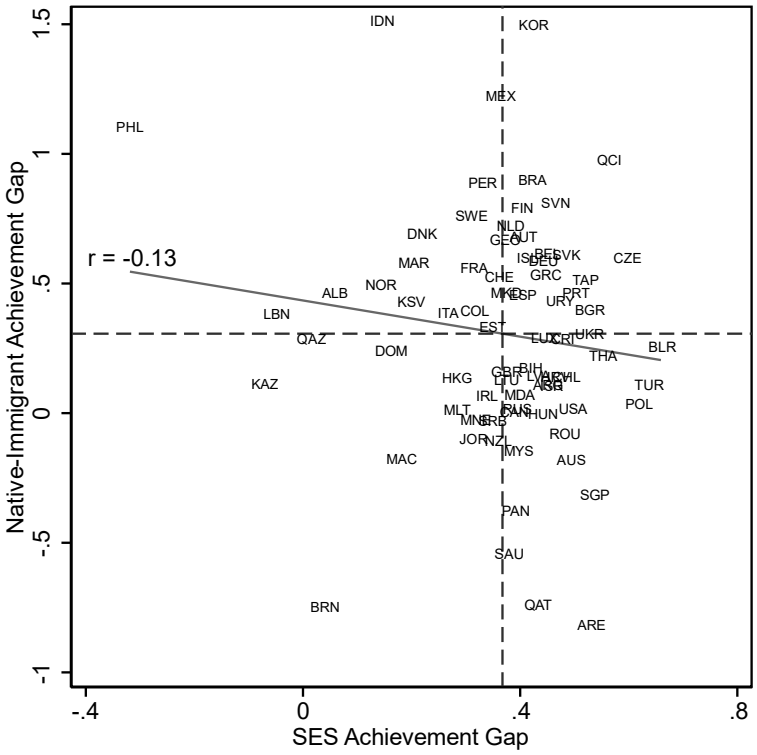
5.2 Correlation of SES, Immigration, and Gender Achievement Gaps Across Countries

The correlational analyses, including all 76 countries, confirm the need to distinguish between SES, immigration, and gender achievement gaps. The correlation between SES and immigration achievement gaps is r=-.13 (non-significant [n.s.]), between SES and gender gaps it is

r=.24 ($p < 0.05$), and between immigration and gender gaps it is r=.18 (n.s.). Figure 2 plots the associations between the three performance gaps, with no evidence of non-linear relationships.

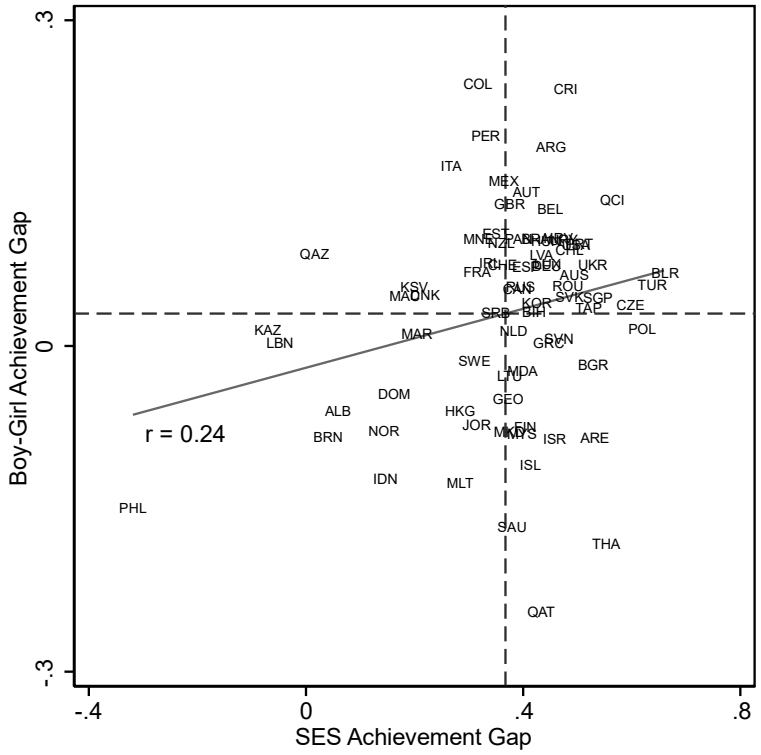
In countries with a small number of immigrants, the gap between immigrant and native students is measured with lower reliability (see the large confidence intervals for some countries in Figure 2b). To address this issue, we restricted the sample to the n=53 countries with more than 100 immigrant students[4]. The results of the correlation analyses with the restricted sample are qualitatively the same as with the total sample. The correlation between SES and native-immigrant achievement gap remains insignificant and the correlation estimate is even lower (r=0.03; n.s.). The correlation between immigrant achievement gap and gender achievement gap is 0.31 ($p < 0.05$). There is a small correlation when removing countries with less precise estimations of the immigrant achievement gap; in this sample there is a small tendency of countries with advantages for native students over immigrants to also show advantages for boys over girls. The correlation between the SES achievement gap and gender achievement gap is lower and is statistically insignificant (r=.19; n.s.). The analyses in the following section use the full sample of countries.

a. Correlation SES achievement gap – Immigrant achievement gap



[4] Countries excluded from the restricted sample of n=53 are Poland, Romania, Peru, Indonesia, Japan, Albania, Colombia, Chinese Taipei, Turkey, Morocco, Brazil, Uruguay, Kosovo, Slovak Republic, Bulgaria, Thailand, Philippines, Georgia, Moldova, North Macedonia, Mexico, and Malaysia.

b. Correlation SES achievement gap – Gender achievement gap



c. Correlation Gender achievement gap – Immigrant achievement gap

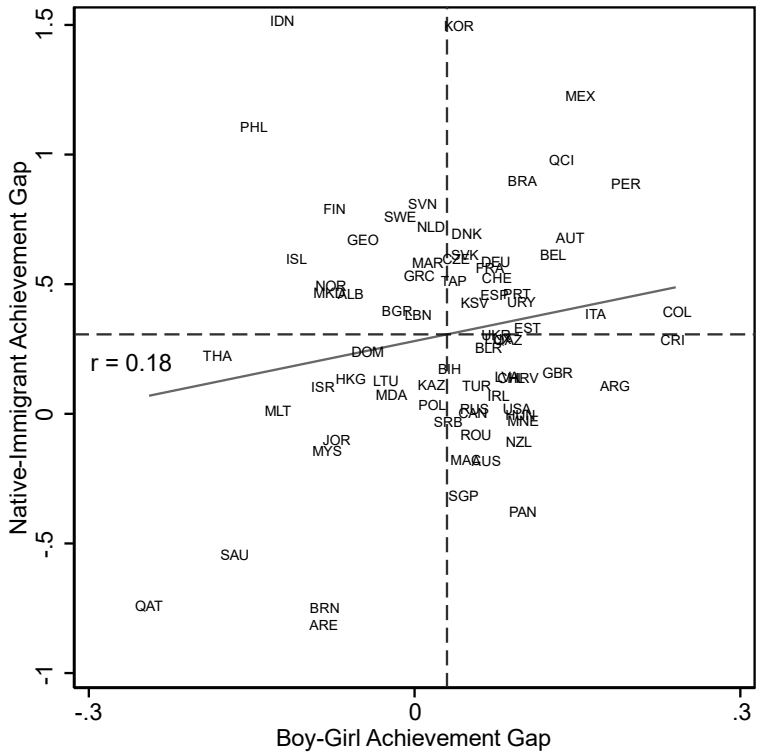


Figure 2. Correlation between achievement gaps

Notes. Horizontal and vertical lines represent the between-countries mean of axis Y and axis X achievement gaps respectively. Solid line represents the correlation using the full sample of countries (N=76).

5.3 Nomological Networks

In the previous section, we presented evidence that the three achievement gaps—by SES, immigration background, and gender—are largely uncorrelated. In this section, as validity analyses that complement the previous section, we explore how the different achievement gaps are associated with different country-level features. If institutional features are associated differently with each achievement gap, this would provide further evidence of the need to differentiate between these three types of inequality when analyzing social inequality. We do not aim to bring substantive conclusions, that would require more theoretical development and a more complex analysis design.

Table 2 shows the results from three regression analyses, where we regressed the three measures of social inequalities on the set of country-level institutional features of education systems. For easier interpretation of the results, we present standardized regression estimates. Models estimated variable-by-variable are available in Appendix, Tables 4 to 6. The comparison reveals that institutional characteristics better explain variation in the gender achievement gap (explaining 48% of the international variation) than in the immigrant achievement gap (32%) and the SES achievement gap (22%).

The main finding of the comparison of regression parameters is that institutional characteristics are differentially associated with different social achievement gaps, providing further evidence that a holistic evaluation of social inequality requires a consideration of different gaps. For example, the economic power of countries, measured as GDP per capita, is associated negatively with the immigrant achievement gap and the gender gap, but it is not associated with the SES achievement gap. The growth mindset cultural indicator is associated with the SES achievement gap and the gender achievement gap, but not with the immigration gap.

Regarding selectivity, using residence as a criteria for selection is negatively associated with SES and gender achievement gaps. Counterintuitively, selecting by performance is only associated with a reduction in the immigration gap. A later tracking is significantly negatively associated only with the immigration gap.

Education systems with a higher percentage of repeating students tend to show a lower SES gap and a higher gender gap.

Table 2. OLS Models on Achievement Gaps

	(1) High – Low SES	(2) Native – Immigrant	(3) Boy – Girl
GDP per Capita	0.037	-0.474***	-0.373***
	(0.008)	(0.019)	(0.004)
Growth Mindset	0.241*	0.032	0.563***
	(0.191)	(0.457)	(0.089)
Tracking Age	-0.188	-0.290**	-0.159
	(0.012)	(0.029)	(0.006)
Selection by Residence	-0.246*	-0.068	-0.274***
	(0.100)	(0.238)	(0.047)
Selection by Performance	-0.203	-0.265*	-0.030
	(0.095)	(0.228)	(0.044)
Repeated grade	-0.201*	0.115	0.316***
	(0.168)	(0.402)	(0.079)
N	73	73	73
r2	0.216	0.303	0.458
r2-adj	0.145	0.240	0.409

Note. Standardized beta coefficients are reported. Standard errors in parentheses. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

6. Discussion

In this study, we aimed to explore the degree to which there is one *umbrella* concept of social inequality, or whether there are substantially different concepts of social inequalities. We explored the correlations between three different social inequalities in achievement: SES, immigration status, and gender. We also compared how different education system-level covariates are associated with each achievement gap.

We highlight several points. First, at least one category of achievement gaps can be observed in every country. Second, while SES gaps were observed in all but four countries (as well as their direction), the size and direction of the immigration and gender gaps vary across countries. In most countries, natives and boys have better mean performance in mathematics than immigrants and girls, but there are several countries where immigrants and girls have an advantage. The variation across countries in the achievement gap by immigration is clearly higher than in the SES and, especially, the gender achievement gap. Also, the share of immigrants is very low in many countries, making it both empirically difficult to study (due the lower efficiency of the estimations) and a less prominent problem in some regions of the world. These findings suggest that, while SES inequality appears to be an almost global phenomenon, immigration and gender are associated with educational disadvantages differently across different countries and regions. Based on these findings, we conclude that the institutional context and social practices in different countries play a role in shaping social inequality. In the following section, we support this interpretation with additional findings.

Second, there is hardly any correlation between the three achievement gaps. This means that one education system can be egalitarian in some category, but profoundly unequal in another. To

properly assess how unequal or egalitarian education systems are, policy-makers, researchers, and other stakeholders need to consider and address different indicators of social inequality.

Third, this lack of correlation is also related to how we study these inequalities. Using the same sample of countries and the same covariates, we showed that each achievement gap is associated with a different set of institutional features. Researchers who aim to study the impact of institutional characteristics on social inequality from a holistic perspective are advised to consider different forms of social inequality. Conclusions from a study on one gap cannot be generalized to other gaps.

6.1 Limitations and future research

This article has some limitations. One set of limitations relate to the measures we considered in the present study. We have only considered three key social categories here—SES, immigration, and gender—but there are other important categories (such as religiosity and ethnicity). These categories are often not highlighted in international assessments, and more comprehensive data is required to explore the gaps associated with these categories. Another limitation concerns the indicators used to measure SES, immigration, and gender. For the sake of simplicity, we considered only one indicator per category; nevertheless we recognize that these indicators have limitations, as there are more ways of operationalizing both immigration and SES. Also, we have not explored intersectionality among the three categories. For example, boys with a migration background may be a particularly disadvantaged group. Such analyses are beyond the scope of this paper but appear important for further research.

It is important to mention that these results only refer to mathematics achievement. There are other cases where achievement gaps could be of different magnitude or different direction. For example, looking at the latest international reports of PIRLS (Mullis et al., 2017) and PISA (OECD, 2019a), girls score significantly higher than boys in most countries, while in no country boys score better than girls. We focused on the achievement gap in mathematics as it illustrated the best the differences between achievement gaps.

Another set of limitations relates to the analysis of the institutional covariates, and natural limitations to the samples in some countries. The analysis of the covariates is based on cross-sectional data, and for this reason we do not make causal inferences. However, such analyses of nomological networks provide useful evidence for the distinction of social inequalities. Furthermore, the results involving immigration gaps are particularly affected by small subsamples of immigrants in certain countries where immigration is uncommon, lowering the measurement efficiency. Lastly, while PISA samples hundreds of thousands of students, the number of countries remains a natural limitation in any cross-national research.

6.2 Conclusion

In conclusion, mind the gap, but consider what gap you are looking at, as not all gaps are equal; depending on the social category, the results are very different. Ranking countries in terms of

just one social inequality category provides a limited picture, at best. SES inequality is mostly a global problem, but immigration is more relevant in some regions than others, while gender gaps follow opposite direction between countries. This has direct consequences on the evaluation of education systems, and on research.

7. References

Andon, A., Thompson, C. G., & Becker, B. J. (2014). A quantitative synthesis of the immigrant achievement gap across OECD countries. *Large-Scale Assessments in Education*, *2*(1). https://doi.org/10.1186/s40536-014-0007-2

Ayalon, H., & Livneh, I. (2013). Educational standardization and gender differences in mathematics achievement: A comparative study. *Social Science Research*, *42*(2), 432–445. https://doi.org/10.1016/j.ssresearch.2012.10.001

Bodovski, K., & Munoz, I. G. (2020). Do education system characteristics moderate the socioeconomic, gender and immigrant gaps in math and science achievement? *International Journal of Sociology of Education*, *9*(2), 122–154. https://doi.org/10.17583/rise.2020.4807

Bouhlila, D. S. (2017). Parents' education and literacy skills: Evidence on inequality of socioeconomic status in Arab countries. *World Development Perspectives*, *5*, 34–43. https://doi.org/10.1016/j.wdp.2017.02.006

Bourdieu, P. (1986). The forms of capital. In J. Richardson (Ed.), *Handbook of Theory and Research for the Sociology of Education* (pp. 241–258). Greenwood.

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

Broer, M., Bai, Y., & Fonseca, F. (2019). *Socioeconomic Inequality and Educational Outcomes* (Vol. 5). Springer International Publishing. https://doi.org/10.1007/978-3-030-11991-1

Buchmann, C., & Parrado, E. A. (2006). Educational achievement of immigrant-origin and native students: A comparative analysis informed by institutional theory. In *International Perspectives on Education and Society* (Vol. 7, pp. 335–366). https://doi.org/10.1016/S1479-3679(06)07014-9

Chmielewski, A. K. (2019). The Global Increase in the Socioeconomic Achievement Gap, 1964 to 2015. *American Sociological Review*, 000312241984716. https://doi.org/10.1177/0003122419847165

Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. *American Journal of Sociology*, *94*, S95–S120. http://www.jstor.org/stable/2780243

Coleman, J. S. (1990). *Foundations of Social Theory*. The Belknap of Harvard University Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. doi:10.1037/h0040957

Dronkers, J., & Levels, M. (2007). Do school segregation and school resources explain region-of-origin differences in the mathematics achievement of immigrant students? *Educational Research and Evaluation*, *13*(5), 435–462. https://doi.org/10.1080/13803610701743047

Eccles, J. S., Jacobs, J. E., & Harold, R. D. (1990). Gender Role Stereotypes, Expectancy Effects, and Parents' Socialization of Gender Differences. *Journal of Social Issues*, *46*(2), 183–201. https://doi.org/10.1111/j.1540-4560.1990.tb01929.x

Ferreira, F. H. G., & Gignoux, J. (2014). The Measurement of educational inequality: Achievement and opportunity. *World Bank Economic Review*, *28*(2), 210–246. https://doi.org/10.1093/wber/lht004

González de San Román, A., & de La Rica, S. (2016). Gender Gaps in PISA Test Scores: The Impact of Social Norms and the Mother's Transmission of Role Attitudes. *Estudios de Economía Aplicada*. http://www.redalyc.org/articulo.oa?id=30143731005

Guiso, L., Monte, F., & Sapienza, P. (2008). Differences in Test Scores Correlated with Indicators of Gender Equality. *Science*, *320*(May), 1–2.

Halpern, D. F. (2012). Sex differences in cognitive abilities, 4th ed. In *Sex differences in cognitive abilities, 4th ed.* Psychology Press.

Hermann, Z., & Kopasz, M. (2019). Educational policies and the gender gap in test scores: a cross-country analysis. *Research Papers in Education*, *00*(00), 1–22. https://doi.org/10.1080/02671522.2019.1678065

Hopfenbeck, T. N., Lenkeit, J., el Masri, Y., Cantrell, K., Ryan, J., & Baird, J. A. (2018). Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research*, *62*(3), 333–353. https://doi.org/10.1080/00313831.2016.1258726

Hyde, J. S. (2014). Gender Similarities and Differences. *Annual Review of Psychology*, *65*(1), 373–398. https://doi.org/10.1146/annurev-psych-010213-115057

Jackson, M. (2012). Bold choices: How ethnic inequalities in educational attainment are suppressed. *Oxford Review of Education*, *38*(2), 189–208. https://doi.org/10.1080/03054985.2012.676249

Jerrim, J., Volante, L., Klinger, D. A., & Schnepf, S. v. (2019). Socioeconomic Inequality and Student Outcomes Across Education Systems. In *Socioeconomic Inequality and Student Outcomes: Cross-National Trends, Policies, and Practices* (pp. 3–16). Springer. https://doi.org/10.1007/978-981-13-9863-6_1

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

Lenkeit, J., Schwippert, K., & Knigge, M. (2017). Configurations of multiple disparities in reading performance: longitudinal observations across France, Germany, Sweden and the United Kingdom. *Assessment in Education: Principles, Policy and Practice*, *25*(1), 52–86. https://doi.org/10.1080/0969594X.2017.1309352

Levels, M., Kraaykamp, G., & Dronkers, J. (2008). Immigrant children's educational achievement in western countries: Origin, destination, and community effects on mathematical performance. *American Sociological Review*, *73*(5), 835–853. https://doi.org/10.1177/000312240807300507

Marks, G. N. (2008). Accounting for the gender gaps in student performance in reading and mathematics: evidence from 31 countries. *Oxford Review of Education*, *34*(1), 89–109. https://doi.org/10.1080/03054980701565279

Meinck, S., & Brese, F. (2019). Trends in gender gaps: using 20 years of evidence from TIMSS. *Large-Scale Assessments in Education*, *7*(1). https://doi.org/10.1186/s40536-019-0076-3

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fischbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science*. TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/international-results/

Mullis, I. V. S., Martin, M. O., Foy, Pierre., & Hooper, M. (2017). *PIRLS 2016 International Results in Reading*. IEA TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. http://timssandpirls.bc.edu/pirls2016/international-results/

Nauck, B. (2019). Ethnic inequality in educational attainment. In *Research Handbook on the Sociology of Education* (pp. 499–518). Edward Elgar Publishing. https://doi.org/10.4337/9781788110426.00038

Neuville, E., & Croizet, J.-C. (2007). Can salience of gender identity impair math performance among 7–8 years old girls? The moderating role of task difficulty. *European Journal of Psychology of Education*, *22*(3), 307–316. https://doi.org/10.1007/BF03173428

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., Bar-Anan, Y., Bergh, R., Cai, H., Gonsalkorale, K., Kesebir, S., Maliszewski, N., Neto, F., Olli, E., Park, J., Schnabel, K., Shiomura, K., Tulbure, B. T., Wiers, R. W., … Greenwald, A. G. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, *106*(26), 10593–10597. https://doi.org/10.1073/pnas.0809921106

OECD. (2019a). PISA 2018 Results (Volume II): Where All Students Can Succeed. In *OECD Publishing: Vol. II*. https://www.oecd.org/pisa/publications/PISA2018_CN_IDN.pdf

OECD. (2019b). *PISA 2018 Technical Report*. https://www.oecd.org/pisa/data/pisa2018technicalreport/

OECD. (2019c). *PISA 2018 Results (Volume III): What School Life Means for Students' Lives*. OECD. https://doi.org/10.1787/acd78851-en

Reilly, D. (2012). Gender, culture, and sex-typed cognitive abilities. *PLoS ONE*, *7*(7), 15–16. https://doi.org/10.1371/journal.pone.0039904

Rosén, M., Steinmann, I., & Wernersson, I. (2022). Gender Differences in School Achievement. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International Handbook of Comparative Large-Scale Studies in Education* (pp. 1–48). Springer. https://doi.org/10.1007/978-3-030-38298-8_46-1

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. 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

Scheeren, L., & Bol, T. (2022). Gender inequality in educational performance over the school career: The role of tracking. *Research in Social Stratification and Mobility*, *77*. https://doi.org/10.1016/j.rssm.2021.100661

Schmid, C. L. (2001). Educational achievement, language-minority students, and the new second generation. *Sociology of Education*, *74*, 71. https://doi.org/10.2307/2673254

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

Söhn, J., & Özcan, V. (2006). The Educational Attainment of Turkish Migrants in Germany. *Turkish Studies*, *7*(1), 101–124. https://doi.org/10.1080/14683840500520626

Steinmann, I., & Rutkowski, L. (2023). The link between gender gaps in school enrollment and school achievement. *Comparative Education Review*. https://doi.org/10.1086/725395

Steinmann, I., Strietholt, R., & Rosén, M. (2023). International reading gaps between boys and girls, 1970–2016. *Comparative Education Review, 67*(2), 298-330.

Stoet, G., & Geary, D. C. (2013). Sex Differences in Mathematics and Reading Achievement Are Inversely Related: Within- and Across-Nation Assessment of 10 Years of PISA Data. *PLoS ONE*, *8*(3), e57988. https://doi.org/10.1371/journal.pone.0057988

Strand, S. (2011). The limits of social class in explaining ethnic gaps in educational attainment. *British Educational Research Journal*, *37*(2), 197–229. https://doi.org/10.1080/01411920903540664

Strand, S. (2014). Ethnicity, gender, social class and achievement gaps at age 16: Intersectionality and "getting it" for the white working class. *Research Papers in Education*, *29*(2), 131–171. https://doi.org/10.1080/02671522.2013.767370

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

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* (Issue January, pp. 51–58). Waxmann Verlag.

Strietholt, R., Gustafsson, J.-E., Hogrebe, N., Rolfe, V., Rosén, M., Steinmann, I., & Hansen, K. Y. (2019). The Impact of Education Policies on Socioeconomic Inequality in Student Achievement: A Review of Comparative Studies. In *Socioeconomic Inequality and Student Outcomes: Cross-National Trends, Policies, and Practices* (pp. 17–38). Springer. https://doi.org/10.1007/978-981-13-9863-6_2

Teltemann, J., & Schunck, R. (2016). Education systems, school segregation, and second-generation immigrants' educational success: Evidence from a country-fixed effects approach using three waves of PISA. *International Journal of Comparative Sociology*, *57*(6), 401–424. https://doi.org/10.1177/0020715216687348

UNESCO. (2018). *Handbook on Measuring Equity in Education*. https://doi.org/10.1016/S0733-8619(03)00096-3

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 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 Hek, M., Buchmann, C., & Kraaykamp, G. (2019). Educational systems and gender differences in reading: A comparative multilevel analysis. *European Sociological Review*, *35*(2), 169–186. https://doi.org/10.1093/esr/jcy054

van Langen, A., Bosker, R., & Dekkers, H. (2006). Exploring cross-national differences in gender gaps in education. *Educational Research and Evaluation*, *12*(2), 155–177. https://doi.org/10.1080/13803610600587016

World Bank. (2022). *GDP per capita (current US$)*. https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

Worrell, F. C. (2014). Theories school Psychologists should know: Culture and academic achievement. *Psychology in the Schools*, *51*(4), 332–347. https://doi.org/10.1002/pits.21756

Zell, E., Krizan, Z., & Teeter, S. R. (2015). Evaluating gender similarities and differences using metasynthesis. *American Psychologist*, *70*(1), 10–20. https://doi.org/10.1037/a0038208

6 Appendix

Table 3. Achievement gaps, correlation between social indicators, by country

	Achievement gap			Correlation		
Country	High – Low SES	Native - Immigrant	Boy - Girl	University - Native	University - Boy	Native - Boy
Albania	.06	.46*	-.06*	-.12*	.07*	-.06
Argentina	.45*	.11	.18*	.13*	.09*	.01
Australia	.49*	-.18*	.07*	-.21*	.02*	.00
Austria	.40*	.68*	.14*	.10*	.02	.00
B-S-J-Z (China)	.56*	.98*	.13*	-.04	.01	.02
Baku (Azerbaijan)	.01	.29*	.09*	-.02	.11*	.00
Belarus	.66*	.26*	.07*	.09*	.00	-.03
Belgium	.45*	.61*	.13*	.12*	.04*	-.03*
Bosnia and Herzegovina	.42*	.18*	.03	-.17*	.09*	-.04
Brazil	.42*	.90*	.10*	-.14*	.07*	.08*
Brunei Darussalam	.04	-.75*	-.08*	-.11*	.16*	-.00
Bulgaria	.53*	.40*	-.02	-.08*	.07*	-.12*
Canada	.39*	.01	.05*	-.12*	.04*	-.03*
Chile	.48*	.14	.09*	-.10*	.08*	.08*
Chinese Taipei	.52*	.51*	.04	-.15*	.03	.10*
Colombia	.32*	.39*	.24*	-.03	.11*	.10*
Costa Rica	.48*	.29*	.24*	.19*	.08*	.00
Croatia	.46*	.14*	.10*	.11*	.09*	.01
Czech Republic	.60*	.60*	.04	-.07*	.03*	.02
Denmark	.22*	.69*	.05	.25*	-.02	-.02
Dominican Republic	.16*	.24*	-.04	-.11*	.13*	-.09*
Estonia	.35*	.33*	.10*	-.13*	-.01	-.10*
Finland	.40*	.79*	-.07*	.12*	.00	.01
France	.31*	.56*	.07*	.11*	.00	-.02
Georgia	.37*	.67*	-.05*	.05	.10*	-.08*
Germany	.44*	.59*	.07*	.16*	.01	-.02
Greece	.45*	.53*	.00	.22*	.08*	-.02
Hong Kong	.28*	.14*	-.06*	.26*	-.05*	-.06*
Hungary	.44*	-.00	.10*	-.08*	.09*	.04
Iceland	.41*	.60*	-.11*	.15*	.03	.10*
Indonesia	.14*	1.51*	-.12*	-.39*	.13*	-.13*
Ireland	.34*	.07*	.08*	-.15*	-.01	.04*
Israel	.46*	.11*	-.08*	-.03*	.13*	-.06*
Italy	.27*	.39*	.17*	.02	.07*	.04
Japan	.48*	.95*	.12*	-.05	.03	.01
Jordan	.31*	-.10*	-.07*	-.08*	.15*	-.00
Kazakhstan	-.07*	.11*	.01	.06*	.08*	-.01
Kosovo	.20*	.43*	.05*	-.09*	.07*	-.03
Latvia	.43*	.14	.08*	-.06*	.08*	-.09*
Lebanon	-.05	.38*	.00	-.20*	.16*	.02
Lithuania	.37*	.13	-.03	-.13*	.11*	-.04
Luxembourg	.44*	.29*	.08*	.12*	.05*	-.04*
Macao	.18*	-.17*	.05	.17*	.01	.05*
Malaysia	.40*	-.14	-.08*	.02	.08*	-.01
Malta	.28*	.01	-.13*	-.30*	.09*	.01
Mexico	.36*	1.22*	.15*	.07	.07*	.20*
Moldova	.40*	.07	-.02	-.17*	.08*	-.09*
Montenegro	.32*	-.02	.10*	-.17*	.16*	.00
Morocco	.20*	.58*	.01	-.33*	.12*	-.17*
Netherlands	.38*	.72*	.01	.22*	.04*	-.00

	Achievement gap			Correlation		
Country	High – Low SES	Native - Immigrant	Boy - Girl	University - Native	University - Boy	Native - Boy
New Zealand	.36*	-.11*	.10*	-.17*	.02	-.07*
North Macedonia	.37*	.47*	-.08*	-.03	.13*	.07*
Norway	.14*	.49*	-.08*	.06*	.05*	.03
Panama	.39*	-.38*	.10*	-.13*	.04*	-.00
Peru	.33*	.89*	.19*	-.18*	.02	.13*
Philippines	-.32*	1.11*	-.15*	-.28*	.18*	-.17*
Poland	.62*	.04	.02	-.16*	-.01	.09
Portugal	.50*	.46*	.09*	-.09*	.05*	-.01
Qatar	.43*	-.74*	-.24*	-.15*	.09*	-.04*
Romania	.48*	-.08	.06*	-.23*	.10*	-.01
Russian Federation	.39*	.02	.05*	-.02	.09*	.04
Saudi Arabia	.38*	-.54*	-.17*	-.07*	.08*	-.08*
Serbia	.35*	-.03	.03	-.04*	.08*	.05*
Singapore	.54*	-.31*	.04*	-.39*	-.02*	.02*
Slovak Republic	.48*	.61*	.05	-.14*	.03	-.07*
Slovenia	.46*	.81*	.01	.31*	.02	-.02
Spain	.40*	.46*	.07*	.13*	.01	.01
Sweden	.31*	.76*	-.01	.11*	.00	-.05*
Switzerland	.36*	.52*	.07*	.09*	.02*	.01
Thailand	.55*	.22	-.18*	.01	.11*	-.05
Turkey	.64*	.11	.06*	-.17*	.07*	-.01
Ukraine	.53*	.31*	.07*	-.11*	.01	-.15*
United Arab Emirates	.53*	-.81*	-.08*	-.32*	.09*	-.01
United Kingdom	.37*	.16*	.13*	-.10*	.05*	.05*
United States	.50*	.02	.09*	.22*	.09*	.00
Uruguay	.47*	.43*	.10*	-.17*	.09*	.02

Note. * = achievement gap / correlation significant $p < 0.05$. Correlations correspond to tetrachoric correlations between dichotomic indicators.

Table 4. OLS models on socioeconomic (High – Low SES) mathematics achievement gaps

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
GDP per Capita	0.037	0.106					
	(0.008)	(0.008)					
Growth Mindset	0.241*		0.341***				
	(0.191)		(0.154)				
Tracking Age	-0.188			-0.203*			
	(0.012)			(0.011)			
Selection by Residence	-0.246*				-0.115		
	(0.100)				(0.095)		
Selection by Performance	-0.203					-0.123	
	(0.095)					(0.077)	
Repeated grade	-0.201*						-0.192
	(0.168)						(0.170)
N	73	73	73	73	73	73	73
r2	0.216	0.011	0.116	0.013	0.015	0.041	0.037
r2-adj	0.145	-0.003	0.104	-0.001	0.001	0.028	0.023

Note. Standardized beta coefficients are reported. Standard errors in parentheses. Significance levels: * p < 0.1; ** p < 0.05; *** p < 0.01.

Table 5. OLS models on immigration (Native – Immigrant) mathematics achievement gaps

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
GDP per Capita	-0.474***	-0.417***					
	(0.019)	(0.018)					
Growth Mindset	0.032		-0.000				
	(0.457)		(0.415)				
Tracking Age	-0.290**			-0.122			
	(0.029)			(0.029)			
Selection by Residence	-0.068				-0.080		
	(0.238)				(0.241)		
Selection by Performance	-0.265*					-0.186	
	(0.228)					(0.193)	
Repeated grade	0.115						0.134
	(0.402)						(0.435)
N	73	73	73	73	73	73	73
r2	0.303	0.174	0.000	0.006	0.035	0.015	0.018
r2-adj	0.240	0.162	-0.014	-0.008	0.021	0.001	0.004

Note. Standardized beta coefficients are reported. Standard errors in parentheses. Significance levels: * p < 0.1; ** p < 0.05; *** p < 0.01.

Table 6. OLS models on gender (Boys - Girls) mathematics achievement gaps

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
GDP per Capita	-0.373***	-0.190					
	(0.004)	(0.004)					
Growth Mindset	0.563***		0.364***				
	(0.089)		(0.086)				
Tracking Age	-0.159			-0.212*			
	(0.006)			(0.006)			
Selection by Residence	-0.274***				-0.246**		
	(0.047)				(0.052)		
Selection by Performance	-0.030					-0.136	
	(0.044)					(0.043)	
Repeated grade	0.316***						0.246**
	(0.079)						(0.094)
N	73	73	73	73	73	73	73
r2	0.458	0.036	0.132	0.060	0.018	0.045	0.060
r2-adj	0.409	0.023	0.120	0.047	0.005	0.032	0.047

Note. Standardized beta coefficients are reported. Standard errors in parentheses. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Article 2. Socioeconomic Inequality in Achievement: Conceptual Foundations and Empirical Measurement

Rolf Strietholt, Andrés Strello

The version of the manuscript printed below is the preprint of the article published in in Nilsen, T., Stancel-Piątak, A., Gustafsson, JE. (eds) *International Handbook of Comparative Large-Scale Studies in Education*, (2022). https://doi.org/10.1007/978-3-030-38298-8_11-1

Abstract

The study of social inequality in student achievement is based on ideas of justice which are often not sufficiently explicated. Furthermore, there is a large set of measures used to quantify socioeconomic inequality in achievement. The first part of this chapter explains conceptual principles underlying measures of social inequality in achievement. For this purpose, we first introduce the concepts of adequacy and equality and discuss how social inequality extends them. In this respect, we emphasize the nature of education and its intrinsic, instrumental individual and societal value. The second part of the chapter discusses key measurement issues researchers deal with when studying achievement gaps between students of different socioeconomic status. We summarize research on commonly used indicators of socioeconomic background and compare children and parent reports. Different sets of statistical measures for continuous, categorical, single and multiple background variables are reviewed, and the distinction between relative and absolute inequality measures is discussed with a focus on the implications for cross-national comparisons, and trend studies within countries over time.

Conceptual foundations

To understand if and why we should be concerned about socioeconomic inequality in student achievement, it is important to briefly review some conceptual philosophical foundations relevant to studying inequality. To understand implicit assumptions in the study of social inequality in student achievement, we first discuss theoretical foundations of political philosophy. The analysis of inequality is based on normative assumptions about justice. In the following, we will try to make these assumptions explicit. We consider it important to develop an understanding that socioeconomic inequality is only one form of inequality.

The provision and distribution of goods such as education, health care, or wealth are public concerns. The notion of excellence and equity introduce a normative dimension to the discussion around the provision and distribution of public goods. On the one hand, the discussions are about the achieved levels of literacy, health, or income as measured by, for example, the mean achievement scores in international student assessments, the average life expectancy, or the gross domestic

product (GDP). On the other hand, the discourse is based on inequality as a measure by, for example, the achievement gap between privileged and disadvantaged children, unequal access to medical care, or the Gini coefficient for income inequality. While the term excellence implies that it is desirable to achieve high achievement levels of a public good (e.g., education, wealth, health etc.), the term equity suggests it is also desirable to minimize inequalities in the distribution of goods (e.g., education, wealth, health etc.). Next, we will discuss how "minimizing inequality" can have very different meanings.

Distributive Rules

The philosophical literature on inequality typically distinguishes between a) an object or public good that is distributed, and b) a distributive rule that is used to assess inequality (see Brighouse, Ladd, Loeb, & Swift, 2018). In this chapter we focus on the object 'education' (or more specifically on 'student achievement') but to illustrate that different distributive rules may be justified in different contexts, it is also useful to think about other public goods such as health care or wealth (Atkinson & Bourguignon, 2000, 2015; Van Doorslaer & Van Ourti, 2011). Distributive principles are central for the study of inequality because they define how to fairly distribute a public good among the members of a group, such as the students in an educational system or the citizens of a state. Different distributive rules are best illustrated with concrete examples. Without much argument, we will focus on three distributive rules: equality, adequacy, and social inequality. Three related popular measures of income inequality are the Gini index, the proportion of the population living in poverty, and the gender payment gap.

Equality in wealth is frequently measured by the Gini index; a coefficient of 0 indicates perfect equality, where everyone has the same income, and a value of 1 indicates maximal inequality where a single person has all possible income and everyone else has none. Another distributive rule is *adequacy*, which is closely related to poverty. A simple measure to quantify poverty is the proportion of people who do not reach a minimum income, however defined. Studying poverty implies that it is considered unfair to distribute income in such a way that some people do not receive a minimum income; but at the same time inequality above a certain threshold is not problematized. In fact, there are good reasons to justify that the variation in income above the poverty threshold is not considered problematic or even wanted. For example, the high incomes of surgeons may be a reward for prior investments into education or for taking responsibility for the lives of other people. Following liberal ideas of individual freedom and choice, it should be left to the individual to decide whether he or she wishes to study for a long period or take on a high level of responsibility. The key differences between equality and adequacy is that the latter concept introduces the distinction between unjust and just inequalities in some public goods.

Gender inequalities in income is yet another form of inequality; the decisive factor here being whether the income of men and women is different. In this example, we look at gender differences, but the same reasoning can also be applied to differences between race or socioeconomic status

(SES). The idea behind measuring gender differences is that an unequal distribution of income is not problematic per se, but only if there are systematic differences between men and women. There are two measures of gender inequality: the gap in the mean income of men and women, and the comparison of the share of men and women in poverty. These measures are somehow linked to the two distributive rules *equality* and *adequacy*. It should be noted, however, that both measures of gender inequalities do not problematize variation in income, nor poverty itself. If half the men and half the women live in poverty, there is no gender inequality, yet is this fair? If we do not wish for people to live in poverty, we should know how many people are living in poverty, regardless of whether they are men or women.

The distribution of public goods is a constant and contentious topic of the political debate. There are no universally valid principles which can be justified in a similar way everywhere, and at any time. Rather, different distribution rules can also be applied to other goods. In the USA, there is a controversial discussion as to whether all Americans should have health insurance. Other countries may have universal health insurance for all of their citizens, but there are discussions on which services are standard for all the insured citizens, and which services are offered as individual, additional benefits. Progressive taxation is a means to increase the tax burden of higher incomes, whilst a universal basic income is a measure to reduce poverty. In some cases, it may be justified to eliminate any differences, but at the same time, it is important to stress that such equalization largely ignores individual freedom (Nozick, 1974).

Different Goods, Different Rules

Hardly anyone would want to tell someone else to go to football at the weekend, but not to the opera. Some drink beer, others prefer wine. Inequalities are not necessarily an indicator of injustice but of a liberal society. At the same time, however, it would not be argued that in liberal societies all inequalities are legitimate. Poverty and access to (at least basic) health care are public concerns, whereas recreational activities are not. What about education? Education plays different roles for individual and social development (Robeynes, 2006; Strietholt, 2014). One can play the flute for pleasure, or to earn money in the orchestra. One can learn a foreign language for fun, or to earn money abroad. In other words, education can have an intrinsic or instrumental value. This applies both on an individual and collective level. If several languages are taught and spoken in a country, the instrumental value is the strengthening of the national economy in a globalized world. The ability to communicate with people from other countries is also an intrinsic value on a collective level, because cultural exchange is related to mutual understanding.

Based on the different roles and functions of education, different distributive rules provide a suitable conceptual framework to study educational inequality. For example, some people take great pleasure in studying classical literature. However, it hardly seems appropriate to demand that every student must read all works of Lessing, Schiller and Goethe schools. It could be argued that liberalism provides a useful approach to justice in this context, in which young people are free to engage in

literature, sports or technology. But what about basic reading literacy? Probably, egalitarianism would find more support here, since reading is a basic prerequisite for cultural, economic or political participation in our society. It seems hardly justifiable to require children (or their parents) to decide for themselves whether they want to learn to read or not.

International large-scale assessments test students at different levels, for example, the Trends in International Mathematics and Science Study (TIMSS) assesses students in grade four and eight and in the last year of secondary education. The so-called Programme for the International Assessment of Adult Competencies (PIAAC) tests adults between 16 and 65 years old. While all these assessments are about mathematics, they differ dramatically in difficulty. The test in primary schools assesses students' basic numeracy, for example simple equations with whole numbers while the tests at the end of secondary education assesses advanced mathematics, such as calculus.

Perhaps we propose it reasonable to demand that all children should perform at around the same level at the end of primary school (*equality*) but we do not demand equality at the end of secondary school. We find it reasonable to demand that all students in secondary school have a basic knowledge about simple equations to solve real world problems, but we do not demand that all students are proficient in calculus (*adequacy*). To level the playing field, it seems fair that all students acquire basic mathematical skills. If one agrees, the standard deviation of achievement scores in international school achievement studies provides a suitable measure of educational inequality. The extent to which students then decide to continue studying mathematics at school or university is an individual decision. Following this argumentation, the standard deviation is no longer a suitable measure, because here implicitly any variation in performance is seen as problematic (including high proficiency). Accordingly, it would be better to focus on how many students do not reach certain minimum standards such as being able to solve simple equations. Any variation beyond this threshold is not problematic.

Socioeconomic Inequality: Implicit Assumptions

As discussed above, the two concepts of *equality* and *adequacy* can be extended by a social dimension. Which inequalities or differences are perceived as problematic or not? To compare the socioeconomic achievement gap, we can compute the mean difference in performance between disadvantaged and privileged children. By definition this measure quantifies differences between children from different socioeconomic backgrounds, although it ignores any other differences. If the performance gap between socioeconomic groups is small, there may still be other gaps, such as between gender, race and so on. In the same vein, we can compare the proportion of socioeconomically disadvantaged children who do not reach a certain basic literacy and numeracy performance level with the proportion of privileged children who do not reach these levels. However, again, even if the differences between social groups are small, that does not assume that all children are literate and numerate. So it is socioeconomic gaps that are of interest, or something else? Brighouse et al. (2019, p 57) questions the gaps in socioeconomic status are the main problem, "what

is really at stake may be the low achievement of members of the low-performing group rather than the size of the gap between the average achievement of the two groups. Here the relevant distributive value may be adequacy (…)".

Achievement gaps between socioeconomic groups receive a tremendous amount of attention in the literature based upon international assessments. It is beyond the scope of this chapter to further discuss whether more concern should be given to other forms of inequality such as *equality* and *adequacy*. It is, however, important to acknowledge the implicit assumptions behind different measures. The crucial issue is the need to provide arguments concerning which information is considered to be relevant: *equality*, *adequacy*, or *socioeconomic inequality*.

Measurement Issues

Indicators of Socioeconomic Status

Mueller and Parcel (1981, p. 14) refer to the broader concepts of social stratification to define socioeconomic status as an individual's position within a society:

> The term "social stratification," for example, is used to describe a social system (usually a society or community) in which individuals, families, or groups are ranked on certain hierarchies or dimensions according to their access to or control over valued commodities such as wealth, power, and status. A case's relative position (and associated score) on a particular hierarchy (or combination of hierarchies) may be referred to as its SES. (Mueller & Parcel, 1981, p. 14)

In studies of child development, the three common commodities used to measure SES are parental income, education, parental occupation, and parental education (Duncan, Featherman, & Duncan, 1972; Gottfried, 1985; Hauser, 1994; Mueller & Parcel, 1981; White, 1982; Sirin, 2005). It is difficult to survey information on these indicators for several reasons. Students as well as their parents are often unable or unwilling to report income reliably (Moore, Stinson, & Welniak, 2000). While these problems do not only apply to the recording of income, further problems arise when measuring occupations and educational qualifications in international studies. It is difficult to establish valid and reliable classification systems to put degrees and occupations into a hierarchy. International classification systems such as the ISCED (International Standard Classification of Education; UNESCO, 1997) and the ISEI (International Socio-Economic Index of Occupational Status; Ganzeboom, De Graaf, & Treiman, 1992) have been developed to address this issue. However, comparable coding is only possible to a certain extent in intercultural surveys due to national differences in the educational and economic systems (Jerrim, Volante, Klinger, & Schnepf, 2019). Furthermore, the coding of occupations is labor-intensive and therefore costly when the information is collected by means of an open question.

In addition to income, occupation, and education, international large-scale assessments typically administer various questions on home possessions such a car, lawnmower, or the number of books. Home possessions are common indicators of SES because questions about the presence of

a car, paintings, lawnmower, or the number of books, are easier to answer than questions about parental income, professions and education. Student data can be used to survey home possessions, which is an advantage in international studies, since parents do not often fill out the questionnaires and thus, the proportion of missing data is very high. However, regional and cultural differences remain an issue in international surveys. Owning a lawnmower is in many countries an indicator for having a garden, but very dry areas often do not have grass, therefore this indicator does not work in all countries. The number of cars is also less meaningful in urban areas than in rural ones. Rutkowski and Rutkowski (2013) provide evidence that the latent structure of items on home possessions varies across countries, so one should be careful when using the same items internationally to measure SES.

The number of books in the family home is probably the most popular home possession indicator used to measure SES. It has been used for more than 100 years in educational research and is a part of the survey material for the majority of international assessments. The indicator is popular because books are theoretically closely linked to education, often correlated with high student achievement in all countries (Brese & Mirazchiyski, 2013; Hanushek & Woessmann, 2011). Engzell (2019), however, argues against the use of the book variable as student and parent data do not always match. He observed that girls often rate the number of books higher than boys and that disadvantaged children tend to underestimate the number of books. Whilst this criticism is important, it is appropriate to point out that all indicators are imperfect; gender difference, for example, is not only observed for the number of books, but also for the student data on parental education.

The meaning of different indicators of SES changes over time. Economic structural change often means the importance and prestige of certain occupations changes with time, while new professions emerge, or gain prominence. Similarly, the importance of educational qualifications is also changing in the context of great educational expansion that can be observed worldwide in the past 100 years. Even within a few years, the significance of certain indicators of social status can change dramatically. For example, TIMSS data reveals that, the share of eighth grade students who report more than 100 books at home decreased from 65 to 42 percent between 1995 and 2011 in Sweden (Beaton et al., 1996; Mullis, Martin, Foy, & Arora, 2012). A possible explanation for such large differences is the spread of eBooks in recent years. Similar changes can be observed for computers and other digital devices.

Watermann, Maaz, Bayer and Roczen (2016) discuss whether SES is a multi- or unidimensional construct. If SES is considered a multidimensional construct, occupation is an indicator of social prestige, education is indicated by cultural resources, and income by financial liberties. On the other hand, if SES is considered to be unidimensional, all indicators are measures of the same latent constructs. International assessments like TIMSS and PISA typically compute and report SES indices which combine information from different components, such as the so-called PISA index of economic, social, and cultural status (ESCS), and the TIMSS index of home resources for learning (HER). Research papers based on international large-scale assessment (ILSA) data, however, use both single indicators as well as complex indices. A recent review of 35 international

studies on SES inequality (Strietholt et al., 2019), reported that around half of the studies used single indicators, and the other half, complex indices. The most common single indicators were the number of books at home and parental education.

Classification of Continuous and Categorical Measures

There are many measures to quantify socioeconomic inequality in achievement which all combine socioeconomic background information with student achievement. Both socioeconomic status and achievement may be measured as categorical or continuous variables. An example of a categorical indicator of the socioeconomic status is the comparison of students that have parents with or without tertiary education, and an example for a continuous indicator is the household income. The achievement scores in studies like PISA and TIMSS are examples of continuous achievement measures. The achievement scale is also divided into so-called "proficiency levels" (PISA) or" international benchmarks" (TIMSS), level 2 in PISA and the low benchmark in TIMSS are sometimes regarded as a baseline level of literacy. Following this approach, a common categorical achievement measure is whether students perform below a certain achievement threshold.

According to Table 1, different measures of socioeconomic performance inequality can be classified into four different types, depending on whether performance and socioeconomic status are measured categorically or continuously: (1) if both status and performance are measured as categorical variables, a simple contingency table can be used to describe inequality and based on this information, measures such as the relative risk or odds ratios can be calculated. For example, if half of the disadvantaged children and a quarter of the privileged children do not reach a certain achievement level, the relative risk of disadvantaged children is two times higher than that of privileged children; (2) if status is measured categorically and performance continuously, the achievement gap may be computed as the simple difference in the average achievement of privileged children, and the average achievement of disadvantaged children; (3) if status is measured continuously and performance categorically, logistic regression can be used to regress the binary performance indicator on a continuous measure of the social status; (4) if both status and performance are measured continuously, the covariance between the two variables, Pearson's correlation or linear regression, can be used to assess the continuous performance level on a continuous measure of the status.

Table 1. Classification of measures of socioeconomic inequality in achievement

		Achievement	
		Categorical	Continuous
Socioeconomic status	Categorical	1	2
	Continuous	3	4

Standardization and Threshold-setting: International Comparability and National Specificity

There is a common distinction between absolute and relative measures of inequality (see Heisig, Elbers, & Solga, 2019). Absolute measures are unstandardized measures of inequality and relative measures are standardized. Unstandardized measures use the metric of the achievement scale to quantify inequality. In studies like TIMSS and PISA, the achievement scale has an international mean of 500 with a standard deviation of 100 so that a SES achievement gap of 50 point corresponds to half an international standard deviation. To be more precise, the metric was set in the years of the first administration of the study and based on the countries that participated in that year; the same metric was used in subsequent years to facilitate trend analyses over time. However, the variation in test results typically varies by country, and standardized measures take these differences into account. If there is no variation in test scores in a county, there cannot be any achievement gaps. On the other hand, there could be a huge variation in test scores within one country. For example, in TIMSS 2015 (Grade 4) the standard deviation of the mathematics test scores was 57 points in the Netherlands and 107 points in Jordan. An achievement gap of 50 points corresponds to a relative achievement gap of about one standard deviation in the Netherlands, but only around half a standard deviation in Jordan. In addition to standardizing the performance variable, the grouping variable SES can also be standardized by country. Indicators for SES such as the number of books at home, parental education, or income, are unequally distributed internationally, and such differences may be taken into account by standardizing the SES indicator.

It is sometimes useful to divide a continuous scale using thresholds to ease the interpretation. For example, the concept of academic resilience focuses on students who succeed against the odds; resilience is defined by low status and high performance. To define low status and high performance either fixed or relative thresholds can be used; relative thresholds vary by country (see Ye, Strietholt, & Blömeke, 2020). An example of a fixed threshold are the so-called benchmark levels in TIMSS. In the TIMSS report, all students who score at least 625 points in the TIMSS test are considered to be of an advanced level (e.g. Mullis, Cotter, Centurino, Fishbein, & Liu, 2016). By applying fixed threshold for all countries, half of the students in countries such as Singapore and Hong Kong are classified at an advanced mathematical level. On the other hand, in several other countries, none or only a few percent of the student population reached this level. To address this, an alternative approach is to classify high performing students in each country separately by using relative thresholds that vary by country. For example, we can use the 75th percentile in each country to identify the 25 percent top performing students in each country. In the same vein fixed or relative thresholds can be used to define disadvantage. A drawback of relative thresholds is the substantive comparability of the groups across countries; in some countries even high performing students have only an understanding of whole numbers, while in other countries, high performance means that students are able to solve linear equations and they also have a solid understanding of geometry.

Empirical Analyses

Data and Variables

We will next use TIMSS 2015 data to study different SES measures. In Grade 4, student achievement tests in mathematics and science and student, parent, teacher, and principal questionnaires were administrated. The student and parent questionnaires cover various items on SES that will be used to compute measures of SES inequality in mathematics. We use data including 245,060 students in 46 countries, in each country around 5,000 students from 150-200 schools were sampled. Data from England and the US was not used because no parent questionnaire was given, and we also excluded the regions Dubai, Abu Dhabi, Ontario, Quebec and Buenos Aires. Martin, Mullis, and Hooper (2016) provide further information on the study design and technical details.

We used seven SES measures to capture a wide variety of indicator measures of parental education and occupation, as well as home possession. An income measure was not used because TIMSS does not include such an item. We consider student and parent data, categorical and continuous information, single items and a composite measure:

(1) Having access to internet (dichotomous variable; student data);

(2) having an own room (dichotomous variable; student survey);

(3) number of books in the home (five ordered categories; student survey);

(4) number of books in the home (five ordered categories; parent survey);

(5) parental occupation (four ordered categories; parent survey);

(6) parental education (five ordered categories; parent survey).

The seventh SES variable is the composite measure that combines five of the previously mentioned indicators (1), (2), (3), (5), (6) and number of children's books in the home (five ordered categories; parent survey). Martin, Mullis, & Hooper (2016, p. 15.33) provide detailed information on how item response theory was used to compute the continuous scale:

(7) Home resources for learning (HRL; continuous HRL scale; parent and student surveys).

The pooled international data from all countries contains 19-27 percent missing data for parent data and 2 to 3 percent for student data. For 20 percent of the students no information on the HRL score is available. The variation in missing items between student and parent data points to a practical issue for the measurement of SES in ILSA. Student data is typically surveyed in the classroom, while parents fill in the questionnaires at home. For this reason, the amount of missing data tends to be much higher for parent data. In some countries and studies, the response rate in the parent surveys is well above 50 percent, which reduces the sample size and may also introduce bias if the parent data is not missing at random.

Correlations Between the Different Measures of SES

How much do the SES variables correlate with each other? Note that we initially only look at the SES indicators themselves, the SES performance gaps will be considered later. Table 2 shows the correlation between the SES measures at the student and country level for Grade 4 TIMSS data.

For the sake of simplicity, we dichotomized the books variables (up to/more than 100 books), parental occupation (white/blue collar), and parental education (with/without tertiary education) in the country-level analyses. We then used these variables to compute the share of students who have access to the internet, have their own room, more than 100 books at home and so forth in each country. For the continuous HLR, we simply computed the country mean.

The individual-level correlations are presented below the diagonal in Table 2. All variables correlate positively, but the strength of the correlations varies considerably. The HRL scale is composed of the individual items and it is thus not surprising that the composite measure shows the highest correlations with the individual indicators. Further, the number of books reported at home by parents, parental occupation, and parental education are more highly correlated with each other than the other measures. In the student measures, access to internet, having their own room, and the number of books at home are more loosely correlated. It is also worth mentioning that the number of books reported by students is relatively highly correlated with the HLE scale.

The country-level correlations are presented above the diagonal in Table 2 and they reveal interesting patterns. First, a more general finding is that the correlations are higher, on average, at the country level. This difference can be explained at least in part by the fact that measurement errors are less significant for aggregated data than at the individual level. Interestingly, the highest correlation with HLR can be observed for the access to the internet indicator. Second, the share of students who have access to the internet is the best proxy for the composite measure HRL on a country level. It should be noted, however, that there are some counties where hardly any students have access to the internet and others where almost all students have. A SES indicator which is useful on a country level is not necessarily equally useful on an individual level.

The decision of which SES indicators should be used in research depends on various reasons. Under the assumption that SES is a latent unidimensional construct, it is useful to combine the information from different items to increase the validity and reliability of the measures. From this perspective, the composite HLR is a particularly useful measure. However, our analyses also indicate that even single items such as the number of books at home (both student and parent reports), parental education, and parental occupation, may be sufficiently highly correlated proxies of SES. In contrast to the book variable, internet access and owning a room are only weakly correlated with the composite measures and, therefore, insufficient proxies of SES.

Table 2. Student- and country-level correlations between different SES measures.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) Possession: Internet (student)	-	.49*	.59*	.66*	.69*	.65*	.87*
(2) Possession: Own room (student)	.25*	-	.34*	.67*	.29*	.30*	.62*
(3) Books at home (student)	.32*	.18*	-	.74*	.45*	.46*	.78*
(4) Books at home (parent)	.34*	.17*	.59*	-	.46*	.42*	.81*
(5) Parental occupation (parent)	.36*	.15*	.36*	.45*	-	.87*	.75*
(6) Parental education (parent)	.39*	.16*	.41*	.51*	.67*	-	.72*
(7) HLR (student & parent)	.42*	.30*	.72*	.67*	.71*	.75*	-

Note. Pooled international data from 46 countries; data sources are listed in parentheses (student or/and parent survey); on student level (below diagonal) polychoric correlation were computed for the correlations between categorical variables (1) to (6); the square root of the R^2 retrieved from one-way ANOVAs were used to measure the correlations between the continuous HRL (7) scale and the other SES indictors; on country level (above diagonal) Pearson's correlations were computed; * = statistically significant at 5%-level.

Correlations Between the Different Measures of SES Inequality in Achievement

Does the degree of inequality in a country depend on which indicators are used to measure SES, or are countries generally more or less unequal regardless of the indicator used to measure SES?

To address this question, we computed different measures of SES inequality in achievement and estimated the correlations between them. Specifically, we first conducted a series of regression analyses where we regressed mathematics achievement on one of the dichotomous, ordered categorical or continuous SES indicators to compute the amount of variance (R^2) each SES indicators explains in mathematics achievement. We replicated the analyses for each country seven times to achieve seven SES inequality measures for each country. The possession items, access to the internet and owning a room explain on average only about 3 percent of the variance in achievement, the two books variables, parental education and occupation explain around 10 percent, and HLE variable explains approximately 15 percent. However, the inequality measures vary across countries and, in a second step, we compute the correlations between these measures that are depicted in Table 3.

Table 3. Correlations between different measures of SES inequality in mathematics achievement.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) Possession: Internet (student)	-	-.07	.29*	.39*	.47*	.56*	.49*
(2) Possession: Own room (student)		-	.20	.10	-.08	.07	.21
(3) Books at home (student)			-	.81*	.77*	.74*	.88*
(4) Books at home (parent)				-	.86*	.81*	.84*
(5) Parental occupation (parent)					-	.87*	.85*
(6) Parental education (parent)						-	.90*
(7) HLR (student & parent)							-

Note. Pooled international data from 46 countries; data sources are listed in parentheses (student or/and parent survey); Pearson's correlation; * = statistically significant at 5%-level.

We observe reasonably strong correlations between the measures of SES inequality in mathematics achievement which are based on parent data. The associations between the inequality measures are higher than the associations between the SES indicators themselves presented above. For example, the correlations between all inequality measures based on the parent survey correlate, r=.8 to .9. Such high correlations suggest that different measures of SES inequity in mathematics achievement lead to a similar ranking of countries. However, as Ascombe's (1973) quartet shows, numerical calculations for correlations can be misleading and distributions can look different when graphed. For example, outliers or clusters of data points can artificially lead to high correlation. Correlations Figure 1 plots the achievement gap by parental education (x-axis) and the gap by the number of books reported by the parent. The figure reveals that the high correlation is at least in part driven by the extreme values in Hungary, Slovakia, and Turkey. The correlations decrease in the middle of the distribution; in Denmark, Italy, Cyprus, and Korea, for example, the achievement gaps between children with up to more than 100 books is at the same level, while the gap between children of parents with and without higher education varies dramatically across these countries. From the perspective of a single country like Denmark or Korea, how SES has been operationalized makes a considerable difference.

With regard to the questions in the student survey, comparably high correlations with the SES measures from the parent survey can only be observed for the student reported number of books variable. The two inequality measures that are based on the SES indicators, access to the internet and owning a room correlate much lower than with the other items.

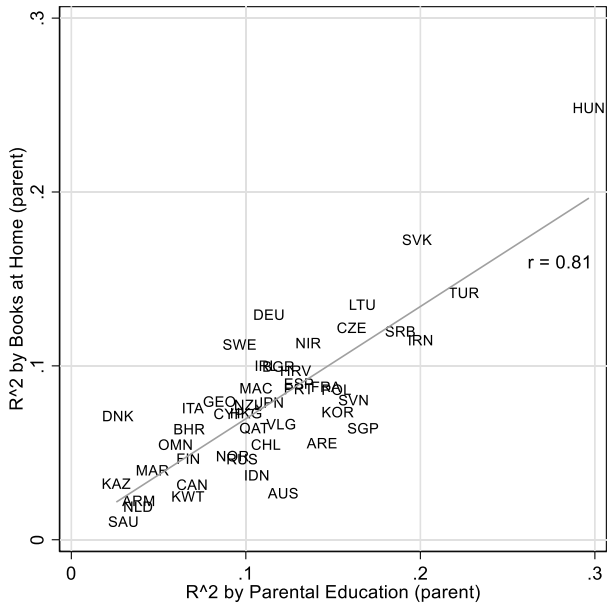


Figure 1. Plot of two measures of SES inequality in mathematics achievement.

Standardization of Inequality Measures: Relative and Absolute Measures

In the pooled international data of TIMSS and other international assessments, the standard deviation of the achievement scale is 100 points, but it varies between countries. For example, the standard deviation was 104 points in Kuwait and 64 in Korea. Figure 2 used two different metrics to

measure the achievement gap by parental education. The x-axis plots the absolute achievement gap defined as the mean differences between children of parents with, versus without, a university degree using the international metric of the achievement scores, and the y-axis shows the relative gap which is standardized by dividing the absolute gaps by the standard deviation in the respective countries. The plot shows that both measures are correlated but the associations are not perfect. For example, the achievement gap in both Kuwait and Korea corresponds to roughly 50 points when using the international achievement scale as a metric for the achievement gaps, while the standardized achievement gaps suggest that the SES inequality is much larger in Korea compared to Kuwait. It should be noted that the previously used R^2 measure is another approach to standardize measures of SES inequality, because here the proportion of explained variance is reported, which can take values between zero and one, independent of how much variance exists within countries. Further, it can be useful to standardize not only achievement, but also the SES to ease the interpretation of associational measures. For example, it is typically easier to interpret the correlation between two continuous variables than their covariance. But in many cases, however, it is difficult to interpret transformed variables. For example, the comparison of blue versus white collar workers is an easy to communicate measure of occupation. At the same time, such a comparison has different meanings in developing and advanced economy.

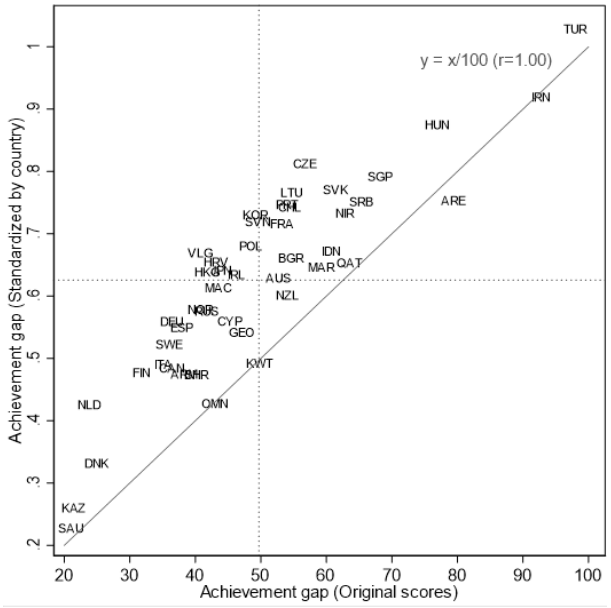


Figure 2. Absolute and relative achievement gaps by parental education.

The variation in achievement is a natural limitation for the SES achievement gaps. If there is no variation within countries, there cannot be an SES achievement gap. To illustrate this, Figure 3 plots the SES gaps with the standard deviation of the achievement scores. In both panels the y-axis shows the standard deviation of the achievement scores within the countries. The left panel shows absolute SES gaps based on the original TIMSS scale, while relative gaps are presented in the right panel. Note that the absolute and relative gaps are the same as that used in Figure 2, except for that

they are now both plotted on the x-axis. The comparison shows that absolute SES performance gaps are larger in those countries where the standard deviation of performance is also large (r=.56). This association vanishes in the right panel with the standardized relative SES gaps (r=.13). The comparison reveals that absolute SES gaps are not only affected by the difference between SES groups, but also by the overall variation in the achievement scores. The relative measure may be conceived as a purer measure of SES inequality which is not affected by the overall variation in achievement.

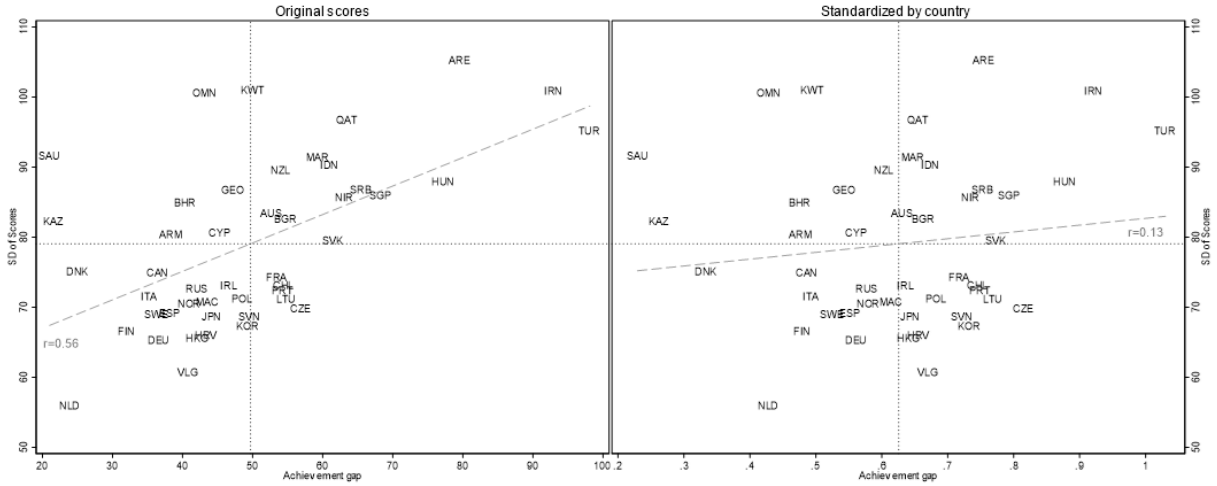


Figure 3. Association between the absolute and relative gaps with the overall variation in achievement

Different measures of SES inequality are best understood and interpreted in context. Whether one should use absolute or relative inequality measures cannot be answered unanimously. Advantages of using the original achievement scales from TIMSS, and other international assessments is that they are well documented and established in the research community. The study reports provide detailed information on the mathematics content that students master at certain levels of the achievement scale. Further, it is well documented that the learning progress of an additional school year towards the end of primary school corresponds to roughly 60 points in TIMSS (Luyten, 2006; Strietholt, Rosén, & Bos, 2013). In the same vein, studies like PIRLS, PISA and TIMSS have now been conducted for 20 or even more years and researchers have developed a reasonably good understanding of how the performance within countries changes from one cycle to another. This kind of interpretability of performance scores is limited as soon as the values are standardized.

Concluding Remarks

One of the most salient findings in international comparative large-scale studies on student achievement are the large SES gaps. These findings have been replicated in several studies (Volante, Schnepf, Jerrim, & Klinger, 2019), and an increasing number of studies investigate the institutional determinates that moderate the association between SES and achievement (Strietholt et al., 2019). Studying SES inequality from a comparative perspective using ILSA data has at least two

methodological advantages. First, many institutional features of educational systems do not vary within a single country (e.g., the existence of central exit exams) so that international comparative studies are the only approach to observe variation in such features. Second, selection mechanisms within educational systems make it difficult to study socioeconomic inequality within a single country. For example, in a tracked school system socially advantaged students are often overrepresented in higher tracks, while disadvantaged children are overrepresented in lower tracks. Analyses at the country level avoid such selection bias and provide a more complete picture of the degree of SES inequality within a country. At the same time, there are several conceptual and methodological challenges for researchers when describing and investigating SES inequality using data from international school achievement studies.

First, different indicators are being used to measure SES. On the one hand, inequality measures based on parental education, parental occupation, home possessions, and composite measures that combine different indicators are fairly high correlated (r=.8 and higher); the ranking of the individual countries is frequently quite different for different indicators. In particular, national policy-makers who are largely interested in mapping their own countries in comparison to others are well advised to consider which SES indicator(s) they consider relevant.

Second, it is impossible to make general recommendations as to which indicators should be used to measure inequality. International classification schemes such as the ISEI and ISCED have been developed to compare occupations and educational degrees internationally, but the cross-cultural validity of these measures is not perfect. Home possessions are a much-needed proxy for income but it is extremely difficult to identify possessions that function similarly in poor and rich countries, as well as in urban and rural areas. For example, there is little variation in the access to the internet within rich and poor countries, in that either everyone or no one has access to the internet. The challenge of finding suitable indicators for ILSAs is also reflected in the constantly changing home possessions scales. In studies like PIRLS, PISA, and TIMSS there are hardly any items that have been administrated continuously across multiple study cycles. An exception is the well-established variable on the number of books at home, which is administrated in all ILSAs. Although this variable is not perfect either, it has important advantages. There is a high face validity since books are important for education and the variable has a variation across a large range of values (there are typically about five categories, e.g., 0-10 books, 11-25 books, …). In contrast, the number of cars in the household, for example, can take a positive value in theory, but in practice, there are hardly any families that have more than two cars.

Third, SES is a moving target. Parental education, their profession and family income have been, and will probably remain, important characteristics for educational careers of children in the future. However, social and economic systems change over time and this change has consequences for the study and analysis of SES. Graduation rates in higher education have risen considerably over the last 100 years in many countries; new professions have emerged while others have lost importance. Home possessions that were good indicators of wealth some years ago, are now

accessible to a wide range of people, and digitalization is replacing printed books with eBooks. SES research must achieve a balance between continuity and change, ensuring comparability over time and making necessary adjustments. We do not want to be misunderstood here; far too often items and instruments are changed only to return to the original version a few years later. Changes must not be an end in themselves; they are only legitimate if there are substantial improvements.

Fourth, this chapter focused on socioeconomic status, but it is clear there are other social categories such as gender and race. In order to investigate educational justice, it is probably insufficient to look only at SES. In the section *Socioeconomic inequality: implicit assumptions* we have raised the question as to whether we should be concerned about SES gaps, or rather children who have been left behind. Of course, the socially disadvantaged are more often left behind, but what really matters here is not the difference between social groups, but the fact that some students are left behind.

References

Anscombe, F. J. (1973). Graphs in Statistical Analysis. *The American Statistician, 27*(1), 17-21.

Atkinson, A. B., & Bourguignon, F. (Eds.). (2000). *Handbook of Income Distribution* (Vol. I). Amsterdam: North-Holland.

Atkinson, A. B., & Bourguignon, F. (Eds.). (2015). *Handbook of Income Distribution* (Vol. II). Amsterdam: North-Holland.

Brighouse, H., Ladd, H. F., Loeb, S., & Swift, A. (2018). *Educational Goods. Values, Evidence, and Decision-Making*. Chicago: The Universtiy of Chicago Press.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzales, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study*: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Brese, F., & Mirazchiyski, P. (2013). *Measuring Students' Family Background in Large-scale International Education Studies.* Hamburg: IEA-ETS Research Institute.

Duncan, O. D., Featherman, D. L., & Duncan, B. (1972). *Socio-economic background and achievement*. New York: Seminar Press.

Ganzeboom, H. B. G., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21*(1), 1-56.

Gottfried, A. (1985). Measures of socioeconomic status in child development research: Data and recommendations. *Merrill-Palmer Quarterly, 31*(1), 85-92.

Hauser, R. M. (1994). Measuring socioeconomic status in studies of child development. *Child Development, 65*(6), 1541-1545.

Hanushek, E. A., & Wößmann, L. (2011). The economics of international differences in educational achievement. In E. A. Hanushek, S. Machin, & L. Wößmann (Eds.), *Handbook of the economics of education.* (Vol. 3). Amsterdam: Elsevier.

Heisig, J. P., Elbers, B., & Solga, H. (2019). Cross-national differences in social background effects on educational attainment and achievement: absolute vs. relative inequalities and the role of education systems. *Compare: A Journal of Comparative and International Education, 50*(2), 165-184. doi:10.1080/03057925.2019.1677455

Jerrim, J., Volante, L., Klinger, D. A., & Schnepf, S. V. (2019). Socioeconomic Inequality and Student Outcomes Across Education Systems. In *Socioeconomic Inequality and Student Outcomes* (pp. 3-16).

Luyten, H. (2006). An empirical assessment of the absolute effect of schooling: regression-discontinuity applied to TIMSS-95. *Oxford Review of Education, 32*(3), 397-429.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). *Methods and Procedures in TIMSS 2015*. Boston College, TIMSS & PIRLS International Study Center.

Moore, J. C., Stinson, L. L., & Welniak, E. J., Jr. (2000). Income Measurement Error in Surveys: A Review. *Journal of Oficial Statistics, 16*(4), 331-361.

Mueller, C. W., & Parcel, T. L. (1981). Measures of Socioeconomic Status: Alternatives and Recommendations. *Child Development, 52*(1). doi:10.2307/1129211

Mullis, I. V. S., Cotter, K. E., Centurino, V. A. S., Fishbein, B. G., & Liu, J. (2016). Using Scale Anchoring to Interpret the TIMSS 2015 Achievement Scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 14.1-14.47). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timss.bc.edu/publications/timss/2015-methods/chapter-14.html

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 International Results in Mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Nozick, R. (1974). *Anarchy, state and utopia*. New York: Basic Books.

Rutkowski, D., & Rutkowski, L. (2013). Measuring Socioeconomic Background in PISA: One Size Might not Fit all. *Research in Comparative and International Education, 8*(3), 259-278. doi:10.2304/rcie.2013.8.3.259

Robeyns, I. (2006). Three models of education: rights, capabilities and human capital. *Theory and Research in Education, 4*(1), 69–84. doi: DOI: 10.1177/1477878506060683.

Sirin, S. R. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research, 75*(3), 417-453. doi:10.3102/00346543075003417

Strietholt, R. (2014). Studying educational inequality: reintroducing normativenotions. In R. Strietholt, W. Bos, J.-E. Gustafsson & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments* (pp. 51–58). Münster/New York: Waxmann.

Strietholt, R., Gustafsson, J. E., Hogrebe, N., Rolfe, V., Rosén, M., Steinmann, I., & Yang-Hansen, K. (2019). The Impact of Education Policies on Socioeconomic Inequality in Student Achievement: A Review of Comparative Studies. In L. Volante, S. V. Schnepf, J. Jerrim, & K. D. A (Eds.), *The Impact of Education Policies on Socioeconomic Inequality in Student Achievement: A Review of Comparative Studies*. Singapore: Springer.

Strietholt, R., Rosén, M., & Bos, W. (2013). A correction model for differences in the sample compositions: the degree of comparability as a function of age and schooling. *Large-scale Assessments in Education, 1*(1), 1–20. doi:10.1186/2196-0739-1-1

UNESCO. (1997). *ISCED 1997: International Standard Classification of Education*: UNESCO – Institute for Statistics.

Van Doorslaer, E., & Van Ourti, T. (2011). Measuring inequality and Inequity in Health and Health Care In S. Glied & P. C. Smith (Eds.), *The Oxford Handbook of Health Economics* (pp. 837-869). Oxford: Oxford University Press.

Volante, L., Schnepf, S. V., Jerrim, J., & Klinger, D. A. (2019). *Socioeconomic Inequality and Student Outcomes. Cross-National Trends, Policies, and Practices*: Springer.

Watermann, R., Maaz, K., Bayer, S., & Roczen, N. (2016). Social Background. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing Contexts of Learning* (pp. 117-145): Springer.

Article 3. Early Tracking and Different Types of Inequalities in Achievement: Difference-in-Differences Evidence from 20 Years of Large-scale Assessments

Andrés Strello, Rolf Strietholt, Isa Steinmann, Charlotte Siepmann

The version of the manuscript printed below is the preprint of the article published in *Educational Assessment, Evaluation and Accountability*, volume 33 (2021). https://doi.org/10.1007/s11092-020-09346-4

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.

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,[5] while countries like the United States do not track students into ability-grouped schools before higher education.

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

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, Trautwein, Lüdtke, and Baumert 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[6] (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 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

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

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.

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, Ladd, Loeb, and Swift 2018; Hanushek 2013; Hanushek, Schwerdt, Wiederhold, and Wößmann 2015). In the second part, we review previous studies on the effects of tracking.

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).

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 towards 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, Autin, Bataillard, and Butera 2018; Hallinan 1994; Holm, Jæger, Karlson, and Reimer 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, Ursprung, and Wößmann 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.

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 and colleagues (2018) 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.

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.

Methodology

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, Bunker, and Wachs 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 Figure 1 by study, cycle, domain, and country (study-by-cycle-by-domain-by-country observations) amounted to 1177.

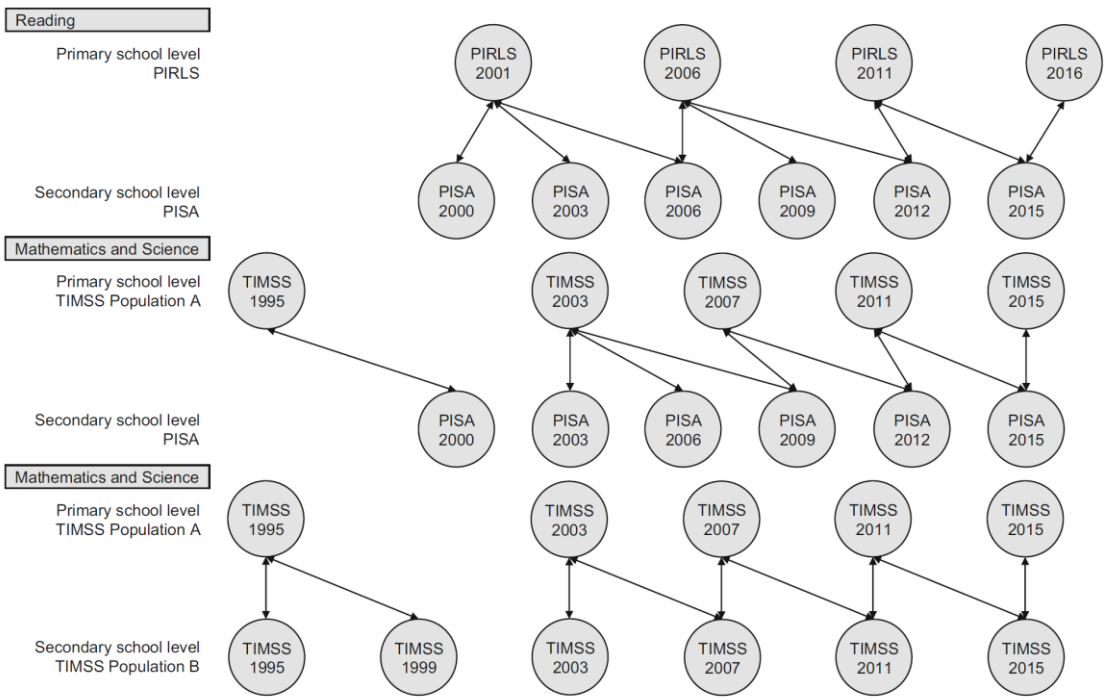


Figure 1. Study Pairs of Large-Scale Assessments at Primary and Secondary School Level.

Notes. 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.

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 standard deviation of 100 (Martin, Mullis, and Hooper 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 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 percent for the books at home variable to 30 percent 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 percent of primary and 12 percent of the secondary school level students did not reach these levels of adequate achievement in the present sample.

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

Note. 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.

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 percent 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, 2013b, 2013a, 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. Annex 1 shows the tracking status for all countries in our sample.

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

						Overall sample	Early tracking	Late tracking
Primary school level data				Secondary school level data		*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

						Overall sample	Early tracking	Late tracking
	Primary school level data			Secondary school level data		*N*	*N*	*N*
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
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

Note. 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.

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 equation (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. Figure 1 and Table 2).

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:

$$w_i = \frac{1}{SE_i^2} \quad (2)$$

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:

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

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

$$SE_{\overline{ES}} = \sqrt{\frac{1}{\sum w_i}} \quad (4)$$

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).

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 Figure 2. Panel A shows, for example, the regression coefficients of the effects of early tracking on dispersion inequality along with the 95 percent 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.

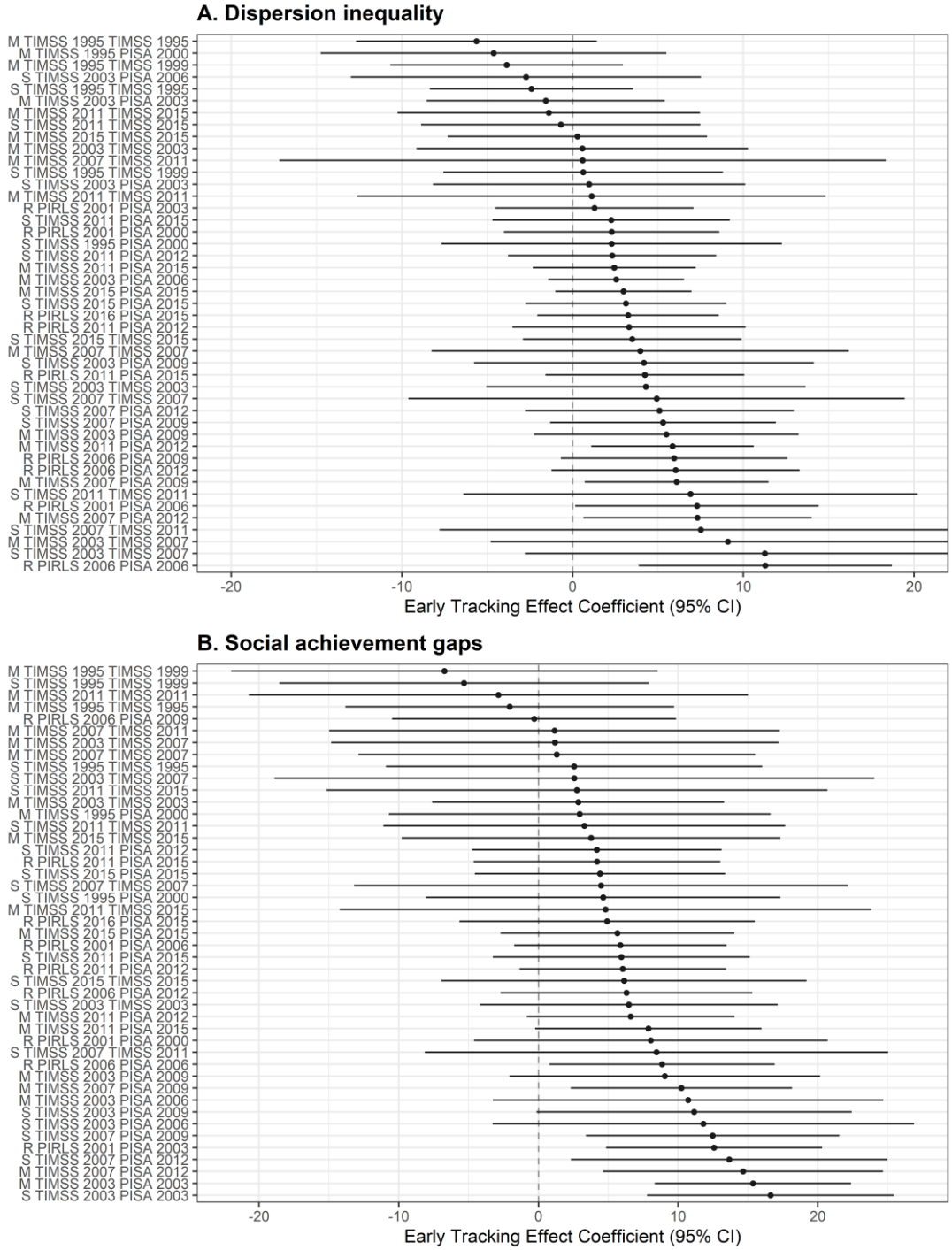


Figure 2. Individual Estimates of the Effects of Early Tracking on the Four Dependent Variables in the 45 Study Pairs

Note. 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.

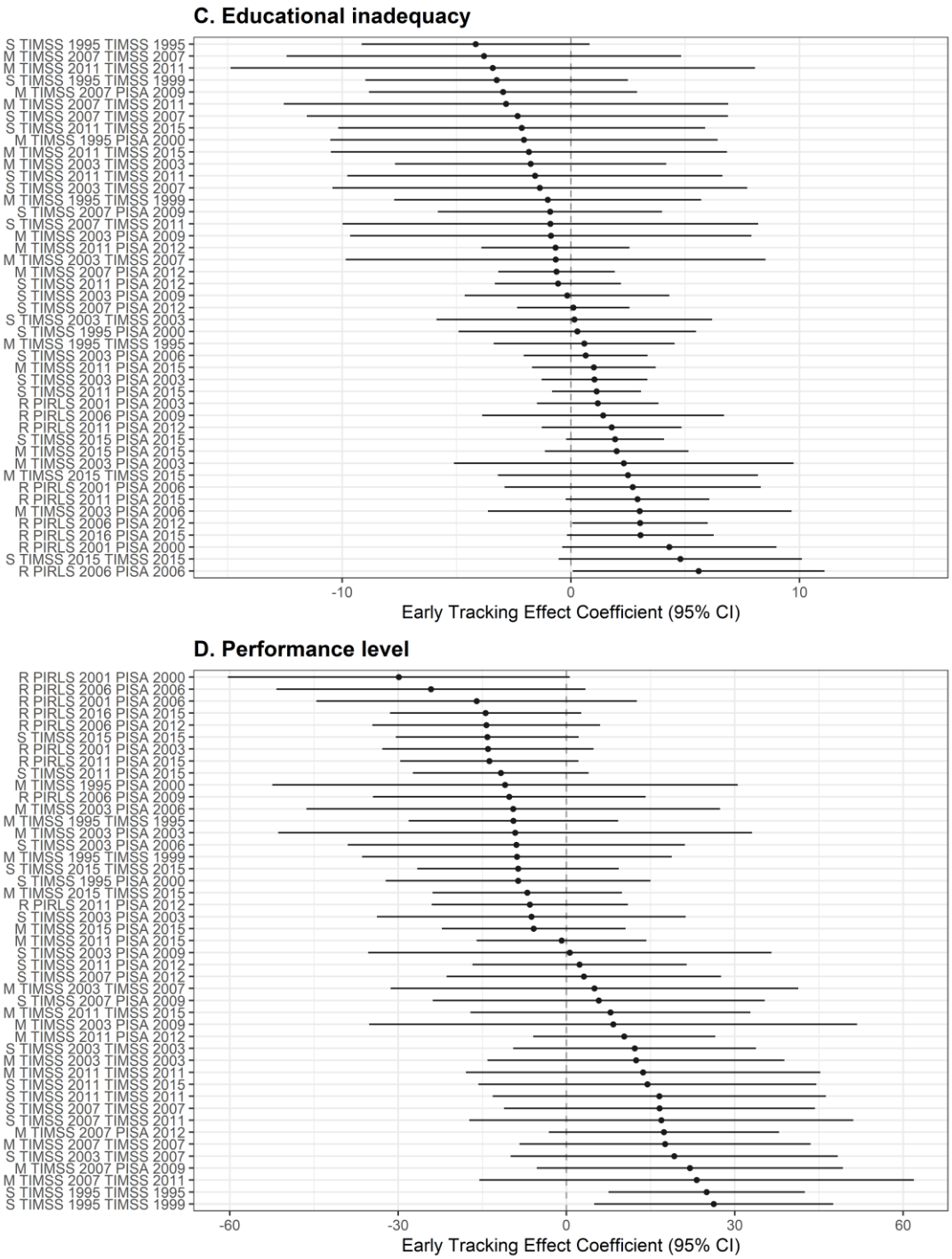


Figure 2. Individual Estimates of the Effects of Early Tracking on the Four Dependent Variables in the 45 Study Pairs (cont.)

Note. 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.

Mean Effects of Early Tracking on Inequalities

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

	(1) All Domains		*(2) Reading*		*(3) Mathematics*		*(4) Science*	
	$\bar{\gamma}$	*SE*	$\bar{\gamma}$	*SE*	$\bar{\gamma}$	*SE*	$\bar{\gamma}$	*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	

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

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 percent. 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 ($p > .05$).

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 Annex 2). The results remained qualitatively the same.

Further Analyses

Table 4. Robustness Checks of the Synthesis of the Effects of Early Tracking ($\bar{\gamma}$) for All Domains in the 45 Study Pairs

	$\bar{\gamma}$		*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 non-dichotomous			
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

Note. The unstandardized parameter $\bar{\gamma}$ 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 *** $p < .001$.

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 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 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 percent 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 percent 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 percent ($p < .01$; 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 five 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.

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.

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.

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.

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.

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, Dumont, and Trautwein 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.

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 ten 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.

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. doi: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. doi: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. doi: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. doi:10.1016/0038-0121(77)90032-5

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*, *92*(4), 1545–1572. doi:10.1093/sf/sou003

Boudon, R. (1974). *Education, Opportunity, and Social Inequality: Changing Prospects in Western Society*. New York: John Wiley & Sons.

Brighouse, H., & Swift, A. (2008). Putting educational equality in its place. *Education Finance and Policy*, *3*(4), 444–466. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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: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. doi: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: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:10.2797/40560

Eurydice. (2013b). *The Structure of the European Education Systems 2013/14: Schematic Diagrams*. doi:10.2797/206797

Eurydice. (2014). *The Structure of the European Education Systems 2014/15: Schematic Diagrams*. doi:10.2797/607957

Gorard, S., & Smith, E. (2004). An international comparison of equity in education systems. *Comparative Education*, *40*(1), 15–28. doi: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. doi: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. doi:10.3368/jhr.47.3.684

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

Hanushek, E. A. (2013). Economic growth in developing countries: The role of human capital. *Economics of Education Review*, *37*, 204–212. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. doi:10.1093/sf/sov006

Koerselman, K. (2013). Incentives from curriculum tracking. *Economics of Education Review*, *32*(1), 140–150. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. 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*(4), 483–505. doi: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. doi: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. doi:10.1787/9789264006416-en

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

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

OECD. (2010). *PISA 2009 Results: What Makes a School Successful?* OECD. doi: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. doi: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. doi:10.1016/j.econedurev.2014.06.002

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi: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. doi: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. doi: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. doi:10.1177/1465116509346387

Schütz, G., Ursprung, H. W., & Wößmann, L. (2008). Education policy and equality of oppurtunity. *KYKLOS*, *61*(2), 279–308. doi: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: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. doi:10.3102/00346543060003471

Solga, H. (2014). Education, economic inequality and the promises of the social investment state. *Socio-Economic Review*, *12*, 269–297. doi: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* [Unpublished manuscript].

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

UNESCO. (2018). *Handbook on Measuring Equity in Education*. doi:10.1016/S0733-8619(03)00096-3

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. doi: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. doi: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. doi: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. doi:10.1002/berj.3172

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

Annex

Annex 1. 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
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		

Country	Tracking age	Tracking grade	Early tracking country in PISA	Early tracking country in TIMSS
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		
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		

Note. 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, 2013b, 2013a, 2014), and OECD reports (2004, 2006, 2008, 2010).

Annex 2. 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

Note: 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.

Article 4. Does Tracking Increase Segregation? International Evidence on the Effects of Between-school Tracking on Social Segregation Across Schools

Andrés Strello, Rolf Strietholt, Isa Steinmann

The version of the manuscript printed below is the preprint of the article published in *Research in Social Stratification and Mobility*, volume 78 (2022). https://doi.org/10.1016/j.rssm.2022.100689

Abstract

In the present study, we focus on institutional segregation as determined by the school system. We investigate the effect of between-school tracking on the countries' social segregation across schools as measured by the variation of the socioeconomic composition of schools. We combine data from international large-scale assessments to estimate difference-in-differences models. Specifically, we regress the degree of social segregation across schools at the secondary school level on a between-school tracking indicator, while controlling for the degree of segregation at the primary school level. The analyses are replicated on 32 different combinations of datasets from three large-scale assessments (Progress in International Reading Literacy Study [PIRLS], Trends in International Mathematics and Science Study [TIMSS], and Programme for International Student Assessment [PISA]), englobing 16 assessment cycles and altogether 75 different countries or regions. The results provide strong evidence for the hypothesis that between-school tracking increases the social segregation of advantaged and disadvantaged students.

1. Introduction

Social segregation between schools means that children from disadvantaged and advantaged families do not attend the same but different schools. The unequal distribution of students across schools is linked to student outcomes because the social context and peers are considered relevant factors of cognitive learning outcomes and for the socialization of the students. For example, there is evidence that a higher proportion of students from advantaged backgrounds in the same school has positive effects on learning for less advantaged students (Sacerdote, 2011). Comparative studies reveal a correlation between the degree of social segregation and the social achievement gap (Burger, 2019; Hindriks, et al. 2010). Segregated school systems are also problematic for the socialisation of children if there is no contact and exchange between children from poorer and richer families. Historically, there have been forms of deliberate school segregation of social groups such as race or gender that persisted for a long time in some countries or that are still promoted today (e.g., Brown vs Board of Education in the US; single-gender school). Unlike segregation by race or gender,

segregation by socioeconomic background has not been publicly and politically propagated, but rather occurs in hidden ways.

Although social segregation is considered an important issue for the socialisation and learning of children, there is little international comparative research that systematically addresses the topic. This lack of research is surprising, since the degree of social segregation across school can only be observed and compared at the country level. There is also hardly any rigor research that aims to determine the institutional factors that explain why the degree of segregation is higher in some countries or school systems, than in others. The present paper is a first attempt to fill this research gap. In particular, we have focused on the question of whether between-school tracking, i.e., the sorting of students into different ability tracks in secondary education, increases the social segregation across schools in education systems.[7]

Between-school tracking is one of the most controversial issues in educational research that addresses the structures of school systems. While proponents of ability tracking argue that it is easier and more effective to teach groups with homogeneous abilities, critics are concerned that tracking reinforces inequalities. When summarizing the empirical findings from previous research, one finds hardly any evidence to support the arguments of the proponents but instead, fairly strong evidence that suggests between-school tracking would increase inequalities in cognitive and non-cognitive learning outcomes, as well as in educational attainment (Hanushek & Wößmann, 2006; van de Werfhorst & Mijs, 2010; Strello et al., 2021; Lavrijsen & Nicaise, 2016; Reichelt et al., 2019; Parker et al., 2016). However, very few studies have investigated how tracking affects the social segregation across schools from a comparative perspective. Although the concepts of inequalities in outcomes and segregation are both discussed in the discourse around social justice, they are distinct phenomena. Theoretically, higher segregation does not automatically translate into a higher outcome inequality because richer and poorer students *may* actually learn more efficiently in more homogeneous groups. Further, studying the determinants of segregation is inherently difficult because the degree of school segregation may be confounded with other institutional features. One example of such an institutional feature is the existing residential segregation, which is likely to be higher in countries with more pronounced social disparities between, for instance, rural and urban areas.

To study the effects of between-school tracking on the degree of school social segregation in education systems, we have applied a difference-in-differences approach, in which we studied the degree of social segregation at the secondary school level, while controlling for the degree of social segregation that is already existent at the primary school level. This approach has allowed us to circumvent the issue of confounding national context factors, such as residential segregation. In our analyses, all cycles of the Organisation for Economic Co-operation and Development's (OECD)

[7] Chmielewski, Dumont, and Trautwein (2013) distinguish between course-by-course tracking, within- and between-school tracking, as different types of differentiation. In the present study, we consider only between-school tracking, as we examine social segregation between schools. Nevertheless, we recognize that even within schools, the other forms of differentiation can segregate certain social groups of students.

Programme for International Student Assessment (PISA), IEA's Trends in International Mathematics and Science Study (TIMSS), and IEA's Progress in International Reading Literacy Study (PIRLS) were used, taking advantage of the repeated sampling of student populations at both primary and secondary school levels from overlapping countries.

This article is divided into five sections. First, a review of the main concepts and previous comparative research on tracking and school segregation was presented. Thereafter, we outlined the analysis plan, i.e., how we aimed to identify the effects of tracking on segregation. In the results section, international comparative findings concerning the degree and distribution of social segregation across schoolswere presented, as well as our main findings on the effects of tracking on segregation. Lastly, we discussed our findings.

2. Literature Review

We follow Allen and Vignoles (2007) and define social segregation across schools as the degree to which members of social groups attend the same or different schools. Under this definition, complete social segregation means that all members of one social group attend other schools than the members of the other social group. Absence of social segregation across schools, by contrast, means that the members of both social groups are equally distributed across all schools of a country. While the social segregation across schools definition can apply to different social groups (e.g., gender, immigration status, religion; Gorard & Smith, 2004), we focus exclusively on the school segregation by socioeconomic status in this article. We acknowledge, however, that social segregation across schools is a complex concept that cannot be fully captured by one single indicator or index (OECD, 2019).

2.1 International Variation in Segregation

International comparative studies provide a unique opportunity to study segregation, since segregation is a phenomenon best observed at the system level. The data from international large-scale assessments contain large representative samples and locally adapted survey instruments to measure the degree of social segregation. Since the release of the first PISA assessment, a few studies have used the data to investigate segregation. The analyses reveal a considerable variation in the degree of social segregation across schools in international comparison, and that these cross-country differences have remained stable over time (Jenkins et al., 2008; Gutiérrez et al., 2019). The high variability between countries, together with the low variability over time, suggest that stable institutional features at the country level are an important determinant of social segregation.

Nevertheless, due to the cross-sectional design of PISA and other international assessments, it is inherently difficult to identify the determinants of social segregation because there may be other confounding variables, such as an already existing residential segregation. If there are large differences between rich and poor regions within countries, the social segregation across schools in the respective regions will differ significantly. Indeed, some national studies confirm that the degree of residential segregation explains some of the variation in the degree of school segregation in

Germany, Spain, Sweden, and the United States (Kristen, 2005; Bonal et al., 2019; Malmberg & Andersson, 2020; Bottia, 2019). However, the correlation between residential and school segregation is not perfect, which implies that there are other institutional features within education systems that promote further segregation processes. Central to social segregation are the transitions that take place within the education system, where the choice of the educational path differs between social groups.

2.2 Transitions and School Choice

The term 'institutional differentiation' covers various forms of formal and informal, as well as internal and external, hierarchies in the education system. While there are different forms of differentiation, the most obvious one is the so-called 'between-school ability tracking'. Between-school tracking means that students are sorted into different types of secondary schools that have either an academic or vocational orientation (Skopek et al., 2019; Dollmann, 2019). Some authors (e.g., Hanushek & Woessmann, 2006) call between-school tracking 'ability tracking', emphasizing that the sorting of students into different tracks should not be based on their social status but on their abilities. There are reasons, however, why between-school tracking might nevertheless increase social segregation. In this regard, Boudon (1974) introduced the useful distinction between primary and secondary effects to explain the mechanisms by which social background influences educational decisions at the transition between different educational stages. In the present case, primary effects refer to the association between student background and achievement before transitioning from primary to secondary school. An achievement-based allocation to different ability tracks automatically leads to social segregation, if background and achievement correlate. Secondary effects refer to background-dependent educational decisions. The basic idea is that, even when comparing students with the same academic achievement in primary school, a greater number of privileged students are more likely to transition to more ambitious academic tracks than socially disadvantaged children.

To understand the mechanisms that underlie secondary effects, Kristen (2005) developed a theoretical model that incorporates a sequence of three stages (see also Hallinan, 1994). These include the perception of different school alternatives, the evaluation of the perceived alternatives, and the selection of and access to the desired school. Kristen (2005) argues that even after controlling for student achievement, privileged families have both more cultural and social resources, better information to evaluate the different educational alternatives, and higher educational expectations than disadvantaged families. At the same time, other important actors, such as teachers, tend to believe that students with more privileged backgrounds are more likely to succeed in e.g., ambitious tracks. Batruch et al. (2018) conducted a vignette experiment, where the description of social status of students varied, whereas the achievement level was the same. They observed that university students and teachers considered a lower track more appropriate for lower SES students than for higher SES students and a higher track more appropriate for higher SES students than for lower SES students, even when achievement was of the low and high SES was the same. Secondary effects are

often considered more objectionable than primary effects. However, it should be recognized that both primary and secondary effects suggest that tracking increases social segregation across schools. This argument is detailed further, below. The sorting of students into tracked schools constitutes one of the most contentious issues in education. While every education system differentiates at some point (e.g., between lower and upper secondary education), the controversy is mainly about the age at which students are tracked for the first time. Some countries like Germany, Austria, and Hungary, track their students as early as the end of grade 4. Opponents of early tracking argues that sorting at such early ages can be considered problematic, since students' abilities may not be reliably assessed and the family background and expectations would have a stronger impact on the school choice of the student at an early age (Horn, 2009; Erikson & Jonsson, 1996). For this reason, the tracking of children can be seen as a driver of social segregation. Especially in tracking countries where the track allocation highly depends on teachers' and parents' judgements and choices, the track allocation may be more highly related to the socioeconomic status of the child. In early tracking countries with a more objective track allocation (e.g., based on central exams), the association should be less pronounced (Korthals and Dronkers, 2016; Bol et al., 2014).

However, there is hardly any robust evidence on effects of tracking on social segregation across schools. Some studies have found that the degree of social segregation tends to be higher in countries with a tracked lower secondary school system, but that high levels of segregation have also been observed in late tracking countries (Gorard & Smith, 2004; Jenkins et al., 2008; Chmielewski & Savage, 2015; Burger 2019; Chmielewski 2014; OECD 2019; Murillo et al., 2018). A central problem is that all these international comparisons are based only on cross-sectional comparisons, making it difficult to draw causal inferences. Burger (2019) points out that social segregation across schools may be high because students from disadvantaged backgrounds may live in more homogeneous areas, and that the social composition of schools just mirrors the social composition of the neighbourhood. The central problem for the empirical analysis of the interplay between tracking and segregation is accounting for such confounding.

3. Research Question and Hypothesis

The aim of this paper was to determine the effect of between-school tracking on social segregation across schools. To this end, data from two large-scale international studies were used to compare the degree of social segregation in tracked and untracked secondary school systems when students are 15 years old (PISA) and in eighth grade (TIMSS). Previous research on school choice suggests that the sorting of students into different ability tracks differs depending on the social background of students. Following this research, we hypothesized that between-school tracking increases the degree of social segregation across schools.

4. Methodology

4.1 Data Sources: Combining Primary and Secondary School Information

To identify the effect of between-school ability tracking on segregation, we compared the degree of segregation between countries with tracked and comprehensive secondary school systems. We controlled for the degree of segregation at the primary school level in order to circumvent possible bias from confounding factors. For this purpose, we combined primary and secondary school data from three international large-scale assessments: PISA, TIMSS and PIRLS. PIRLS was conducted in 2001, 2006, 2011, and 2016, with a target population of fourth grade students. TIMSS assessed a target population of fourth grade students (Population A) and eighth grade students (Population B) in 1995, 1999 (only Pop. B), 2003, 2007, 2011, and 2015. PISA was administered in 2000, 2003, 2006, 2009, 2012, and 2015 and tested 15-year-old secondary school students. We constructed pseudo-panel datasets at the country level by combining primary and secondary school data to study change in segregation prior and after the transition from primary to secondary school (Cordero et al., 2018).

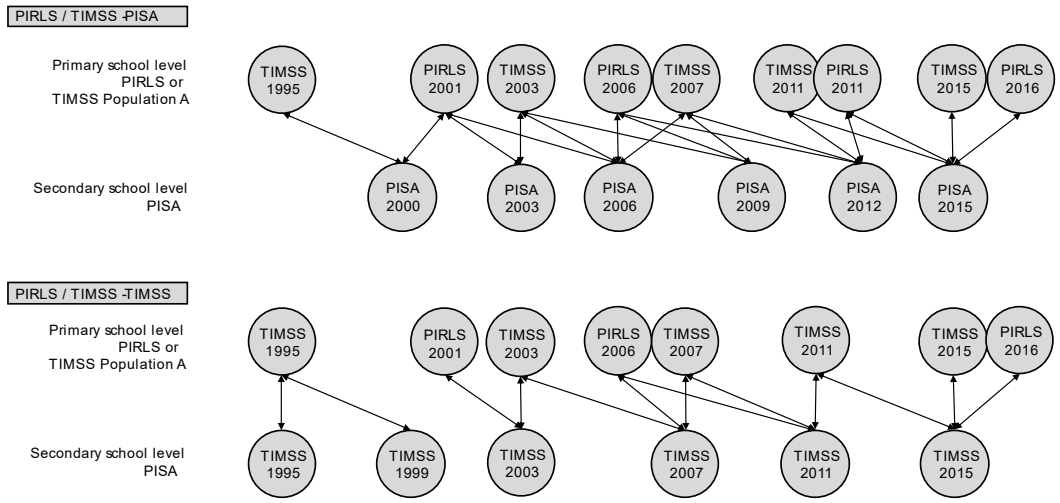


Figure 1. Survey Combinations of Large-Scale Assessments at Primary and Secondary School Level

Following previous research (Strello et al. 2021; Ruhose & Schwerdt 2016), we replicated the analyses with several combinations of primary and secondary school studies in order to maximize the sample size and the number of included countries. We matched primary and secondary school studies that were conducted in roughly the same years (e.g., PIRLS 2001 with PISA 2000) or that surveyed roughly the same student cohorts (e.g., grade 4 students in PIRLS 2001 with 15-year-olds in 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 (see Blanchard et al., 1977). The resulting combinations are depicted in Figure 1. Every arrow reflects a pair of datasets at primary and secondary school level used in our empirical analyses. The overall 32 study pairs contain data from all countries or regions that participated in both assessments (see also

Table 2). In 19 of the study pairs, we matched primary school studies with secondary school data from PISA, and in 13 with secondary school data from TIMSS (Pop. B). Each survey combination includes between 19 and 37 countries (see Table 1), with 75 unique countries or regions being sampled in at least one survey combination. In Table A.4 we report the mean number of students per country, mean number of schools per country, and the mean number of students per school per country for each of the study pairs.

4.4 Variables

4.2.1. Social Segregation Across Schools Indexes

To determine the level of social segregation across schools, we estimated two different indexes. The first is the Dissimilarity Index D (Duncan & Duncan, 1955). D is an unevenness segregation measure[8] that compares, in this case, how advantaged and disadvantaged children are distributed across schools within a country. The Dissimilarity Index is defined as:

$$D_c = \frac{1}{2} \sum_{i=1}^{s} \left| \frac{a_i}{A_c} - \frac{b_i}{B_c} \right| \quad (1)$$

A_c and B_c are the total numbers of advantaged and disadvantaged students in country c, and a_i and b_i are the total number of advantaged and disadvantaged students in school i. In each country, s denotes the number of schools. D can be interpreted as the proportion of children from both groups that would need to be reallocated to other schools in order to attain an even distribution in all schools. If D is 0, all schools have the exact same proportion of advantaged and disadvantaged students. If D is 1, advantaged and disadvantaged students never attend the same schools, i.e., a country shows complete social segregation across schools. If D is 0.5, half of the students would need to be allocated to other schools in order to attain an even distribution. This measure is easy to compute and interpret, and it is the widely used non-spatial measure of segregation in education studies (e.g. Jenkins et al., 2008; Gutiérrez et al., 2019; Gorard & Smith, 2004).

Another aspect to consider is the dependency of segregation indexes on marginal distribution, since analysis on change between time or between spaces (e.g., countries) may be biased due to differences on the distribution of groups (e.g., low or high socioeconomic status) or units (e.g., schools). Regarding this, another advantage of D index is that is a measure independent of the marginal distribution of the groups (i.e., the sizes of low and high socioeconomic students) (Elbers, 2021), although it is dependent of the distribution of the units. In further analysis, we made further robustness analyses to approach this problem.

While D is the most commonly reported unevenness segregation measure, it has the disadvantage that it does not fulfil the 'principle of transfer', that is "if a student with a low social

[8] Unevenness indexes measure how far the distribution of the advantaged and disadvantaged children across schools is far from what would be the even share. Since we are comparing between countries with different socioeconomic realities, we estimated such measures are a better fit for this study that alternatives, such as exposition/isolation measures (c.f. Reardon et al., 2014)

position moves from a school with a higher share of low-social-position children to a school with a lower share, then overall segregation must fall" (Jenkins et al., 2008, p. 25). This criterion is regarded as important for segregation measures (Allen & Vignoles, 2007; James & Taeuber, 1985). Therefore, we also replicated the analyses by using the alternative Hutchens' Square-Root Index or H (Hutchens, 2001, 2004), that accounts for the principle of transfer. H is defined as:

$$H_c = \sum_{i=1}^{S} \left(\frac{a_i}{A_c} - \sqrt{\frac{a_i\, b_i}{A_c\, B_c}} \right) \tag{2}$$

The notation is the same as in Equation 1. In summary, H reflects how many more disadvantaged students than advantaged students are at each school i. Like D, the values of H range from 0 (even distribution of advantaged and disadvantaged students at all schools of a country) to 1 (complete social segregation across schools).

PISA, TIMSS, and PIRLS provide sampling weights to account for the complex sample designs and non-response. We adjusted both D and H by applying the survey weights to be representative of the target population. All proportions A_c, B_c, a_i, and b_i were weighted before estimating D and H for every country-by-cycle-by-study.

4.2.2. Measuring Socioeconomic Advantage and Disadvantage

We used two different variables to measure socioeconomic status. At first, we used the number of books at home as reported by students. Students with 100 or more books at home were defined as advantaged, and students with less than 100 books as disadvantaged. The biggest advantage of this socioeconomic status indictor is that it is the only one available for all cycles of PISA, PIRLS, and TIMSS. The share of missing values was on average 3% across studies and cycles. However, since the number of books at home variable has been criticized in the past (e.g., Engzell 2019), we replicated the analyses using parental education as a second socioeconomic status indicator. Parental education was reported by parents in the primary school studies and by the students in the secondary school studies. However, a disadvantage of using this variable concerns missingness. First, the parental education variable was not administered in all studies. While the number of books variable is available for all 32 study pairs, parental education is only available for 17 (see Table A.3). Second, there is a high rate of missing data in the parent questionnaires (on average 23% across studies and cycles for the number of books variable and 30% for the parental education variable) compared to a low rate in the student questionnaires (on average 3% across studies and cycles). We computed both segregation indexes D and H for both socioeconomic status measures and for each country-by-cycle-by-study observation (see Table 2).

We imputed missing data into the two socioeconomic status indicators using predictive mean matching (e.g. Rubin, 1987) in the R package mice (Buuren & Groothuis-Oudshoorn, 2011). The imputation model used information on age, gender, parental education, number of books at home, country of birth of parents, language at home, and achievement scores. We conducted a single imputation across all studies and cycles that used the respective indicator (all 32 study pairs for books

at home and 17 study pairs for parental education). We replicated our analyses using non-imputed dataset and the results remain stable since we are working with country-level aggregated data (non-reported). We then used the completed datasets to estimate the segregation indexes D and H for each (possible) country-by-cycle-by-study observation.

4.2.3. Tracking Indicator

Education systems sort their students into different ability tracks at different ages and grades. To determine the grade and age at which countries applied tracking, we followed the categorization proposed by Strello et al. (2021), who combined information from UNESCO-IBE (2007, 2012), Eurydice (2005, 2011, 2014, 2013b, 2013a), and OECD (2004, 2006, 2008, 2010).

Based on the tracking information, we constructed the two variables tracking grade and tracking age (i.e., modal age in tracking grade) that indicate whether students were already tracked when the secondary school studies were administered or not. In the analyses with TIMSS Pop. B secondary school data, we used the tracking grade as a cut-off criterion (i.e., grade 8 students are in a tracked system = tracked; tracking occurs after grade 8 = untracked). In the analyses with PISA secondary school data, we used the tracking age instead (i.e., students in the grade with the most 15-year-olds are in a tracked system = tracked; tracking occurs after the grade with the most 15-year-olds are = untracked). Table A.1. shows the tracking status for all countries in our sample. Thirteen out of the overall 75 observed countries apply tracking before the TIMSS Pop. B assessment takes place, and 17 before the PISA assessment takes place. Table 1 shows the number of early tracking countries per survey combination.

4.3. Analysis Method

In our analyses, we made use of the fact that none of the observed countries applied tracking at the primary school system, but on average 24% of the countries applied tracking before the secondary school assessments TIMSS Pop. B and PISA were conducted. This enabled us to compare the degree of social segregation across schools before and after the tracking took place, using late tracking countries with a comprehensive secondary school system as a control group. To increase the power, we replicated the analyses for different combinations of primary and secondary school data (see Figure 1 and Table 1).

Table 1. Overall Number of Countries and Divided by Tracking Status by Survey Combinations

No.	Survey combinations		Number of countries		Number of students	
	Primary school level data	Secondary school level data	All countries	Early tracking countries	Primary school	Secondary school
TIMSS Pop. A/PIRLS – PISA						
1	TIMSS Pop. A 1995	PISA 2000	19	6	67,566	60,607
2	PIRLS 2001	PISA 2000	21	7	84,268	118,768
3	PIRLS 2001	PISA 2003	18	7	75,131	117,541
4	PIRLS 2001	PISA 2006	23	8	92,751	146,013
5	TIMSS Pop. A 2003	PISA 2003	12	3	53,601	71,485
6	TIMSS Pop. A 2003	PISA 2006	14	3	61,149	97,524
7	TIMSS Pop. A 2003	PISA 2009	16	4	71,798	115,907
8	PIRLS 2006	PISA 2006	24	8	113,622	154,892
9	PIRLS 2006	PISA 2009	29	10	137,591	194,504
10	PIRLS 2006	PISA 2012	26	9	125,202	181,455
11	TIMSS Pop. A 2007	PISA 2006	22	7	97,198	137,723
12	TIMSS Pop. A 2007	PISA 2009	25	8	106,229	170,449
13	TIMSS Pop. A 2007	PISA 2012	24	8	106,229	167,644
14	TIMSS Pop. A 2011	PISA 2012	34	11	177,321	248,287
15	TIMSS Pop. A 2011	PISA 2015	34	11	176,966	232,63
16	PIRLS 2011	PISA 2012	32	10	189,029	254,883
17	PIRLS 2011	PISA 2015	35	11	201,371	255,522
18	TIMSS Pop. A 2015	PISA 2015	33	11	189,63	241,516
19	PIRLS 2016	PISA 2015	33	11	191,802	237,871
TIMSS Pop. A/PIRLS – TIMSS Pop. B						
20	TIMSS Pop. A 1995	TIMSS Pop. B 1995	26	6	94,732	94,488
21	TIMSS Pop. A 1995	TIMSS Pop. B 1999	18	4	71,623	81,661
22	PIRLS 2001	TIMSS Pop. B 2003	26	5	99,923	111,117
23	TIMSS Pop. A 2003	TIMSS Pop. B 2003	27	4	120,594	123,784
24	TIMSS Pop. A 2003	TIMSS Pop. B 2007	21	2	96,397	89,233
25	PIRLS 2006	TIMSS Pop. B 2007	25	3	111,734	108,322
26	PIRLS 2006	TIMSS Pop. B 2011	24	2	110,489	127,069
27	TIMSS Pop. A 2007	TIMSS Pop. B 2007	32	3	128,912	138,468
28	TIMSS Pop. A 2007	TIMSS Pop. B 2011	27	2	108,046	143,273
29	TIMSS Pop. A 2011	TIMSS Pop. B 2011	37	2	177,321	248,287
30	TIMSS Pop. A 2011	TIMSS Pop. B 2015	34	3	189,325	205,857
31	TIMSS Pop. A 2015	TIMSS Pop. B 2015	35	4	198,476	212,071
32	PIRLS 2016	TIMSS Pop. B 2015	32	4	186,406	195,704

4.3.1. Identification Strategy

Some countries have a comprehensive secondary school system while others have a tracked secondary school system. Simple comparisons of the degree of social segregation across countries with tracked and untracked school systems may be biased because the observed differences may have existed even before the students were tracked. For example, the degree of residential segregation may be higher in countries with a tracked secondary school system so that the social segregation across schools simply reflects the residential segregation in these countries. Furthermore, even in

education systems with a comprehensive school system, there may be transitions between separately operated primary and secondary schools, which may be socially biased. To address such possible heterogeneity, we followed Hanushek and Wößmann's (2006) approach to identify the effect of tracking using a difference-in-differences framework. This approach is based on the observation that no country has a tracked primary school system as early as elementary school. When comparing the degree of social segregation across schools in countries with tracked and untracked secondary school systems, we took into account the extent of social segregation across school that already existed before the sorting of students after primary school. Within this framework, the effect of tracking, y, is defined as the difference in social segregation across schools, *S*, between primary and secondary schools in countries with tracked and untracked secondary school systems:

$$\gamma = \left(S_{tracked,secondary} - S_{tracked,primary}\right) - \left(S_{untracked,secondary} - S_{untracked,primary}\right) \quad (3)$$

It seems useful to reiterate that in equation 3 the indices *tracked* and *untracked* refer to the secondary school system; in primary education all countries have a comprehensive school system. To estimate the effect of tracking on social segregation as defined in equation 3,we used a simple regression model, where we regressed the degree of social segregation in country *j* on a tracking indicator, *T* (1=tracked, 0=untracked) while controlling for the degree of segregation at the primary school level before the tracking took place (see Hanushek & Wößmann, 2006):

$$S_{secondary,j} = \alpha + \beta S_{primary,j} + \gamma T_{secondary,j} + e_j \quad (4)$$

The key parameter of interest in Equation 4 is γ, which is the effect of early tracking on social segregation across schools. The equation does not include the tracking status at the primary school level because no country in our sample applied tracking before grade 4. We replicated the analyses for all combinations of primary and secondary school studies (see Figure 1 and Table 1), as well as for both segregation indexes (*D* and *H*) and both socioeconomic status indicators (number of books and parental education).

4.3.2. Mean of Effects

We computed weighted mean effect sizes to summarize the findings of the $i = 32$ replications (respectively i 17 replications when the segregation index is based on the parental education variable) across the different combinations of primary and secondary school data. We did this separately for the two types of secondary school datasets, the two segregation indexes, and both socioeconomic status indicators.

For this purpose, we used a formula provided by Card (2012) to combine effect sizes in meta-analyses. The basic idea is that some effect estimates are more reliable than others, 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:

$$w_i = \frac{1}{SE_i^2} \quad (5)$$

The mean effects are defined as the weighted sum of the effect sizes (γ_i) from the up to 32 replications:

$$\overline{\gamma} = \frac{\sum(w_i * \gamma_i)}{\sum w_i} \quad (6)$$

The weights were also used to compute a standard error for $\overline{\gamma}$. For this purpose, we used the square root of the inverse value of the sum of the weights:

$$SE_{\overline{\gamma}} = \sqrt{\frac{1}{\sum w_i}} \quad (7)$$

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.

5. Results

5.1. Descriptive Analysis

Table 2 shows descriptive statistics across all countries as well as separately for early and late tracking countries. The overview shows that the differences in segregation between primary and secondary school level are usually smaller in late tracking countries than in early tracking countries. Following the social segregation level measured by number of books, the segregation between primary and secondary schools remains relatively constant between levels within late tracking countries. At the same time, the segregation level increases considerably within early tracking countries; this pattern is observed for both the D and H measures and both PISA and TIMSS Pop. B datasets. Social segregation measured by parental education results in a similar conclusion. In addition, there are no noticeable differences of social segregation levels between early and late tracking countries at primary school level but there are large differences in social segregation across schools at secondary level, suggesting the importance of early tracking to explain social segregation across schools.

Table 2. Descriptive Statistics of the Indexes of Social Segregation across School at Primary and Secondary Level in the Overall Country Sample and Divided by Tracking Status

		Segregation based on books at home				Segregation based on parental education			
		M	SD	Min	Max	M	SD	Min	Max
Dissimilarity Index (*D*)									
PIRLS/TIMSS Pop. A – PISA									
All countries	Primary school	0.313	0.063	0.142	0.564	0.371	0.068	0.237	0.601
	Secondary school	0.326	0.065	0.195	0.476	0.353	0.065	0.217	0.613
Late tracking	Primary school	0.310	0.067	0.142	0.564	0.370	0.069	0.249	0.559
	Secondary school	0.307	0.061	0.195	0.470	0.344	0.065	0.217	0.613
Early tracking	Primary school	0.318	0.051	0.238	0.516	0.374	0.068	0.237	0.601
	Secondary school	0.366	0.053	0.199	0.476	0.371	0.062	0.219	0.523
PIRLS/TIMSS Pop. A – TIMSS Pop. B									
All countries	Primary school	0.325	0.082	0.142	0.631	0.389	0.091	0.250	0.649
	Secondary school	0.325	0.071	0.150	0.556	0.345	0.081	0.202	0.609
Late tracking	Primary school	0.326	0.086	0.142	0.631	0.389	0.095	0.250	0.649
	Secondary school	0.324	0.074	0.150	0.556	0.343	0.080	0.202	0.609
Early tracking	Primary school	0.316	0.043	0.244	0.455	0.387	0.060	0.295	0.507
	Secondary school	0.332	0.045	0.260	0.442	0.371	0.084	0.245	0.479
Square-Root Index (*H*)									
PIRLS/TIMSS Pop. A – PISA									
All countries	Primary school	0.335	0.087	0.143	0.720	0.391	0.087	0.239	0.676
	Secondary school	0.334	0.075	0.195	0.548	0.359	0.077	0.206	0.736
Late tracking	Primary school	0.333	0.094	0.143	0.720	0.384	0.085	0.256	0.668
	Secondary school	0.316	0.075	0.195	0.548	0.350	0.077	0.206	0.736
Early tracking	Primary school	0.338	0.069	0.246	0.607	0.405	0.089	0.239	0.676
	Secondary school	0.371	0.058	0.200	0.495	0.380	0.072	0.219	0.558
PIRLS/TIMSS Pop. A – TIMSS Pop. B									
All countries	Primary school	0.352	0.117	0.143	0.827	0.407	0.124	0.255	0.817
	Secondary school	0.341	0.091	0.151	0.733	0.349	0.092	0.205	0.717
Late tracking	Primary school	0.355	0.123	0.143	0.827	0.407	0.129	0.255	0.817
	Secondary school	0.342	0.095	0.151	0.733	0.346	0.092	0.205	0.717
Early tracking	Primary school	0.327	0.058	0.246	0.532	0.408	0.082	0.296	0.591
	Secondary school	0.333	0.050	0.251	0.489	0.372	0.091	0.246	0.484

Note. Number of books at home indicates whether a household contains at least 100 books. Parental education indicates whether at least one parent reports attaining a university degree. All statistics were estimated based on 474 (PIRLS/TIMSS Pop. A – PISA) respectively 364 (PIRLS/TIMSS Pop. A – TIMSS Pop. B) country-by-cycle-by-study observations (see Table 1).

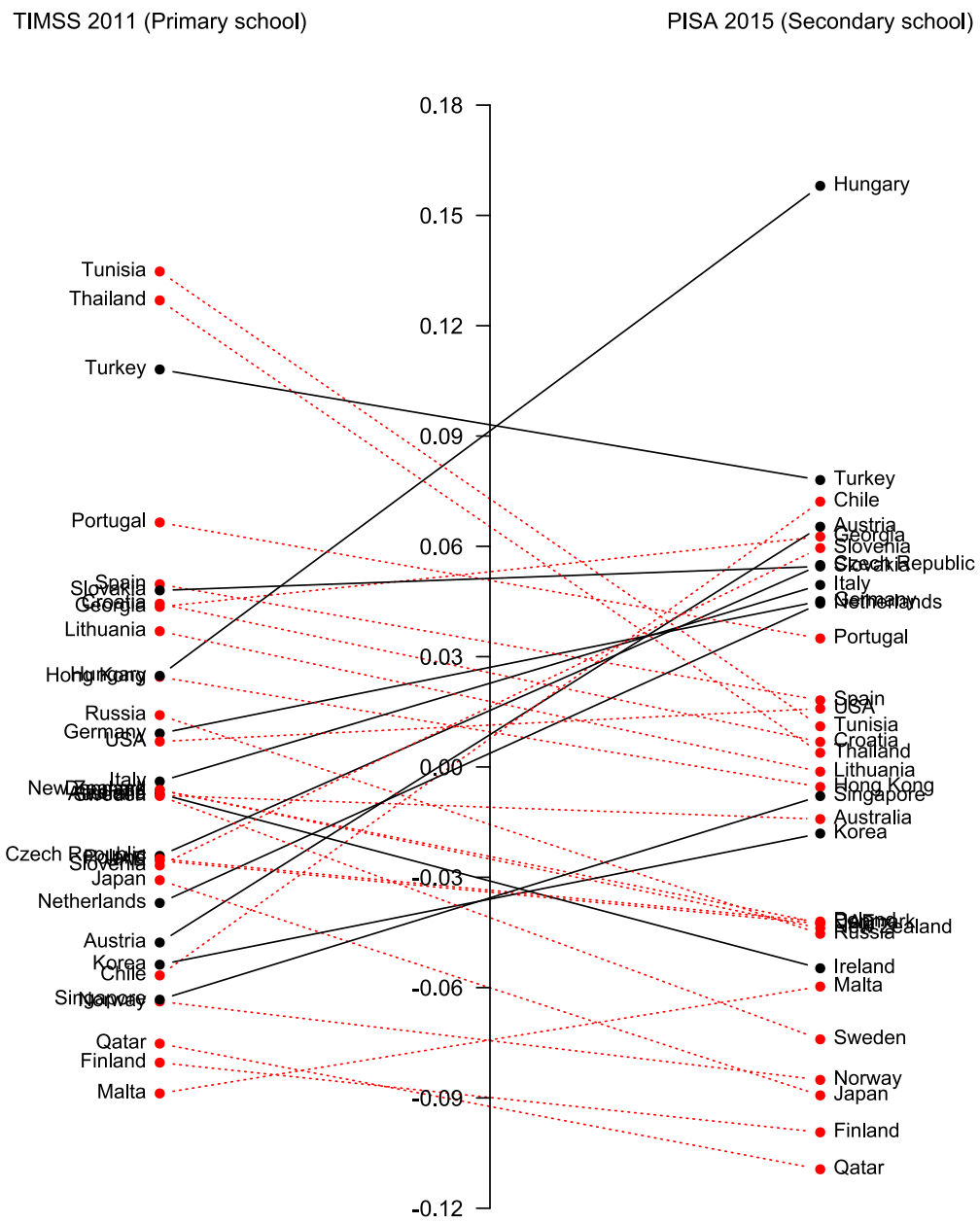


Figure 2. Social Segregation across Schools (Measured as *D* Based on Books at Home) on Primary and Secondary School Level (TIMSS Pop. A 2011 – PISA 2015). Solid lines correspond to tracking countries, dotted lines correspond to late tracking countries.

To further illustrate our analysis, in Figure 2, we plotted the difference in social segregation levels between primary and secondary school level for countries with a tracked and untracked secondary school system. The left side shows the segregation levels at primary school and the right side shows the segregation levels at secondary school. To ease the interpretation, the segregation levels are centralized on the mean (e.g., 0 = mean of segregation level across countries). This example shows a similar pattern to the one seen in Table 2 but looking only at the combination sample of TIMSS Pop. A 2011 and PISA 2015. In most countries with a tracked secondary school system we observed an increase in social segregation levels between primary and secondary school, while that

pattern is not observable at all in countries with an untracked secondary school system. The next section studies this further by means of difference-in-difference models that are replicated across 32 survey combinations.

5.2. Main Results

For each of the up to 32 survey combinations (see Figure 1 and Table 2) and separately for the two segregation indexes (D and H) and two socioeconomic status indicators (books at home and parental education), we replicated a regression model. In this model, we regressed the social segregation across schools at the secondary level on the tracking dummy variable, while controlling for the social segregation at the primary school level. The results of the single replications are available in Table A.2 (segregation based on number of books) and Table A.3 (segregation based on parental education). As main results, we report summaries of these results, since the single replications contained different samples, which often contained few countries. To ease the interpretation, we further computed the standardized effect size Cohen's d by dividing $\overline{\gamma}$ by the SD of the segregation indexes at the secondary school level (see Table 2). To interpret this effect size, we considered values up to $d = 0.20$ to be small, values up to $d = 0.50$ to be moderate, and values up to $d = 0.80$ as large (Cohen 1969).

The weighted mean effects of between-school tracking on social segregation across schools are summarized in Table 3. These appear separately for the two types of secondary school level datasets (TIMSS Pop. A/PIRLS – PISA and TIMSS Pop. A/PIRLS – TIMSS Pop. B), two segregation measures (D and H), and two socioeconomic status indicators (books at home and parental education).

Table 3. Weighted Mean Effects of Early Between-school Tracking on Social Segregation across Schools for Different Types of Survey Combinations, Measured as Dissimilarity Index D and Square-Root Index H

	Segregation based on books at home			Segregation based on parental education		
	Estimate	*d*	*SE*	Estimate	*d*	*SE*
Dissimilarity Index *D*						
PIRLS/TIMSS Pop. A – PISA	0.058***	0.892	0.005	0.026***	0.400	0.006
PIRLS/TIMSS Pop. A – TIMSS Pop. B	0.019*	0.268	0.008	0.042**	0.519	0.012
Square-Root Index *H*						
PIRLS/TIMSS Pop. A – PISA	0.057***	0.760	0.005	0.018**	0.234	0.007
PIRLS/TIMSS Pop. A – TIMSS Pop. B	0.014	0.154	0.009	0.039**	0.424	0.014

Note. Significance levels: *** $p < 0.001$; ** $p < 0.010$; * $p < 0.050$

The main results in Table 3 provide compelling evidence that tracking reinforces social segregation across schools. When focusing on the Dissimilarity Index D and the socioeconomic status indicator books at home first, we observe a mean effect of tracking of 0.058 in the analyses with PISA as secondary school data. This effect corresponds to a standardized effects size of $d = 0.892$. The same analyses with TIMSS Pop. B secondary school data replicate the previously

presented finding, with a smaller but still statistically significant effect of 0.019 or $d = 0.268$. The analyses using the Square-Root Index H instead confirm the previously reported findings, with very similar effects for both PISA and TIMSS Pop. B data.

The analyses using parental education as an indicator of the social background contribute further evidence to the main finding that tracking reinforces segregation. Interestingly, the effect is larger for TIMSS Pop. B data ($d = 0.519$) than for the analyses with PISA data ($d = 0.400$) in this case. However, these differences could also be due to the fact that the data basis differed, as the questions on parental education were not administered across all studies. Again, the analyses with the Square-Root Index H were qualitatively the same.

5.3. Further Analyses

We ran alternative specifications to test the robustness of our analyses. We focused on the Dissimilarity Index D in reporting these findings, since the analyses using H led to very similar results. The results of these specifications are summarized in Table 4.

Table 4. Weighted Mean Effects of Early Between-school Tracking on Social Segregation across Schools, Measured as Dissimilarity Index D for Both Types of Survey Combinations in Further Analyses

	Segregation based on books at home			Segregation based on parental education		
	Estimate	*d*	*SE*	Estimate	*d*	*SE*
PIRLS/TIMSS Pop. A – PISA						
(1) At least 15 students per school	0.057***	0.877	*0.005*	0.028***	0.431	*0.006*
(2) At least 20 students per school	0.058***	0.892	*0.005*	0.023***	0.321	*0.006*
(3) Alternative specification: 25+ books at home	0.076***	1.179	*0.005*			
(4) Alternative specification: 200+ books at home	0.050***	0.579	*0.006*			
(5) Alternative specification: Parental educ. over ISCED2				0.019	0.118	*0.012*
(6) Alternative specification: ICC non-recoded variable	0.084***	1.178	*0.006*	0.032***	0.442	*0.006*
(7) Only same-cohort comparisons	0.054***	0.831	*0.008*	0.020*	0.308	*0.009*
(8) Only same-period comparisons	0.058***	0.892	*0.007*	0.021*	0.323	*0.008*
(9) Unweighted mean effects	0.058	0.892		0.030	0.462	
PIRLS/TIMSS Pop. A – TIMSS Pop. B						
(1) At least 15 students per school	0.017*	0.239	*0.008*	0.037**	0.457	*0.012*
(2) At least 20 students per school	0.015*	0.210	*0.008*	0.041***	0.446	*0.012*
(3) Alternative specification: 25+ books at home	0.024**	0.345	*0.008*			
(4) Alternative specification: 200+ books at home	0.016	0.173	*0.008*			
(5) Alternative specification: Parental educ. over ISCED2				0.050*	0.411	*0.021*
(6) Alternative specification: ICC non-recoded variable	0.028***	0.440	*0.007*	0.044**	0.476	*0.014*
(7) Only same-cohort comparisons	0.015	0.211	*0.013*	0.036	0.444	*0.019*
(8) Only same-period comparisons	0.022*	0.310	*0.009*	0.046**	0.568	*0.017*
(9) Unweighted mean effects	0.016	0.225		0.038	0.469	

Note. Significance levels: *** $p < 0.001$; ** $p < 0.010$; * $p < 0.050$

5.3.1. Alternative Socioeconomic Groups

The estimation of segregation measures may be sensitive to the grouping criteria of the socioeconomic variables. To address this, we replicated the results using alternative specifications of the socioeconomic groups.

Regarding number of books at home, we used alternatively as cutoff having over 25 books at home versus 25 or less books, in contrast with the main analyses' cutoff of more than 100 books. As a second alternative we also replicated the results grouping students between more than 200 books at home versus 200 or less books. Due to the original categories present on the surveys and the incompatibility with the other surveys' original categories, we dropped from these further analyses PISA 2000 and, therefore, two study pairs (PIRLS 2001 – PISA 2000, and TIMSS 1995 – PISA 2000). Regarding parental education, we replicated the analyses grouping students with parents maximum ISCED level 2 (roughly completing lower secondary school) versus less than that level (incomplete ISCED level 2, ISCED level 1 or less). Most of the results are consistent with the main analyses (see rows (3) to (5) in Table 4), but there are some exceptions where there are non-significant effects, in specific, the alternative to the recoding of parental education when using PISA data, and the alternative to the recoding of number of books at home (200+ books) when using TIMSS Pop. B data.

In addition, we replicated our results using the intraclass correlation (ICC) as a proxy of school segregation, making use of the full variability of the variables instead of dichotomizing on two disadvantaged and advantaged groups. The ICC was calculated in base of a multilevel null model predicting both number of books at home and parental education clustering by school and it can be interpreted as how much the school accounts for the variance of both variables. The disadvantage of this method is that the original variables are categorical rather than continuous, plus the categories between each study and each cycle tend to differ especially on the earlier versions of the studies. Therefore, we recommend caution when interpreting these analyses. Our results (see row (6) in Table 4) indicate that the ICC of both number of books at home and parental education increase further in tracking countries. Overall, these extra sensitivity analyses tend to confirm our main results.

5.3.2. Size Restriction of School Samples

The estimation of segregation measures is sensitive to the sizes of student samples in each school. Therefore, we replicated our analyses restricting them to schools with at least 15 students and to schools with at least 20 students (see row (1) and (2), respectively, in Table 4). When applying these restrictions, the mean effects were very similar to the main findings.

5.3.3. Cohort Effects, Period Effects, and Sample Inflation

We applied two strategies to combine primary and secondary school studies: studies that were administered in approximately the same year (i.e., same-year comparisons) and studies that sampled approximately the same student cohort in primary and secondary school (i.e., same-cohort

comparisons). The first strategy is subject to cohort effects while the second approach may be affected by period effects (see Blanchard et al., 1977). Further, using the same data multiple times leads to sample inflation. To investigate whether these issues actually had an impact on our analyses, we calculated two different mean effects in which each assessment cycle was only included in one comparison and where only same-cohort (see rows (7) in Table 4) and respective same-year (see rows (8) in Table 4) comparisons were made. Again, the effect of early between-school tracking is robust across the different specifications and we did not observe large variations in the mean effect estimates.

5.3.4. Unweighted Mean Effects

In the main analyses, the estimation of the mean effects was weighted by the inverse of the standard error (Card, 2012). This implies that the more pronounced tracking effects or the more participating countries a study pair had, the more it contributed to the mean effects. As an alternative, we estimated unweighted mean effects (see row (9) in Table 4). The weighted and unweighted mean effects were similar, further confirming the robustness of our main analyses.

6. Discussion and Conclusion

Previous research on the effects of tracking focused almost exclusively on inequalities in students' academic performance. We extended this line of research to focus on the effects of between-school tracking on social segregation across schools to provide a more holistic evaluation of the effects of early tracking. We believe that the integration of students from various social backgrounds is a valuable good in democratic societies. Against this background, we consider it problematic if socially privileged and disadvantaged children have little contact with each other in the school context from an early age. In our opinion, the integration of different social classes is a core task of schools and a from an early age on socially segregated school system contradicts this claim.

Our analyses found strong evidence that between-school tracking increases social segregation across school. Previous research applied purely correlative approaches based on cross-sectional data and found that tracking was one of the characteristics that was associated with large degrees of social segregation. We provide more robust evidence for a causal relationship because our difference-in-differences approach circumvents effects of unobserved covariates. Controlling for the segregation that already exists at the primary school level permitted us to control for other important sources of social segregation across schools that are independent from tracking, such as the residential segregation and other mechanisms of school choice. However, it is important to consider that these interpretations rest on assumptions that we can not test under our current design. These limitations are discussed on greater detail below.

Proponents of early between-school ability tracking argue that there would be a trade-off between the efficiency and the equality of learning opportunities at school. Previous studies that used international comparative data challenged this supposed trade-off. Our study complements this research by providing further evidence on *how* tracking widens the social divide between

socioeconomically disadvantaged and advantaged students. Taken together, tracking decisions seem to depend on the socioeconomic status of children and therefore increase social segregation across school at the secondary school level (see findings of the present study), and arguably partially independent of the previous academic achievement although we cannot assess that through our study. Considering that enhanced academic school tracks provide better academic learning opportunities (Skopek et al., 2019; Dollmann, 2019) and disadvantaged students profit from advantaged schoolmates (Benito et al., 2014), early tracking increases achievement gaps between social groups (Strello et al., 2021; Lavrijsen & Nicaise, 2015; van de Werfhorst & Mijs, 2010; van de Werfhorst, 2018). At the same time, there is currently no robust evidence for early tracking increasing general achievement levels (Hanushek & Wößmann, 2006; Strello et al., 2021; Lavrijsen & Nicaise, 2016).

6.1. Limitations and Future Research

Our results contribute evidence on the impact of between-school tracking on social segregation across schools. We acknowledge, however, that this is just one type of differentiation among others, such as course-by-course ability tracking within schools (Skopek et al., 2019; Hillmert & Jacob, 2010). In the present study, we consider both countries with a purely comprehensive and countries within-school tracking systems as the control group of untracked secondary school systems. It seems plausible that the effects of between-school tracking would be even stronger if within-school tracking education systems would have been excluded from the control group of late tracking countries. In the same vein, we did not differentiate between early tracking countries with more or less objective track allocation procedures because this information was not available and the number of tracked system was small. It is plausible that the effects of tracking are higher in early tracking countries where teachers and parents have a greater influence on the track allocation decision than in countries with more objective procedures (e.g., decision based on central exams).

Another set of limitations are related to the sampling structure. Since our analyses require international comparative data, using PISA, TIMSS, and PIRLS is most adequate. However, all three international large-scale assessments' sampling strategies are designed for representative student samples, while the samples may be less representative for schools (e.g., small schools can be excluded). In addition, PIRLS and TIMSS follow different sampling strategies than PISA. While the former two sample one classroom per school in most countries, the latter samples students of different classrooms. The approach of TIMSS and PIRLS restrict school heterogeneity (assuming non-random sorting of students between classrooms) but are more comparable between them than PISA. Another possible limitation are the different school sizes at the primary and secondary level. In some countries, primary schools are much smaller than secondary schools, for instance, which might affect the proportions and therefore segregation measures. These three limitations do not invalidate our analyses, since school samples are representative regardless of their reduced size and some exclusions. The segregation levels do not seem to vary drastically between TIMSS Pop. B and PISA

and the segregation levels seem to be stable between primary and secondary schools in countries without early tracking (see Table 2). This suggests that the data comparisons are adequate.

Studying school segregation with international large-scale assessments brings some more issues due the designs of small samples per cluster (in this case, schools). Segregation measures are subject to bias when the minority groups are particularly small within each cluster, even if the total sample of individuals and clusters are big (Carrington & Troske, 1998; Winship, 1977). However, we did not study extreme minorities but rather big groups. The minority groups in average represented 31-33% of the sample (for the grouping by number of books and parental education, respectively), meaning 8.9 and 9.6 minority students per school. Therefore, this research shall not be as affected as other segregation studies. Future articles dealing with school segregation using large-scale assessments may decide to address this issue by estimating the difference between the observed segregation and the random segregation (as suggested by Winship [1977]). Another alternative may be using the information from the school questionnaires (answered by the directors) that is not affected by the sampling, but it was not possible to implement in this article due our combination of different studies and cycles.

Future research on tracking effects on school segregation may address issues that we could not address with the data at hand. First, future studies may aim for a more comprehensive measurement of the school structure and school choice mechanisms in the education systems (e.g., within-school tracking, selection based on recommendation vs. central entrance exams). Second, it may be interesting to disentangle the primary and secondary effects of socioeconomic status. Segregation related to primary effects may be more defensible than due to secondary effects. Likewise, it would be interesting to investigate to what extent the increases in social segregation across schools have direct consequence for the achievement gap between disadvantaged and advantaged students, and how much of these increases are as a direct consequence of the school track decisions. Finally, a question that has not been directly touched upon by previous research relates to the consequences of social segregation caused by early tracking on social inequalities in education (e.g., in student achievement, civic knowledge, educational attainment, learning motivations, or academic aspirations).

6.2. Conclusion

Regardless of the remaining questions, this article contributes to understanding the effects of institutional features on the school segregation across schools within education systems. Most previous studies on early tracking effects focus on student achievement outcomes, while the joint schooling of social groups is also an important asset in democratic and egalitarian societies. The stratification of school systems reinforces the separation of students from different social origins, which must be considered by policymakers.

7. References

Allen, R., & Vignoles, A. (2007). What should an index of school segregation measure? *Oxford Review of Education*, *33*(5), 643–668. https://doi.org/10.1080/03054980701366306

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

Benito, R., Alegre, M. À., & Gonzàlez-Balletbò, I. (2014). School Segregation and Its Effects on Educational Equality and Efficiency in 16 OECD Comprehensive School Systems. *Comparative Education Review*, *58*(1), 104–134. https://doi.org/10.1086/672011

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, 92*(1), 1545–1572. https://doi.org/10.1093/sf/sou003

Bonal, X., Zancajo, A., & Scandurra, R. (2019). Residential segregation and school segregation of foreign students in Barcelona. *Urban Studies*, *56*(15), 3251–3273. https://doi.org/10.1177/0042098019863662

Bottia, M. C. (2019). *Immigrant Integration and Immigrant Segregation*. https://prrac.org/immigrant-integration-and-immigrant-segregation-the-relationship-between-school-and-housing-segregation-and-immigrants-future-in-the-u-s-martha-cecilia-bottia-april-2019/

Burger, K. (2019). The socio-spatial dimension of educational inequality: A comparative European analysis. *Studies in Educational Evaluation*, *62*(May), 171–186. https://doi.org/10.1016/j.stueduc.2019.03.009

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

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

Carrington, W. J., & Troske, K. R. (1998). Interfirm segregation and the black/white wage gap. *Journal of Labor Economics*, *16*(2), 231-260. https://doi.org/10.1086/209888

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

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., & Savage, C. (2015). Socioeconomic Segregation Between Schools in the United States and Latin America, 1970–2012. In G. W. McCarthy, G. K. Ingram, & S. A. Moody (Eds.), *Land and the City* (pp. 394–423). Lincoln Institute of Land Policy. https://www.lincolninst.edu/publications/conference-papers/socioeconomic-segregation-between-schools-united-states-latin-america

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Science*. Academic Press.

Cordero, J. M., Cristóbal, V., & Santín, D. (2018). Causal inference on education policies: A survey of empirical studies using PISA, TIMSS and PIRLS. *Journal of Economic Surveys*, *32*(3), 878–915. https://doi.org/10.1111/joes.12217

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

Duncan, O. D., & Duncan, B. (1955). A Methodological Analysis of Segregation Indexes. *American Sociological Review*, *20*(2), 210. https://doi.org/10.2307/2088328

Elbers, B. (2021). A method for studying differences in segregation across time and space. *Sociological Methods & Research*. https://doi.org/10.1177%2F0049124121986204

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

Erikson, R., & Johnson, J.O. (1996). Explaining class inequality in education: The Swedish test case. In R. Erikson & J.O. Jonsson (Eds.), *Can education be equalized?* (pp. 1-65). Oxford: Westview Press

Eurydice. (2005). Key Data on Education in Europe 2005. In *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*. https://doi.org/10.2797/40560

Eurydice. (2013b). *The Structure of the European Education Systems 2013/14: Schematic Diagrams*. https://doi.org/10.2797/206797

Eurydice. (2014). *The Structure of the European Education Systems 2014/15: Schematic Diagrams*. 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

Gutiérrez, G., Jerrim, J., & Torres, R. (2019). School Segregation Across the World: Has Any Progress Been Made in Reducing the Separation of the Rich from the Poor? *Journal of Economic Inequality*. https://doi.org/10.1007/s10888-019-09437-3

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., & 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

Hindriks, J., Verschelde, M., Rayp, G., & Schoors, K. (2010). School tracking, social segregation and educational opportunity: evidence from Belgium. In *CORE Discussion Paper 2010/81*. https://uclouvain.be/fr/node/26415#Alfresco

Hutchens, R. (2001). Numerical measures of segregation: desirable properties and their implications. *Mathematical Social Sciences*, *42*(1), 13–29. https://doi.org/10.1016/S0165-4896(00)00070-6

Hutchens, R. (2004). One measure of segregation. *International Economic Review*, *45*(2), 555–578. https://doi.org/10.1111/j.1468-2354.2004.00136.x

James, D. R., & Taeuber, K. E. (1985). Measures of Segregation. *Sociological Methodology*, *15*, 1–32. https://doi.org/10.2307/270845

Jenkins, S. P., Micklewright, J., & Schnepf, S. V. (2008). Social segregation in secondary schools: How does England compare with other countries? *Oxford Review of Education*, *34*(1), 21–37. https://doi.org/10.1080/03054980701542039

Korthals, R. A. & Dronkers, J. (2016). Selection on performance and tracking. *Applied Economics, 48*(30), 2836-2851. https://doi.org/10.1080/00036846.2015.1130789

Kristen, C. (2005). *School Choice and Ethnic School Segregation*. Waxmann.

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

Malmberg, B., & Andersson, E. K. (2020). How Well Do Schools Mix Students from Different Neighborhoods? School Segregation and Residential Segregation in Swedish Municipalities. *Geographical Analysis*, 1–25. https://doi.org/10.1111/gean.12233

Murillo, F. J., Hernández-Castilla, R., Martínez-Garrido, C., & Hidalgo, N. (2018). Una Panorámica de la Segregación Social de los Centros de Educación Secundaria en Iberoamérica. In F. J. Murillo (Ed.), *Avances en Democracia y Liderazgo Distribuido en Educación: Actas del II Congreso Internacional de Liderazgo y Mejora de la Educación* (pp. 559–564). http://hdl.handle.net/10486/683111

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. 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. (2019). Balancing School Choice and Equity: An International Perspective Based on PISA. In *PISA*. OECD Publishing. https://doi.org/10.1787/2592c974-en

Parker, P. D., Jerrim, J., Schoon, I., & Marsh, H. W. (2016). A Multination Study of Socioeconomic Inequality in Expectations for Progression to Higher Education: The Role of Between-School Tracking and Ability Stratification. *American Educational Research Journal*, *53*(1), 6–32. https://doi.org/10.3102/0002831215621786

Reardon, S. F., & Owens, A. (2014). 60 years after Brown: Trends and consequences of school segregation. *Annual Review of Sociology*, *40*, 199-218. https://doi.org/10.1146/annurev-soc-071913-043152

Reichelt, M., Collischon, M., & Eberl, A. (2019). School tracking and its role in social reproduction: reinforcing educational inheritance and the direct effects of social origin. *British Journal of Sociology*, *70*(4). https://doi.org/10.1111/1468-4446.12655

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. 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

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). https://doi.org/10.4337/9781788110426.00022

Strello, A., Strietholt, R., Steinmann, I., & Siepmann, C. (2021). Early tracking and different types of inequalities in achievement: difference-in-differences evidence from 20 years of large-scale assessments. *Educational Assessment, Evaluation and Accountability*, *33*(1), 139-167. https://doi.org/10.1007/s11092-020-09346-4

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 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

Winship, C. (1977). A revaluation of indexes of residential segregation. *Social Forces*, *55*(4), 1058-1066. https://doi.org/10.1093/sf/55.4.1058

8. Appendices

Table A.1. Between-school Tracking Status According to Age and Grade in All Countries and Regions

Country	Tracking age	Tracking grade	Tracked	Country (cont.)	Tracking age	Tracking grade	Tracked
Abu Dhabi, ARE	15	9		Kuwait	18	12	
Alberta, CAN	18	12		Latvia	16	9	
Algeria	15.5	9		Lithuania	15	8	
Argentina	15	9		Luxembourg	12	6	Yes
Armenia	15	9		Macedonia	15	8	
Australia	16	10		Malta	16	11	
Austria	10	4	Yes	Moldova	15	10	
Bahrain	15	9		Mongolia	16	8	
Flanders, BEL	12	6	Yes	Morocco	15	9	
Brit. Col., CAN	18	12		Netherlands	12	6	Yes
B. Aires, ARG	12	6	Yes	New Zealand	16	11	
Bulgaria	14	7	Yes	Norway	16	10	
Canada	18	12		Oman	16	10	
Chile	16	10		Ontario, CAN	18	12	
Colombia	15	9		Philippines	16	10	
Croatia	15	8		Poland	15	9	
Cyprus	15	9		Portugal	15	9	
Czech Republic	11	5	Yes	Qatar	15	9	
Denmark	16	10		Quebec, CAN	18	12	
Dubai, UAE	15	9		Romania	14	8	Yes[a]
El Salvador	16	9		Russian Fed.	15	9	
England	16	11		Saudi Arabia	15	9	
Finland	16	9		Scotland	16	11	
France	15	9		Serbia	15	8	
Georgia	15	9		Singapore	12	6	Yes
Germany	10	4	Yes	Slovakia	10	4	Yes
Greece	15	9		Slovenia	15	9	
Hong Kong	16	11		Spain	15	9	
Hungary	10	4	Yes	Sweden	16	9	
Iceland	16	10		Taiwan	15	9	
Indonesia	16	9		Thailand	15	9	
Iran	15	9		Trinidad & Tob.	11	5	Yes
Ireland	12	6	Yes	Tunisia	16	10	
Israel	15	10		Turkey	14	8	Yes[a]
Italy	14	8	Yes[a]	Ukraine	15.5	9	
Japan	15	9		Un. Arab Emir.	15	9	
Kazakhstan	15	9		United States	18	12	
Korea	14	9	Yes[a]				

Note. [a] Country is considered a tracked secondary school system in PISA analyses but an untracked system in TIMSS Pop. B analyses. Tracking age reflects the modus age in the grade after tracking takes place. Tracking age and grade depict the year after which the first between-school differentiation in each country or region (i.e. the last year in which students are still in an untracked system). Source: Strello et al. (2021)

Table A.2. Regression Coefficients of the Effect of Tracking on Social Segregation across Schools Based on Number of Books at Home by Survey Combination

		Dissimilarity Index (*D*)		Square-Root Index (*H*)		
	Survey combination	Estimate	*SE*	Estimate	*SE*	*N*
1	TIMSS Pop. A 1995 – PISA 2000	0.028	0.027	0.026	0.029	19
2	PIRLS 2001 – PISA 2000	0.068*	0.028	0.070*	0.029	21
3	PIRLS 2001 – PISA 2003	0.101***	0.020	0.098***	0.021	18
4	PIRLS 2001 – PISA 2006	0.035	0.026	0.028	0.029	23
5	TIMSS Pop. A 2003 – PISA 2003	0.107**	0.025	0.103**	0.027	12
6	TIMSS Pop. A 2003 – PISA 2006	0.066+	0.032	0.058	0.040	14
7	TIMSS Pop. A 2003 – PISA 2009	0.060+	0.031	0.053	0.035	16
8	PIRLS 2006 – PISA 2006	0.062*	0.023	0.063*	0.026	24
9	PIRLS 2006 – PISA 2009	0.036	0.022	0.036	0.024	29
10	PIRLS 2006 – PISA 2012	0.062**	0.016	0.061**	0.018	26
11	TIMSS Pop. A 2007 – PISA 2006	0.058*	0.023	0.056+	0.028	22
12	TIMSS Pop. A 2007 – PISA 2009	0.058*	0.022	0.057*	0.027	25
13	TIMSS Pop. A 2007 – PISA 2012	0.062**	0.020	0.059*	0.025	24
14	PIRLS 2011 – PISA 2012	0.057**	0.017	0.057**	0.019	34
15	TIMSS Pop. A 2011 – PISA 2012	0.056**	0.016	0.057**	0.018	34
16	PIRLS 2011 – PISA 2015	0.043*	0.018	0.041*	0.020	32
17	TIMSS Pop. A 2011 – PISA 2015	0.066**	0.017	0.068**	0.019	35
18	TIMSS Pop. A 2015 – PISA 2015	0.051**	0.018	0.049*	0.019	33
19	PIRLS 2016 – PISA 2015	0.036+	0.020	0.037+	0.021	33
	Weighted mean effect (PIRLS/TIMSS Pop. A – PISA)	0.058***	0.005	0.057***	0.005	
20	TIMSS Pop. A 1995 – TIMSS Pop. B 1995	0.032	0.020	0.024	0.020	26
21	TIMSS Pop. A 1995 – TIMSS Pop. B 1999	0.015	0.030	0.013	0.030	18
22	PIRLS 2001 – TIMSS Pop. B 2003	0.020	0.023	0.016	0.023	26
23	TIMSS Pop. A 2003 – TIMSS Pop. B 2003	0.022	0.024	0.014	0.026	27
24	TIMSS Pop. A 2003 – TIMSS Pop. B 2007	0.004	0.040	-0.003	0.051	21
25	PIRLS 2006 – TIMSS Pop. B 2007	0.021	0.036	0.010	0.047	25
26	PIRLS 2006 – TIMSS Pop. B 2011	0.000	0.035	-0.011	0.040	24
27	TIMSS Pop. A 2007 – TIMSS Pop. B 2007	-0.011	0.036	-0.013	0.046	32
28	TIMSS Pop. A 2007 – TIMSS Pop. B 2011	0.005	0.035	0.001	0.035	27
29	TIMSS Pop. A 2011 – TIMSS Pop. B 2011	0.020	0.033	0.012	0.035	37
30	TIMSS Pop. A 2011 – TIMSS Pop. B 2015	0.034	0.031	0.027	0.036	34
31	TIMSS Pop. A 2015 – TIMSS Pop. B 2015	0.030	0.025	0.025	0.030	35
32	PIRLS 2016 – TIMSS Pop. B 2015	0.014	0.024	0.014	0.027	32
	Weighted mean effect (PIRLS/TIMSS Pop. A – TIMSS Pop. B)	0.019*	0.008	0.014	0.009	

Note. The estimate columns report the effects of between-school tracking on social segregation across schools at secondary school level. The estimates of the control variable social segregation across schools at primary school level are omitted. *N* = number of countries included in survey combination. Significance levels: *** $p < 0.001$; ** $p < 0.010$; * $p < 0.050$; + $p < 0.100$.

Table A.3. Regression Coefficients of the Effect of Tracking on Social Segregation across Schools Based on Parental Education by Survey Combination

	Dissimilarity Index (*D*)		Square-Root Index (*H*)		
Survey combination	Estimate	*SE*	Estimate	*SE*	*N*
1 TIMSS Pop. A 1995 – PISA 2000					
2 PIRLS 2001 – PISA 2000	0.070*	0.025	0.073*	0.030	21
3 PIRLS 2001 – PISA 2003	0.107***	0.021	0.109**	0.032	18
4 PIRLS 2001 – PISA 2006	0.051*	0.021	0.048	0.031	23
5 TIMSS Pop. A 2003 – PISA 2003					
6 TIMSS Pop. A 2003 – PISA 2006					
7 TIMSS Pop. A 2003 – PISA 2009					
8 PIRLS 2006 – PISA 2006	0.037	0.025	0.021	0.033	24
9 PIRLS 2006 – PISA 2009	0.003	0.026	-0.008	0.032	29
10 PIRLS 2006 – PISA 2012	0.010	0.021	0.002	0.025	26
11 TIMSS Pop. A 2007 – PISA 2006					
12 TIMSS Pop. A 2007 – PISA 2009					
13 TIMSS Pop. A 2007 – PISA 2012					
14 PIRLS 2011 – PISA 2012	0.026	0.018	0.023	0.020	34
15 TIMSS Pop. A 2011 – PISA 2012	0.025	0.022	0.023	0.025	34
16 PIRLS 2011 – PISA 2015	0.008	0.017	0.005	0.018	32
17 TIMSS Pop. A 2011 – PISA 2015	0.020	0.018	0.016	0.018	35
18 TIMSS Pop. A 2015 – PISA 2015	0.002	0.017	0.003	0.018	33
19 PIRLS 2016 – PISA 2015	0.002	0.018	0.000	0.018	33
Weighted mean effect (PIRLS/TIMSS Pop. A – PISA)	0.026***	0.006	0.018**	0.007	
20 TIMSS Pop. A 1995 – TIMSS Pop. B 1995					
21 TIMSS Pop. A 1995 – TIMSS Pop. B 1999					
22 PIRLS 2001 – TIMSS Pop. B 2003					
23 TIMSS Pop. A 2003 – TIMSS Pop. B 2003					
24 TIMSS Pop. A 2003 – TIMSS Pop. B 2007					
25 PIRLS 2006 – TIMSS Pop. B 2007					
26 PIRLS 2006 – TIMSS Pop. B 2011	0.012	0.047	0.014	0.051	24
27 TIMSS Pop. A 2007 – TIMSS Pop. B 2007					
28 TIMSS Pop. A 2007 – TIMSS Pop. B 2011					
29 TIMSS Pop. A 2011 – TIMSS Pop. B 2011	0.012	0.031	0.012	0.035	37
30 TIMSS Pop. A 2011 – TIMSS Pop. B 2015	0.041+	0.021	0.036	0.023	34
31 TIMSS Pop. A 2015 – TIMSS Pop. B 2015	0.071*	0.029	0.066+	0.034	35
32 PIRLS 2016 – TIMSS Pop. B 2015	0.051+	0.027	0.051	0.031	32
Weighted mean effect (PIRLS/TIMSS Pop. A – TIMSS Pop. B)	0.042**	0.012	0.039**	0.014	

Note. The estimate columns report the effects of between-school tracking on social segregation across schools at secondary school level. The estimates of the control variable social segregation across schools at primary school level are omitted. *N* = number of countries included in survey combination. Significance levels: *** $p < 0.001$; ** $p < 0.010$; * $p < 0.050$; + $p < 0.100$.

Table A.4. Overall Mean Number of Students, Schools, and Number of Students per School Divided by Survey Combinations

No.	Survey combinations		Mean number of students per country		Mean number of schools per country		Mean number of students per school per country	
	Primary school level data	Secondary school level data	Primary school	Secondary school	Primary school	Secondary school	Primary school	Secondary school
TIMSS Pop. A/PIRLS - PISA								
1	TIMSS Pop. A 1995	PISA 2000	3,556.1	5,730.5	160.5	213.7	25.1	31.3
2	PIRLS 2001	PISA 2000	4,012.8	5,655.6	162.0	210.5	29.4	31.5
3	PIRLS 2001	PISA 2003	4,173.9	6,530.1	164.9	256.3	30.3	30.9
4	PIRLS 2001	PISA 2006	4,032.7	6,348.4	160.4	249.1	30.1	30.3
5	TIMSS Pop. A 2003	PISA 2003	4,466.8	5,957.1	167.4	214.2	30.7	30.4
6	TIMSS Pop. A 2003	PISA 2006	3,947.7	7,070.2	160.6	256.0	29.2	30.2
7	TIMSS Pop. A 2003	PISA 2009	4,487.4	7,244.2	166.8	258.8	30.6	30.5
8	PIRLS 2006	PISA 2006	4,734.2	6,453.8	166.9	238.7	36.0	38.6
9	PIRLS 2006	PISA 2009	4,744.5	6,707.0	166.0	244.6	36.0	37.1
10	PIRLS 2006	PISA 2012	4,815.5	6,979.0	168.0	263.1	35.9	38.4
11	TIMSS Pop. A 2007	PISA 2006	4,448.1	6,291.0	164.1	232.9	34.9	32.0
12	TIMSS Pop. A 2007	PISA 2009	4,577.8	6,818.0	166.6	249.6	34.8	32.6
13	TIMSS Pop. A 2007	PISA 2012	4,597.4	6,985.2	167.6	282.1	34.5	32.5
14	TIMSS Pop. A 2011	PISA 2012	5,215.3	7,302.6	187.0	280.6	32.8	32.3
15	TIMSS Pop. A 2011	PISA 2015	5,204.9	6,842.1	185.9	247.2	33.3	36.5
16	PIRLS 2011	PISA 2012	5,907.2	7,965.1	218.7	314.8	33.2	31.8
17	PIRLS 2011	PISA 2015	5,753.5	7,300.6	211.9	266.0	34.0	36.1
18	TIMSS Pop. A 2015	PISA 2015	5,746.4	7,318.7	204.8	267.8	33.6	35.0
19	PIRLS 2016	PISA 2015	5,812.2	7,208.2	216.4	266.3	33.9	36.2
TIMSS Pop. A/PIRLS – TIMSS Pop. B								
20	TIMSS Pop. A 1995	TIMSS Pop. B 1995	3,643.5	3,634.2	158.3	139.1	25.4	28.1
21	TIMSS Pop. A 1995	TIMSS Pop. B 1999	3,979.1	4,536.7	161.8	154.7	27.1	31.4
22	PIRLS 2001	TIMSS Pop. B 2003	3,843.2	4,273.7	153.0	153.6	28.7	31.9
23	TIMSS Pop. A 2003	TIMSS Pop. B 2003	4,466.4	4,584.6	162.9	153.1	31.7	33.8
24	TIMSS Pop. A 2003	TIMSS Pop. B 2007	4,590.3	4,249.2	165.4	157.9	31.9	31.9
25	PIRLS 2006	TIMSS Pop. B 2007	4,469.4	4,332.9	158.7	154.1	34.0	35.7

No.	Survey combinations		Mean number of students per country		Mean number of schools per country		Mean number of students per school per country	
	Primary school level data	Secondary school level data	Primary school	Secondary school	Primary school	Secondary school	Primary school	Secondary school
26	PIRLS 2006	TIMSS Pop. B 2011	4,603.7	5,294.5	164.1	177.2	33.9	36.5
27	TIMSS Pop. A 2007	TIMSS Pop. B 2007	4,285.2	4,327.1	156.9	150.4	34.6	36.0
28	TIMSS Pop. A 2007	TIMSS Pop. B 2011	4,306.0	5,306.4	161.4	179.7	34.2	36.0
29	TIMSS Pop. A 2011	TIMSS Pop. B 2011	5,463.4	5,695.3	191.7	187.7	33.2	36.3
30	TIMSS Pop. A 2011	TIMSS Pop. B 2015	5,568.4	6,054.6	193.0	181.9	34.1	40.2
31	TIMSS Pop. A 2015	TIMSS Pop. B 2015	5,670.7	6,059.2	202.3	185.8	33.3	37.9
32	PIRLS 2016	TIMSS Pop. B 2015	5,825.2	6,115.8	216.5	185.1	33.7	40.0

Article 5. The Effects of Early Between-School Tracking on Gender Segregation and Gender Gaps in Achievement: A Differences-in-Differences Study

Isa Steinmann, Andrés Strello, Rolf Strietholt

The version of the manuscript printed below is the preprint of the article published in *Research in School Effectiveness and School Improvement*, volume 34(2) (2023).
https://doi.org/10.1080/09243453.2023.2165510

Abstract

We investigate effects of tracking students into higher, more academic and lower, less academic school types immediately after primary school (early tracking) instead of having a comprehensive secondary school system (late tracking) on school gender segregation and gender gaps in achievement outcomes. We assume that, in early tracking countries, girls are more frequently selected into more academic school types, which leads to more school segregation by gender and achievement advantages of girls over boys. In a differences-in-differences design, we compare secondary-school level gender inequalities between early and late tracking countries, after controlling for primary-school level differences. We investigate n = 787 country-by-year observations in 33 matches of primary- and secondary-school level datasets from three international large-scale assessments. As expected, we find that early tracking increases the degree of school gender segregation. Not conforming to expectations, the evidence does not indicate that tracking has effects on gender gaps in achievement.

Across the world, boys and girls differ in a variety of educational areas, such as student achievement, school-related attitudes and school-related behaviours (e.g. Mullis et al., 2017, 2020; OECD, 2015). Since such educational gender inequalities likely play a role in educational transitions, we will investigate the effects of early between-school tracking on school gender segregation and gender gaps in student achievement.

Between-school tracking means dividing students into different school types instead of having one comprehensive school type for all. The decisions to allocate students to higher, more academic, or lower, less academic school types are usually based on teachers' evaluations of students' scholastic aptitudes (i.e. school marks[9]) and parents' preferences. While all educational systems apply between-school tracking at some point in students' school careers, some undertake it earlier (e.g. after Grade 4, around age 10) and others later (e.g. after Grade 10, around age 16). Track placement has far-reaching long-term consequences for students' education and labour market outcomes (Borghans et al., 2019; Dockx & De Fraine, 2019; Luyten et al., 2003).

[9] We use the term "school mark" instead of "school grade" to avoid confusion with grade levels.

Proponents of early between-school tracking argue that tracking helps to tailor learning environments to the students' needs and leads to optimal learning outcomes for all. The idea is that curricula, learning materials and didactics can be adjusted to the students' differential aptitudes, allowing teachers to challenge high-achieving students without overwhelming low-achieving ones, for example. Critics of between-school tracking, by contrast, fear that the different school types do not promote learning equally due to differences in learning resources, teacher quality, school climate and educational expectations, etc. They argue that in early tracking countries, low-achieving students might be left behind more than in comprehensive systems (Dockx et al., 2019; Maaz et al., 2008; Retelsdorf et al., 2012). Another central criticism is that tracking decisions do not just depend on students' scholastic aptitudes but also on their membership of social groups, such as gender or socioeconomic status groups (Batruch et al., 2019; Maaz et al., 2008; Timmermans et al., 2015). If social groups differ in their school track allocation and if school tracks foster learning to a different degree, between-school tracking might increase educational inequalities between those groups (Strello et al., 2021, 2022; Contini & Cugnata, 2020; Lavrijsen & Nicaise, 2015).

In conclusion, early tracking may lead to greater between-school segregation of social groups at secondary school level than late tracking (e.g. Strello et al., 2022). This conflicts with the aim of compulsory primary and secondary schooling, namely to include all students regardless of their social background and ideally reduce social divides (cf. Strello et al., 2022a; Reichelt et al., 2019). Further, if higher, more academic, and lower, less academic school tracks differ in how effectively they promote learning, achievement gaps between the social groups should increase more in early than in late tracking countries (Strello et al., 2021; Lavrijsen & Nicaise, 2015; Scheeren & Bol, 2021). Such increases in social inequalities in education would be another problematic effect of early tracking. In contrast to previous research, which has often focused on socioeconomic status as a social category (e.g. Strello et al., 2021, 2022; Lavrijsen & Nicaise, 2015), this study pertains to gender. We assume that gender is relevant for the decision on which school type students choose after primary school because previous evidence suggests gender differences in achievement levels and school marks (Rosén et al., 2022; OECD, 2015), and a higher chance of girls to attend academic tracks than boys (Bacher, 2009; Caro et al., 2009; Jürges & Schneider, 2011; Klapproth et al., 2013; Róbert, 1991; Timmermans et al., 2015).

Literature Review

We will review the literature with three questions in mind: Is there evidence (1) on gender differences in early tracking decisions, (2) on the effects of early tracking on gender segregation and (3) on the effects of early tracking on gender gaps in student achievement?

Gender Gaps in Early Tracking Decisions

Since early tracking decisions are typically predicated on teacher evaluations of students' scholastic aptitudes, we will first review literature on primary-school level gender gaps in school achievement. The international large-scale assessments PIRLS (Progress in International Reading

Literacy Study) and TIMSS (Trends in International Mathematics and Science Study) assess Grade 4 students' achievement in standardised tests. In the latest cycle of PIRLS, girls outperformed boys in reading in most countries or benchmarking participants,[10] while there were no significant gaps in others (Mullis et al., 2017). According to TIMSS 2019, countries' gender gaps in mathematics and science varied – in some cases, boys had mean advantages, while in others, girls had mean advantages (Mullis et al., 2020). Thus, the patterns differed vastly between domains and countries.

In contrast to the heterogeneous findings on achievement gender gaps, findings on teacher-awarded school marks usually point to advantages for girls across domains and countries that are more pronounced in languages than in mathematical subjects (O'Dea et al., 2018; Voyer & Voyer, 2014). The only international large-scale study to assess school marks was PISA (Programme for International Student Assessment) in 2000. In reading, girls reported higher mean marks in all countries, even after controlling for score differences in the reading test. In mathematics, all but one country showed that girls had either better average marks or no significant gaps (OECD, 2015). There are multiple theories on why girls enjoy more consistent advantages in their school marks than in achievement scores in standardised tests (Hadjar & Buchmann, 2016; Kenney-Benson et al., 2006). One prominent explanation is that alongside academic achievement, school marks also capture behavioural components, in which girls outperform boys (Bowers, 2011; Geven et al., 2017; Kenney-Benson et al., 2006).

Besides outperforming boys in their school marks, girls receive more favourable school track recommendations than boys in Germany (Caro et al., 2009; Jürges & Schneider, 2011) as well as Luxembourg (Klapproth et al., 2013). In the Netherlands, teachers have lower academic expectations for primary school boys (Timmermans et al., 2015). In Austria, girls more frequently report that they will transition to a higher track at the end of primary school than boys (Bacher, 2009). In Hungary, girls are more likely to attend academic secondary education than boys (Róbert, 1991).

Together, the evidence points to average female advantages in early tracking decisions.

Early Tracking and the Between-School Gender Segregation

If tracking decisions differ for boys and girls, early tracking should increase between-school gender segregation at the secondary school level, that is, the degree to which boys and girls attend separate schools. Kriesi and Imdorf (2019) have argued that it is "likely that a high level of educational differentiation increases gender segregation even more in educational systems with early tracking in adolescence" (p. 204). However, to the best of our knowledge, there is no previous empirical study that has tested this hypothesis with empirical data (cf. Kriesi & Imdorf, 2019). Imdorf et al. (2015) found that vocational programmes are more gender-segregated than academic programmes, and that Germany showed a higher educational gender segregation than Norway and Canada. This could potentially be explained by the early tracking system in Germany but the study

[10] In the following, the term "country" is short for both country and benchmarking participant.

design does not allow to empirically test this hypothesis. Wiseman (2008) categorised countries according to the degree of between-school gender segregation at the secondary school level but did not investigate their tracking status. Indeed, between-school gender segregation has only been investigated in relation to single-sex schooling, an education policy aiming at absolute between-school gender segregation (cf. Robinson et al., 2021) that operates independently of tracking policies.

Thus, we found no empirical literature on the association between early tracking and school gender segregation.

Early Tracking and Gender Gaps in Achievement at the Secondary School Level

Several studies have investigated associations between tracking policies and gender gaps in achievement using cross-sectional correlational designs. These have reported heterogeneous results (Bedard & Cho, 2010; Bodovski et al., 2020; Marks, 2008; van Hek et al., 2019; van Langen et al., 2006). However, such studies cannot disentangle between-country differences in prior achievement gender gaps from differences due to tracking. The countries could have differed in gender gaps before the streaming of students into different ability tracks due to other institutional features such as prosperity or the role of women in the respective societies. In other words, the results of cross-sectional studies are difficult to interpret due to potential selection bias.

We are aware of only two studies (Hermann & Kopasz, 2019; Scheeren & Bol, 2021) that have applied a more robust design that could identify the effect of tracking on gender gaps in student achievement. Following Hanushek and Wößmann's (2006) seminal study on the effects of tracking on educational inequalities, Hermann and Kopasz (2019) combined primary school level data from PIRLS 2006 and TIMSS 2007 (population A) with secondary school level data from PISA 2012 to compare early and late tracking countries at the secondary school level, while controlling for differences at the primary school level (i.e. before tracking took place). Hence, the study matched primary- and secondary-school level data for roughly 30 countries. It investigated the associations between achievement scores, gender, tracking status and the assessment level (primary or secondary school level) in multilevel models with two- and three-way interaction terms. These models were run separately for reading, mathematics and science outcomes and included further control variables. Their findings suggest that, between primary and secondary school, gender gaps shifted more to the advantage of girls in early than in late tracking countries (Hermann & Kopasz, 2019). Scheeren and Bol (2021) ran a similar model with reading and mathematics as outcomes. They found that early tracking shifted achievement gaps to the advantage of girls in reading, but the effects were not statistically significant for mathematics. They included data from 9 early tracking countries and 12 late tracking ones in overall 26 matches of primary (PIRLS cycles 2001–2011 and TIMSS population A cycles 1995–2015) and secondary school level data (TIMSS cycles population B 1995–2015 and PISA cycles 2000–2015).

The literature therefore suggests that early tracking can – in line with hypotheses – lead to performance advantages among girls. However, unlike the seminal paper by Hanushek and

Wößmann (2006), which used country-level models, the two prior differences-in-differences studies (Hermann & Kopasz, 2019; Scheeren & Bol, 2021) used individual-level models. This approach leads to biased results, since the primary- and secondary-school level achievement tests used in PIRLS, TIMSS populations A and B and PISA are scaled independently, meaning that the achievement scores lie on separate scales (Contini & Cugnata, 2020).[11]

The Present Study

The present study addresses two gaps in the tracking literature, the lack of studies on early tracking effects on school gender segregation and the lack of robust country-level differences-in-differences studies on early tracking effects on achievement gender gaps. Specifically, we investigate two research questions:

(1) Does early tracking increase between-school segregation among boys and girls at the secondary school level?
(2) Does early tracking shift achievement gender gaps so that girls gain a relative advantage at the secondary school level?

We hypothesised that early tracking would increase between-school gender segregation and shift achievement gender gaps so that girls gain relative advantages at the secondary school level. This is because girls can be expected to be more often selected into higher secondary school tracks. Since higher tracks can be assumed to be more favourable to learning than lower ones, early tracking should shift achievement gender gaps to the advantage of girls. To test our hypotheses, we applied the country-level differences-in-differences design as in Hanushek and Wößmann (2006) and combined all available PIRLS, TIMSS and PISA data.

Materials and Methods

Sample

We combined international large-scale assessment data that is representative for the countries' student populations. At the primary school level, we used data from PIRLS and TIMSS (population A), which target Grade 4 student populations. At secondary school level, we used data from PISA, which samples 15-year-old students, and from TIMSS (population B), which targets Grade 8 students. The PIRLS assessment has been repeated every 5 years since 2001 and PISA has been repeated every three years since 2000. Both TIMSS assessments have been repeated every 4

[11] For the estimation of individual-level models, data from different studies are pooled and the achievement scores are treated as if they had the same metric. This is generally not the case. For example, the PIRLS and PISA achievement scales were each transformed to an international mean of 500 with a standard deviation of 100. This does not imply that the actual performance of primary and secondary students is on average the same but is merely the result of the (arbitrary) transformation of the achievement scales. The two-step approach on country-level models circumvents this issue as the secondary school measures are on the right-hand side of the equation and the primary school measures are on the left-hand side (see Contini & Cugnata, 2020 for further detail).

years since 1995, except in 1999, where only population B was assessed. The studies applied multi-stage stratified sampling approaches; they first sampled schools and then, within schools, students based on their grade (PIRLS and TIMSS) or age (PISA). The datasets are available online along with extensive technical background information (OECD, 2021; TIMSS & PIRLS International Study Center, 2019a, 2019b).

Our analytical approach explained below, the country-level differences-in-differences design as in Hanushek and Wößmann (2006), rests on the assumption that, except for tracking, no other policies change between the early and late tracking countries between the primary and secondary school level. In the present case of gender-related effects, the most plausible other policy type that might affect gender segregation and gender gaps in achievement concerns single-sex schooling. If early tracking countries were more likely to have single-sex secondary schools than late tracking countries, or vice versa, this would have distorted analyses of the tracking effect. We therefore excluded countries that had large shares (more than 25%) of single-sex schools (i.e. schools with only girls or only boys)[12] in the samples because such single-sex schooling practices can lead to distorted estimates of tracking effects.

We matched data from primary and secondary school assessments that were conducted at approximately the same time or later, meaning we could consider both possible cohort and period effects (Blanchard et al., 1977). For instance, we combined data from PIRLS 2001 with data from PISA 2000 (same period) as well as with data from PISA 2006 (same cohort). To maximise the power of our analyses, we used all available PIRLS, PISA and TIMSS cycles. Figure 1 displays the 33 ways of matching primary and secondary school data for the three types of study combinations (PIRLS → PISA, TIMSS → PISA and TIMSS → TIMSS). For each match, we included all countries that participated in both the primary and secondary school assessments. If countries participated with more than one target grade (e.g. Norway in PIRLS 2006), we only included the main target observation. We treated benchmarking participants (e.g. the Canadian provinces Quebec and Alberta) as separate entities.

Figure 1 shows the number of countries per match in parentheses. Our 33 study matches generated a total sample of $n = 787$ country-by-year observations. These observations stemmed from 72 countries that repeatedly participated in the studies. Overall, the included data covers information from more than four million students.

[12] Excluded late tracking countries with large shares of single-sex schools were Bahrain, Iran, Kuwait, Malta, New Zealand, Oman, Qatar, Saudi Arabia and the United Arab Emirates. Excluded early tracking countries were Ireland, the Republic of Korea and Trinidad and Tobago.

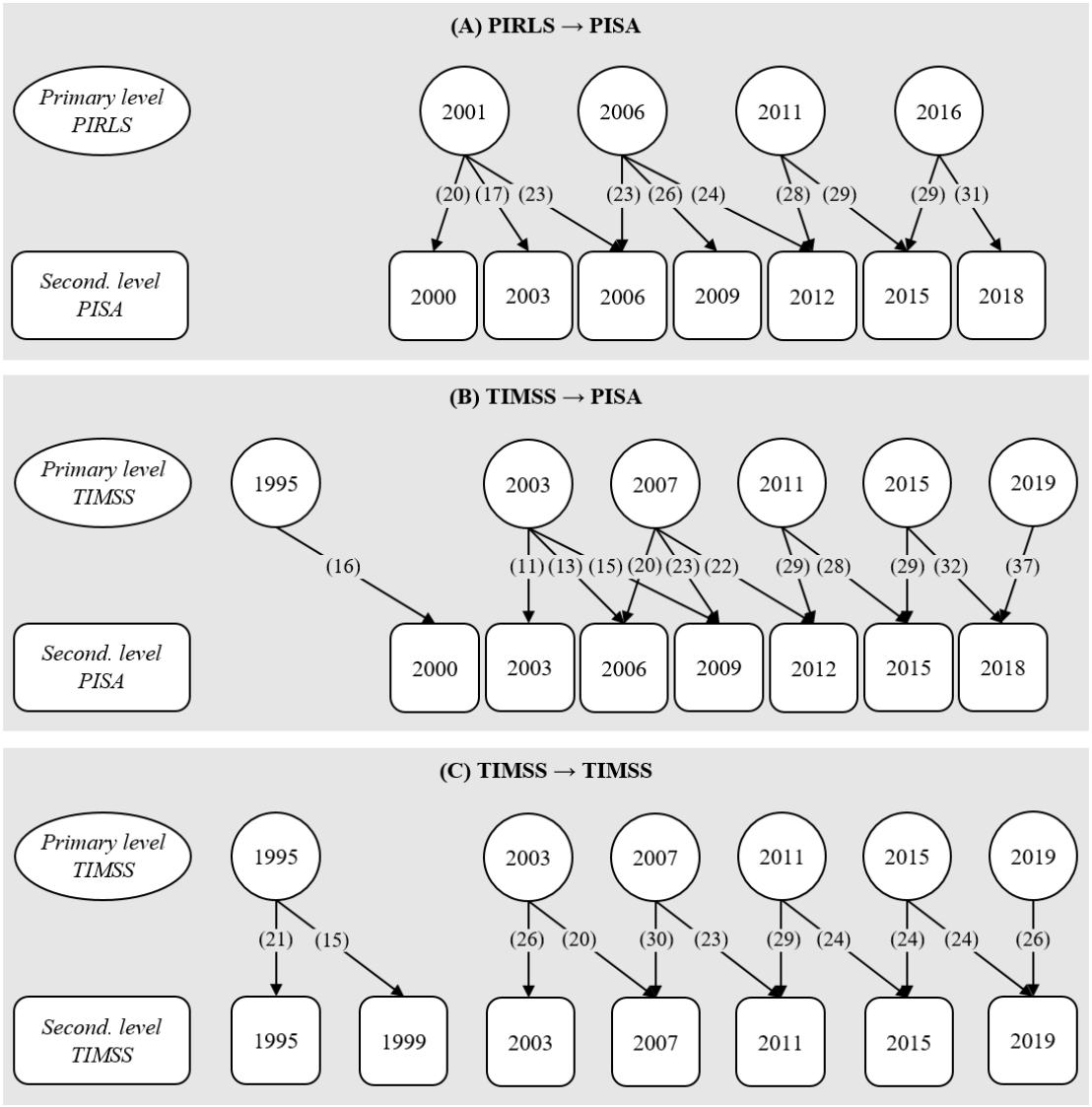


Figure 1. Overview of 33 study matches

Note. Every arrow reflects one match of a primary and secondary school assessment cycle. The number of countries per match are displayed in parentheses.

Instruments

School Gender Segregation

As a measure of school gender segregation, we estimated the dissimilarity index D (Duncan & Duncan, 1955) for all country-by-year observations c:

$$D_c = 0.5 \sum_{i=1}^{k_c} \left| \frac{f_{ic}}{F_c} - \frac{m_{ic}}{M_c} \right| \quad (1)$$

In equation 1, F_c is the number of female students and M_c is the number of male students. In each school i of a country-by-year observation c, f_{ic} represents the number of girls and m_{ic} the number of boys in the sample. The sample contains k_c schools. Thus, D_c reflects the extent to which schools are segregated by gender. A value of 0 indicates the same proportion of girls in all schools

(i.e. no gender segregation) and a value of 1 indicates that no school includes both boys and girls (i.e. total gender segregation).

We applied student sampling weights when estimating D_c to account for the stratified sampling approaches. We excluded students for whom there was no gender information from the estimation of D_c. The gender variables contained between 0% and 8% of missing data across all country-by-year observations.

Gender Gaps in Student Achievement

We estimated the gender gaps in student achievement in reading, mathematics and science separately as standardised mean differences for every country-by-year observation c:

$$G_c = \frac{Y_{fc} - Y_{mc}}{0.5\ (SD_{fc} + SD_{mc})} \tag{2}$$

Y_{fc} and SD_{fc} represent the average achievement score and standard deviation for female students and Y_{mc} and SD_{mc} for male students. By implication, a G_c value of 0 means that boys and girls had equal scores. A value of 1 implies that girls outperformed boys by 1 standard deviation (of the country-by-year observation) and a value of -1 that boys outperformed girls. Thus, G_c can be interpreted like the effect size measure Cohen's *d*.

The large-scale assessments provide multiple plausible values for achievement scores in reading, mathematics and science, estimated based on students' responses to test items and background information using item-response-theory-based conditioning techniques (von Davier et al., 2009). We used all available plausible values and applied Rubin's (1987) rules to estimate the means and standard deviations of the achievement scores of boys and girls in the country-by-year observations. We applied student sampling weights so that we could generalise our findings to the underlying student populations and excluded students with missing gender information when estimating G_c.

Early Tracking Indicator

We determined the target student age and grade at which comprehensive education ended and between-school tracking began using information from different reports (Eurydice, 2015; OECD, 2004, 2006; UNESCO-IBE, 2012). Based on this information, we created two dummy variables that indicated whether a country tracked students before Grade 8 for matches with TIMSS (population B) data or before the age of 15 for matches with PISA data (0 = *late tracking*, 1 = *early tracking*). In some countries, tracking takes place between grade 8 (TIMSS) and when students are 15 years old (PISA); to account for this, these countries are considered late tracking countries in analyses with TIMSS and early tracking countries in analyses with PISA. Benchmarking participants were assigned the same values as the countries they belonged to. Table 1 depicts the 62 unique countries (i.e., without displaying benchmarking participants separately) in our study and whether they were considered early or late tracking countries in study matches with TIMSS and PISA data.

Table 1. Overview of early and late tracking countries

	Tracking					Tracking			
Country name	Grade	Early (TIMSS)	Age	Early (PISA)	Country name	Grade	Early (TIMSS)	Age	Early (PISA)
Albania	9		15		Japan	9		15	
Algeria	9		15		Kazakhstan	9		15	
Argentina	6	X	12	X	Latvia	9		16	
Armenia	9		15		Lithuania	8		15	
Australia	10		16		Luxembourg	6	X	12	X
Austria	4	X	10	X	Macau	9		14	X
Belgium	6	X	12	X	Moldova	10		15	
Bosnia & Herz.	9		14	X	Mongolia	8		16	
Botswana	10		16		Montenegro	9		14	X
Bulgaria	7	X	14	X	Morocco	9		15	
Canada	12		18		Netherlands	6	X	12	X
Chile	10		16		North Maced.	8		15	
Chinese Taipei	9		15		Norway	10		16	
Colombia	9		15		Philippines	10		16	
Croatia	8		15		Poland	9		15	
Cyprus	9		15		Portugal	9		15	
Czech Republic	5	X	11	X	Romania	8		14	X
Denmark	10		16		Russ. Fed.	9		15	
El Salvador	9		16		Serbia	8		15	
Finland	9		16		Singapore	6	X	12	X
France	9		15		Slovak Rep.	4	X	10	X
Georgia	9		15		Slovenia	9		15	
Germany	4	X	10	X	South Africa	9		15	
Greece	9		15		Spain	9		15	
Honduras	6	X	12	X	Sweden	9		16	
Hong Kong	11		16		Thailand	9		15	
Hungary	4	X	10	X	Tunisia	10		16	
Iceland	10		16		Turkey	8		14	X
Indonesia	9		16		Ukraine	9		15	
Israel	10		15		Un. Kingdom	11		16	
Italy	8		14	X	United States	12		18	

Note. Tracking grade reflects the grade after which between-school tracking takes place. Tracking age reflects the typical age in the tracking grade. Early tracking countries are countries with a tracking grade below 8 for matches with TIMSS data and tracking age below 15 for matches with PISA data. Benchmarking participants (e.g. the Canadian provinces Quebec and Alberta) are not displayed separately.

Differences-in-Differences Estimations

To estimate the effect of between-school tracking on school gender segregation and gender gaps in student achievement, we applied differences-in-differences analyses. The basic idea of differences-in-differences analyses is to conduct longitudinal analyses on the country level by comparing secondary school data between early and late tracking countries, while controlling for

primary school data. It is important to note that at the primary school level, all countries have a comprehensive system.

Since the international large-scale assessments have different, independent achievement scales, it is not legitimate to pool primary school data from PIRLS or TIMSS (population A) with secondary school data from PISA or TIMSS (population B) at the student level (Contini & Cugnata, 2020). Unlike previous studies (Hermann & Kopasz, 2019; Scheeren & Bol, 2021), we therefore applied a country-level regression approach, as proposed by Hanushek and Wößmann (2006). We first estimated within-country gender segregation and achievement gap measures (see equations 1 and 2) and then ran country-level regression analyses.

Tracking and School Gender Segregation

The first research question concerned the tracking effect on school gender segregation. The differences-in-differences model for these analyses was:

$$D_{cqs} = \alpha_q + \beta_q T_{cq} + \gamma_q D_{cqp} + \varepsilon_{cq} \quad (3)$$

Here, D_{cqs} reflects the gender segregation measure of a country-by-year observation c in the study match q at the secondary school level s. T_{cq} depicts the binary tracking status (0 = *late tracking*, 1 = *early tracking*). D_{cqp} depicts the gender segregation measure at primary school level p. We regressed the secondary school level segregation measure on the tracking indicator while controlling for the primary school level segregation measure. We repeated the regression analyses for all q = 33 study matches (see Figure 1).

For each study match q, the intercept α_q reflects the estimated school gender segregation at the secondary school level for late tracking countries, if there is no gender segregation at the primary school level. The regression coefficient β_q is the central parameter of interest, as it reflects the differences in school gender segregation between early and late tracking countries at the secondary school level, after controlling for differences at the primary school level. The coefficient γ_q reflects the association between the segregation measures at the primary and secondary school levels.

Tracking and Gender Gaps in Student Achievement

The second research question concerned tracking effects on gender gaps in student achievement. The estimated model was:

$$G_{cqs} = \alpha_q + \beta_q T_{cq} + \gamma_q G_{cqp} + \varepsilon_{cq} \quad (4)$$

As in equation 3, we regressed the average secondary school level gender gaps in student achievement G_{cqs} of a country-by-year observation c on a binary tracking indicator T_{cq}, while controlling for the primary school level gender gap G_{cqp} for every study match q. For the differences-in-differences analyses of gender gaps in reading achievement, we replicated the regression analyses for the q = 10 PIRLS → PISA study matches (see Panel A in Figure 1). To assess gender gaps in

mathematics and science, we replicated the regressions for the q = 12 TIMSS → PISA matches (see Panel B in Figure 1) as well as the q = 11 TIMSS → TIMSS matches (see Panel C in Figure 1).

Here, too, the regression coefficient β_q is the central parameter of interest, as it depicts the difference that early tracking makes in secondary school achievement gender gaps compared to late tracking and after controlling for differences at the primary school level.

Summarising the Effect Estimates

Due to the replications across study matches, the differences-in-differences analyses resulted in q = 33 sets of regression coefficients r for school gender segregation, q = 10 sets of coefficients for gender gaps in reading achievement and q = 23 sets of coefficients for gender gaps in mathematics and science achievement. To synthesise these coefficients per outcome, we computed weighted means of the regression coefficients α_q, β_q and γ_q over the q replications. We applied the formulas that Card (2012) developed for meta-analyses. For each outcome, regression coefficient r and replication q, we first estimated the inverse standard error as a weight for the respective effect estimate:

$$w_{rq} = \frac{1}{SE_{rq}^2} \tag{5}$$

Therefore, w_{rq} reflects that the regression coefficients were not estimated with the same precision in the different study matches (e.g. due to different numbers of countries; see Figure 1). We then estimated weighted mean effects and their standard errors for each outcome:

$$\bar{r} = \frac{\sum(w_{rq} r_q)}{\sum w_{rq}} \tag{6}$$

$$SE_{\bar{r}} = \sqrt{\frac{1}{\sum w_{rq}}} \tag{7}$$

Since $\bar{r}$ and $SE_{\bar{r}}$ followed a normal distribution, we were able to test whether the mean effects differed significantly from zero (Card, 2012). We computed the synthesised regression coefficients $\bar{\alpha}$, $\bar{\beta}$ and $\bar{\gamma}$ separately for the four outcomes: school gender segregation, reading, mathematics and science gender gaps.

Results

Descriptive Findings

Table 2 provides descriptive statistics for all four outcome variables. Inferential results follow in the next section.

Table 2. Descriptive statistics of the four outcome variables at primary and secondary school level in all countries and divided by tracking status

		Primary school level				Secondary school level			
Outcome variable Country sample	n	M	SD	Min	Max	M	SD	Min	Max
School gender segregation									
All countries	787	0.17	0.03	0.07	0.32	0.24	0.10	0.05	0.59
Early tracking countries	203	0.18	0.03	0.10	0.32	0.31	0.10	0.13	0.59
Late tracking countries	584	0.17	0.03	0.07	0.25	0.21	0.09	0.05	0.51
Reading gender gap									
All countries	250	0.20	0.08	-0.02	0.38	0.38	0.12	0.13	0.72
Early tracking countries	86	0.16	0.06	0.01	0.30	0.35	0.10	0.13	0.62
Late tracking countries	164	0.22	0.08	-0.02	0.38	0.40	0.13	0.14	0.72
Mathematics gender gap									
All countries	537	-0.05	0.09	-0.28	0.32	-0.05	0.10	-0.44	0.23
Early tracking countries	117	-0.08	0.08	-0.28	0.10	-0.09	0.09	-0.31	0.16
Late tracking countries	420	-0.05	0.09	-0.28	0.32	-0.04	0.10	-0.44	0.23
Science gender gap									
All countries	537	-0.03	0.09	-0.26	0.22	-0.03	0.12	-0.46	0.27
Early tracking countries	117	-0.09	0.08	-0.26	0.09	-0.06	0.11	-0.42	0.16
Late tracking countries	420	-0.01	0.08	-0.18	0.22	-0.02	0.12	-0.46	0.27

Note. The school gender segregation variable can have values between 0 (absence of segregation) and 1 (total segregation). The achievement gender gap variables are standardised mean differences with positive values indicating that girls have higher achievement scores and negative values indicating that boys have higher achievement scores. Column *n* contains the number of country-by-year observations.

School Gender Segregation

At the primary school level, early and late tracking countries exhibited roughly the same, low degrees of school gender segregation, as measured by the dissimilarity index (see Table 2). Note that by chance, we would not expect boys and girls to be exactly equally distributed in all schools in all samples. Interestingly, there were higher degrees of school gender segregation at secondary school level in both types of countries. Overall, the school gender segregation ranged considerably between countries, especially at the secondary school level. When interpreting these descriptive statistics, it is important to note that we removed the countries with pronounced single-sex schooling practices from our analyses. If they had been included, we would have expected higher means and standard deviations.

Gender Gaps in Student Achievement

Positive average reading gender gaps are evident in both early and late tracking countries at the primary and secondary school levels, implying that girls score higher than boys. At the primary school level, the reading gender gaps varied – in some countries, boys had slight average advantages, while in others, girls had moderate average advantages. At the secondary school level, the reading gender gaps indicated that girls had small to large average advantages. The female reading advantage was more pronounced in late tracking countries than in early tracking countries.

In terms of mathematics gender gaps, boys had small average advantages over girls at the primary and secondary school level in both early and late tracking countries. The mathematics gender gaps varied – in some cases, boys had moderate average advantages while in other cases, girls showed

such moderate average advantages. The average male advantage was more pronounced in early tracking countries than in late tracking ones.

In science, the average primary and secondary school level gender gaps pointed to small advantages for boys. Once again, these gender gaps varied – with boys having moderate average advantages in some cases and girls having moderate average advantages in others. In early tracking countries, boys' average advantages were more pronounced than in late tracking countries.

Note that the primary- and secondary-school level gender gaps in student achievement in Table 2 are not directly comparable because PIRLS, PISA and TIMSS populations A and B have independent achievement scales.

Differences-in-Differences Results

Findings per Study Match

In the differences-in-differences models in equations 3 and 4, we regressed the secondary-school-level outcome measures on a binary tracking indicator (0 = *late tracking*, 1 = *early tracking*) and a primary school measure of the outcomes. We replicated the analyses for all study matches and four outcomes. The coefficients of the tracking indicators show the difference that early tracking makes compared to late tracking at the secondary school level, after accounting for differences at the primary school level.

Figure 2 depicts the estimated tracking effects as well as their confidence intervals per study match and outcome. The figure shows that the individual analyses relied on a small number of country observations, as seen in the wide confidence intervals. However, while individual estimates were subject to pronounced uncertainty, clear trends emerged when we considered the estimates together. Regarding the early tracking effect on school gender segregation (see panel A in Figure 2), the estimated differences-in-differences parameters were positive in almost all study match replications. About half of the parameters were significantly different from zero. Concerning the early tracking effects on reading, mathematics and science gender gaps (see panels B–D in Figure 2), with one exception, none of the tracking effects were significantly different from zero in the single study matches. The estimated parameters were mostly negative in the case of the mathematics and science gender gap outcomes and more mixed in the case of the reading gender gap outcome.

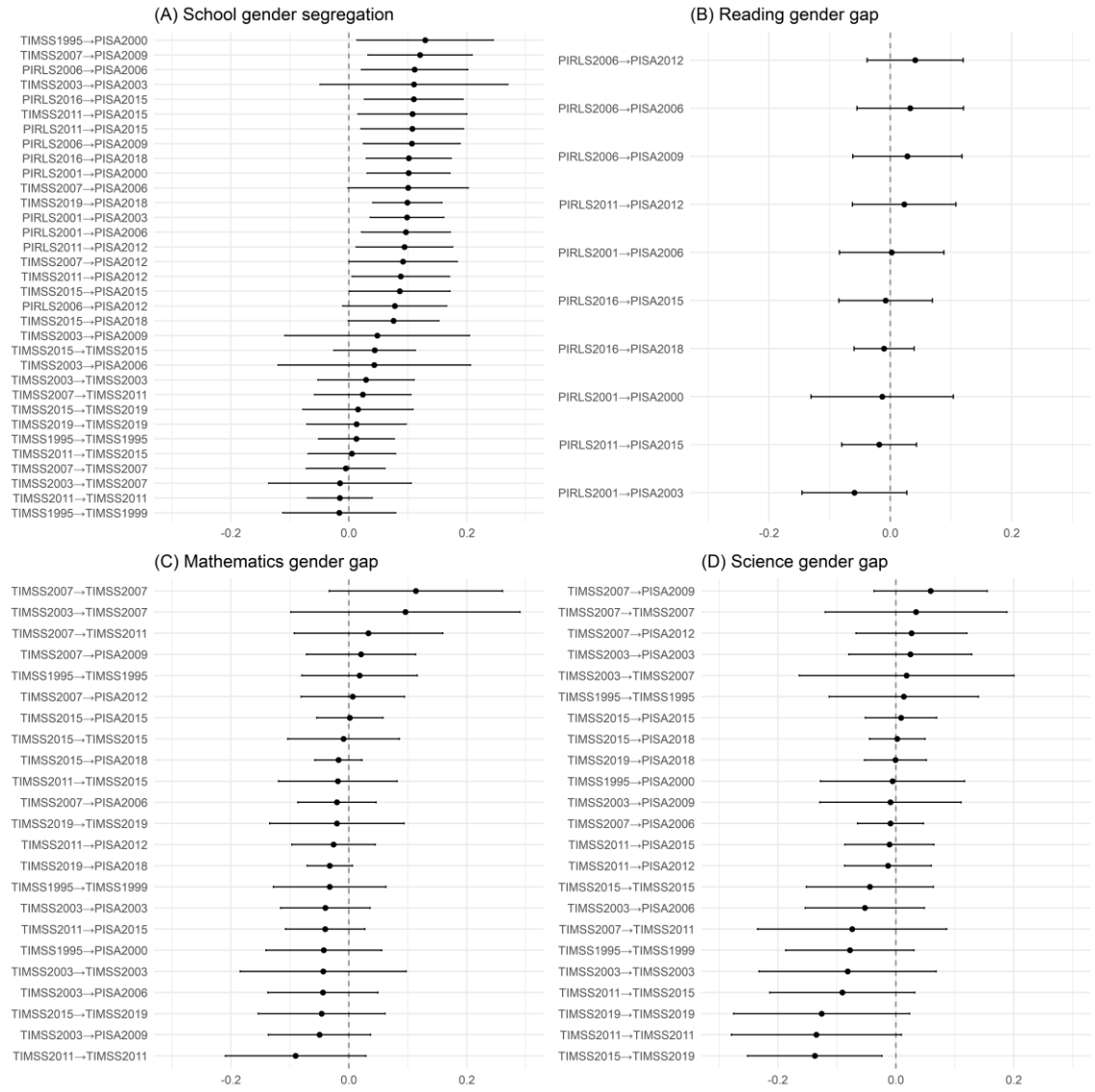


Figure 2. Overview of tracking effects per study match and outcome

Note. The regression coefficients β_q and their confidence intervals per study match q (rows) are displayed here. These coefficients reflect the central effects of interest, as they depict the difference that early tracking makes at the secondary school level compared to late tracking and after controlling for differences at the primary school level. The panels A–D display the results per outcome variable. Positive effects on gender segregation imply more school gender segregation in early than in late tracking countries. Positive effects on achievement gender gaps imply shifting achievement gender gaps to the relative advantage of girls in early compared to late tracking countries.

Summarised Findings across Study Matches

Due to the small number of country observations per study match, we summarised the differences-in-differences regression results across study matches using Card's (2012) formulas (see equations 5, 6 and 7). Table 3 depicts the synthesised intercepts $\bar{\alpha}$, tracking coefficients $\bar{\beta}$ and primary level coefficients $\bar{\gamma}$.

As Table 3 shows, early tracking countries had significantly higher degrees of school gender segregation at the secondary school level than late tracking countries, when controlling for segregation differences at the primary school level. The early tracking effect amounted to a dissimilarity index value of 0.063, implying that 6.3% of boys and girls would have to change schools to obtain the same degree of segregation in early and late tracking countries. This is a strong tracking effect, considering the low country-level variability of the dissimilarity index (SD = 0.10 on secondary school level, see Table 2).

The early tracking effects on gender gaps in student achievement were very close to zero and only significant in the case of mathematics, not reading and science achievement (see Table 3). The statistically significant effect on mathematics gender gaps amounted to -0.021 standard deviations, implying that early tracking gave boys a very small advantage over girls compared to late tracking countries, and given gender gap differences at the primary school level.

Furthermore, the primary- and secondary-school level measures were positively associated for all four outcomes (see coefficients $\bar{\gamma}$ in Table 3). Countries with a higher primary school level degree of school gender segregation and more positive gender gaps in student achievement showed higher secondary school level degrees of school gender segregation and more positive gender gaps in student achievement.

Table 3. Synthesised differences-in-differences regression results per outcome

Outcome variable	Number of study matches q	Number of country-by-year obs. c	Intercept $\bar{\alpha}$	Tracking coefficient $\bar{\beta}$	Primary level coefficient $\bar{\gamma}$
School gender segregation	33	787	0.041* (0.015)	0.063* (0.007)	0.968* (0.088)
Reading gender gap	10	250	0.218* (0.017)	-0.001 (0.012)	0.770* (0.075)
Mathematics gender gap	23	537	-0.026* (0.004)	-0.021* (0.008)	0.441* (0.039)
Science gender gap	23	537	0.006 (0.004)	-0.014 (0.009)	0.590* (0.048)

Note. The results display here are synthesised results of the differences-in-differences regressions with standard errors in parentheses. The rows display the results per outcome variable. The intercept reflects secondary school level outcome values in late tracking countries, after controlling for primary school level outcome values. The tracking coefficient reflects differences between early and late tracking countries at the secondary school level, after controlling for differences at primary school level. The primary level coefficient reflects the association between outcome values at primary and secondary school level. * $p < .050$.

Findings of Further Analyses

As discussed above, it was crucial for our study to ensure that the treatment (i.e. the tracking status) was not confounded with other potentially relevant treatments such as single-sex schooling policies. Thus, we excluded countries with large shares of single-sex schools from the main analyses. However, we furthermore tested whether the included early and late tracking countries (see Table 1)

differed in terms of their shares of single-sex schools between the primary and secondary school levels. Thus, we ran the same differences-in-differences model, but with the share of single-sex schools as the outcome variable. We computed the share of single-sex schools as the percent of sampled schools that included only boys or only girls. Like for the other outcomes, we summarised the effect estimates across the 33 study matches. In both the early and late tracking countries, we found a very low average share of 2% single-sex schools at the primary school level. These shares were slightly higher at the secondary school level in both early and late tracking countries (8% and 6%, respectively). However, early and late tracking countries did not differ significantly in the share of single-sex schools at the secondary school level when controlling for the share at the primary school level ($p \geq .050$). We interpreted this as underscoring the applicability of the differences-in-differences approach in our study.

Furthermore, we included all countries with a tracking grade below 8 (in TIMSS matches) respectively a tracking age below 15 years (in PISA matches) as early tracking countries in the main analyses. We ran further analyses where we applied the stricter early tracking indicator of tracking grade below Grade 8 to all matches, also the ones with PISA data at the secondary level. We ran the same differences-in-differences analyses and summarised the effect estimates across the study matches as in the main analyses. The effect of very early tracking on the segregation outcome was similar to the one in the main analyses ($\bar{\beta} = 0.043$, $SE = 0.008$, $p < .050$). In the case of gender gaps in student achievement, none of the tracking effect estimates were significant ($p \geq .050$). Thus, this analysis did not confirm the effect on mathematics gender gaps from the main analyses.

Discussion

The aim of this study was to investigate the effect of early between-school tracking on school gender segregation as well as achievement gender gap outcomes. We applied the same country-level differences-in-differences design as Hanushek and Wößmann (2006). We combined all available PIRLS, TIMSS and PISA data in 33 study matches of primary and secondary school data and compared early and late tracking countries in terms of changes between the primary and secondary school levels. Consistent with our expectations, we found strong empirical evidence that early between-school tracking increased school gender segregation at the secondary school level. Yet, not in line with our hypotheses, we found no robust evidence that early tracking shifted gender gaps in student achievement. The effects were very close to zero and either not statistically significant or not robust across the main and further analyses.

This is the first study that has investigated the effect of early tracking on school gender segregation. Two previous studies that investigated early tracking effects on achievement gender gap outcomes found that early tracking increased gender gaps in reading, mathematics and science (Hermann & Kopasz, 2019) or in reading but not significantly in mathematics (Scheeren & Bol, 2021). These studies treated the achievement scales from different international assessments as the same, even though the tests were developed independently and not linked onto the same scale. By

ignoring such differences and estimating student-level differences-in-differences models, these studies may generate more biased estimates than our country-level approach (Contini & Cugnata, 2020). Furthermore, we used data from more countries and assessment cycles than those previous studies. Naturally, it is difficult to identify effects in country-level differences-in-differences studies due to the small sample sizes, especially for early tracking countries. However, our segregation outcome analyses (see Figure 2), alongside previous research on socioeconomic inequalities that used the same approach (Strello et al., 2021; Lavrijsen & Nicaise, 2015), showed that such effects can be identified if they are pronounced and consistent enough.

Implications

Our findings suggest that early tracking increases the degree of school gender segregation. We assumed that since girls have been found to get more favourable teacher evaluations at the primary school level (e.g. Caro et al., 2009; O'Dea et al., 2018; Voyer & Voyer, 2014), they would be placed in higher secondary school tracks more frequently (e.g. Bacher, 2009; OECD, 2015; Róbert, 1991), which would increase the segregation of boys and girls at the secondary school level. This is an important consequence of early tracking, although it is highly understudied in previous research. It is unlikely that the observed early tracking effect on school gender segregation is due to other mechanisms because school gender segregation should not be confounded with residential segregation, for instance (cf. Strello et al., 2022a). Boys and girls should have the same probability of being born in socioeconomically advantaged and disadvantaged families or in rural and urban areas, etc. Furthermore, we excluded countries with high proportions of single-sex schools to avoid a confounding of tracking and single-sex schooling policies.

Our evidence does not support the assumption that the increased school gender segregation translates into achievement advantages among girls in early tracking countries. The tracking effects on gender gaps in reading and science achievement were not significant and in mathematics achievement, they were only significant in one out of two specifications. Furthermore, all tracking coefficients were very close to zero. There are various plausible explanations for these findings. One possible explanation is that the achievement-related effects are simply too small to be detected by our design. However, there are also more substantive explanations. For instance, despite higher school gender segregation in early tracking countries, boys, and not girls, may be selected into the more academic secondary school types – at least in some early tracking countries. It is also possible that even though girls are selected into more academic school types, those schools might not actually be more effective at promoting learning than others. Furthermore, it might be that even though girls are selected into more learning-promoting school types, this advantage is compensated otherwise for boys in early tracking countries. We could, however, not investigate these possible explanations with the data at hand. For instance, the TIMSS and PISA data contain no information about the attended school types. However, it seems promising for future research to explore these potential mechanisms in the different early tracking countries.

Limitations

Even though our study investigated four different outcome variables using all available data from multiple cycles of three international large-scale assessments, its scope is limited. For instance, while we focused on school gender segregation and achievement gender gap outcomes, there are further outcomes, such as gender gaps in school attainment outcomes, that are likewise interesting (cf. Scheeren & Bol, 2021). Furthermore, we only compared early and late tracking countries, without considering the number of tracks in early tracking countries or the presence of within-school tracking policies (cf. Bodovski et al., 2020; van Langen et al., 2006). Chmielewski (2014) found differential effects of between-school tracking and within-school tracking on school socioeconomic segregation and achievement inequality. Thus, it is not clear how effects of within-school tracking on gender segregation and achievement gaps would look like. Since our study design did not allow to investigate this, further research is needed to address this question. We did, however, run an additional analysis where we applied a stricter early tracking definition. Furthermore, we focused on tracking between the primary and secondary school level and not on tracking effects at later educational stages (e.g. Scheeren & Bol, 2021).

Conclusions

This is the first study to investigate the effects of early tracking on school gender segregation at the secondary school level. It analyses the effects of early tracking on achievement gender gaps for the first time using the robust country-level differences-in-differences approach that was introduced by Hanushek and Wößmann (2006). The present study complements research on the effects of early tracking on between-school segregation (Strello et al., 2022) and achievement inequality outcomes (Strello et al., 2021; Lavrijsen & Nicaise, 2015), which has previously mostly focused on socioeconomic groups and not gender. We believe that this study adds an important additional perspective that should be considered when policymakers evaluate early tracking policies and consider reforms.

References

Bacher, J. (2009). Soziale Ungleichheit, Schullaufbahn und Testleistungen [social inequality, school career, and test scores]. In B. Suchan, C. Wallner-Paschon, & C. Schreiner (Eds.), *PIRLS 2006: Die Lese-Kompetenz am Ende der Volksschule. Österreichischer Expertenbericht [PIRLS 2006: The reading competence at the end of primary school. Austrian expert report]* (pp. 79–102). Leykam.

Batruch, A., Autin, F., Bataillard, F., & Butera, F. (2019). School selection and the social class divide: How tracking contributes to the reproduction of inequalities. *Personality & Social Psychology Bulletin*, *45*(3), 477–490. https://doi.org/10.1177/0146167218791804

Bedard, K., & Cho, I. (2010). Early gender test score gaps across OECD countries. *Economics of Education Review*, *29*(3), 348–363. https://doi.org/10.1016/j.econedurev.2009.10.015

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

Bodovski, K., Munoz, I., Byun, S.-Y., & Chykina, V. (2020). Do education system characteristics moderate the socioeconomic, gender and immigrant gaps in math and science achievement? *International Journal of Sociology of Education*, *9*(2), 122. https://doi.org/10.17583/rise.2020.4807

Borghans, L., Diris, R., Smits, W., & de Vries, J. (2019). The long-run effects of secondary school track assignment. *PLOS ONE*, *14*(10), e0215493. https://doi.org/10.1371/journal.pone.0215493

Bowers, A. J. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research and Evaluation*, *17*(3), 141–159. https://doi.org/10.1080/13803611.2011.597112

Card, N. A. (2012). *Applied meta-analysis for social science research*. Guilford.

Caro, D. H., Lenkeit, J., Lehmann, R., & Schwippert, K. (2009). The role of academic achievement growth in school track recommendations. *Studies in Educational Evaluation*, *35*(4), 183–192. https://doi.org/10.1016/j.stueduc.2009.12.002

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

Contini, D., & Cugnata, F. (2020). Does early tracking affect learning inequalities? Revisiting difference-in-differences modeling strategies with international assessments. *Large-Scale Assessments in Education*, *8*(1). https://doi.org/10.1186/s40536-020-00094-x

Dockx, J., & De Fraine, B. (2019). On track for unemployment? Long-term effects of tracks. *School Effectiveness and School Improvement*, *30*(2), 131–154. https://doi.org/10.1080/09243453.2018.1537292

Dockx, J., de Fraine, B., & Vandecandelaere, M. (2019). Does the track matter? A comparison of students' achievement in different tracks. *Journal of Educational Psychology*, *111*(5), 827–846. https://doi.org/10.1037/edu0000305

Duncan, O. D., & Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review*, *20*(2), 210. https://doi.org/10.2307/2088328

Eurydice. (2015). The structure of the European education systems 2013/14. In *Schematic diagrams*. https://op.europa.eu/en/publication-detail/-/publication/b50438fa-b64a-40f8-b319-1b5246f52a6e/language-en

Geven, S., Jonsson, J. O., & Tubergen, F. (2017). Gender differences in resistance to schooling: The role of dynamic peer-influence and selection processes. *Journal of Youth & Adolescence*, *46*(12), 2421–2445. https://doi.org/10.1007/s10964-017-0696-2

Hadjar, A., & Buchmann, C. (2016). Education systems and gender inequalities in educational attainment. In A. Hadjar & C. Gross (Eds.), *Education Systems and Inequalities* (pp. 159–184). Policy Press. https://doi.org/10.1332/policypress/9781447326106.003.0009

Hanushek, E. A., & Wößmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal*, *116*(510), C63–C76. https://doi.org/10.1111/j.1468-0297.2006.01076.x

Hermann, Z., & Kopasz, M. (2019). Educational policies and the gender gap in test scores: A cross-country analysis. *Research Papers in Education*, 1–22. https://doi.org/10.1080/02671522.2019.1678065

Imdorf, C., Hegna, K., Eberhard, V., & Doray, P. (2015), Educational systems and gender segregation in education: A three-country comparison of Germany, Norway and Canada. In Imdorf, C., Hegna, K., & Reisel, L. (Eds.) *Gender Segregation in Vocational Education* (*Comparative Social Research, Vol. 31*) (pp. 83-122). Emerald Group Publishing Limited, Bingley. https://doi.org/10.1108/S0195-631020150000031004

Jürges, H., & Schneider, K. (2011). Why young boys stumble: Early tracking, age and gender bias in the German school system. *German Economic Review*, *12*(4), 371–394. https://doi.org/10.1111/j.1468-0475.2011.00533.x

Kenney-Benson, G. A., Pomerantz, E. M., Ryan, A. M., & Patrick, H. (2006). Sex differences in math performance: The role of children's approach to schoolwork. *Developmental Psychology*, *42*(1), 11–26. https://doi.org/10.1037/0012-1649.42.1.11

Klapproth, F., Glock, S., Krolak-Schwerdt, S., Martin, R., & Böhmer, M. (2013). Prädiktoren der Sekundarschulempfehlung in Luxemburg: Ergebnisse einer Large-Scale-Untersuchung [predictors of the secondary school track recommendation in Luxembourg: Results of a large-scale assessment]. *Zeitschrift für Erziehungswissenschaft [journal for education sciences]*, *16*(2), 355–379. https://doi.org/10.1007/s11618-013-0340-1

Kriesi, I., & Imdorf, C. (2019). Gender segregation in education. In R. Becker (Ed.), *Research Handbook on the Sociology of Education* (pp. 193–212). Edward Elgar Publishing. https://doi.org/10.4337/9781788110426.00020

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

Luyten, H., Bosker, R., Dekkers, H., & Derks, A. (2003). Dropout in the lower tracks of Dutch secondary education: Predictor variables and variation among schools. *School Effectiveness and School Improvement*, *14*(4), 373–411. https://doi.org/10.1076/sesi.14.4.373.17158

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. (2008). Accounting for the gender gaps in student performance in reading and mathematics: Evidence from 31 countries. *Oxford Review of Education*, *34*(1), 89–109. https://doi.org/10.1080/03054980701565279

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 international results in reading*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and the International Association for the Evaluation of Educational Achievement (IEA).

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science*. https://timssandpirls.bc.edu/timss2019/international-results/

O'Dea, R. E., Lagisz, M., Jennions, M. D., & Nakagawa, S. (2018). Gender differences in individual variation in academic grades fail to fit expected patterns for STEM. *Nature Communications*, *9*(1), 3777. https://doi.org/10.1038/s41467-018-06292-0

OECD. (2004). *Learning for tomorrow's world*. OECD Publishing. https://doi.org/10.1787/9789264006416-en

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

OECD. (2015). *The ABC of gender equality in education: Aptitude, behaviour, confidence*. OECD Publishing. https://doi.org/10.1787/19963777

OECD. (2021). *PISA: Programme for International Student Assessment*. https://www.oecd.org/pisa/

Reichelt, M., Collischon, M., & Eberl, A. (2019). School tracking and its role in social reproduction: Reinforcing educational inheritance and the direct effects of social origin. *The British Journal of Sociology*, *70*(4), 1323–1348. https://doi.org/10.1111/1468-4446.12655

Retelsdorf, J., Becker, M., Köller, O., & Möller, J. (2012). Reading development in a tracked school system: A longitudinal study over 3 years using propensity score matching. *The British*

Journal of Educational Psychology, 647–671. https://doi.org/10.1111/j.2044-8279.2011.02051.x

Róbert, P. (1991). Educational transition in Hungary from the post-war period to the end of the 1980s. *European Sociological Review*, *7*(3), 213–236. http://www.jstor.org/stable/522691

Robinson, D. B., Mitton, J., Hadley, G., & Kettley, M. (2021). Single-sex education in the 21st century: A 20-year scoping review of the literature. *Teaching and Teacher Education*, *106*, 103462. https://doi.org/10.1016/j.tate.2021.103462

Rosén, M., Steinmann, I., & Wernersson, I. (2022). Gender differences in achievement. In T. Nilsen, A. Stancel-Piątak, J.-E. Gustafsson (Eds.): *International Handbook of Comparative Large-Scale Studies in Education. Perspectives and Findings*. Springer. https://doi.org/10.1007/978-3-030-38298-8_46-1

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

Scheeren, L., & Bol, T. (2021). Gender inequality in educational performance over the school career: The role of tracking. *Research in Social Stratification and Mobility*, 100661. https://doi.org/10.1016/j.rssm.2021.100661

Strello, A., Strietholt, R., Steinmann, I., & Siepmann, C. (2021). Early tracking and different types of inequalities in achievement: difference-in-differences evidence from 20 years of large-scale assessments. *Educational Assessment, Evaluation and Accountability*, *33*(1), 139-167. https://doi.org/10.1007/s11092-020-09346-4

Strello, A., Strietholt, R., & Steinmann, I. (2022). Does tracking increase segregation? International evidence on the effects of between-school tracking on social segregation across schools. *Research in Social Stratification and Mobility*, *78*. https://doi.org/10.1016/j.rssm.2022.100689

Timmermans, A. C., Kuyper, H., & van der Werf, G. (2015). Accurate, inaccurate, or biased teacher expectations: Do Dutch teachers differ in their expectations at the end of primary education? *The British Journal of Educational Psychology*, 459–478. https://doi.org/10.1111/bjep.12087

TIMSS & PIRLS International Study Center. (2019a). *PIRLS: Progress in International Reading Literacy Study*. https://timssandpirls.bc.edu/pirls-landing.html

TIMSS & PIRLS International Study Center. (2019b). *TIMSS: Trends in International Mathematics And Science Study*. https://timssandpirls.bc.edu/timss-landing.html

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 Hek, M., Buchmann, C., & Kraaykamp, G. (2019). Educational systems and gender differences in reading: A comparative multilevel analysis. *European Sociological Review*, *35*(2), 169–186. https://doi.org/10.1093/esr/jcy054

van Langen, A., Bosker, R., & Dekkers, H. (2006). Exploring cross-national differences in gender gaps in education. *Educational Research & Evaluation*, *12*(2), 155–177. https://doi.org/10.1080/13803610600587016

von Davier, M., Gonzalez, E. J., & Mislevy, R. (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *2*, 9–36.

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, *140*(4), 1174–1204. https://doi.org/10.1037/a0036620

Wiseman, A. W. (2008). A culture of (in)equality? A cross-national study of gender parity and gender segregation in national school systems. *Research in Comparative and International Education*, *3*(2). http://dx.doi.org/10.2304/rcie.2008.3.2.179

III. Overall Discussion

Summary of Findings

Correlation Between Inequalities

The first research question addresses the correlation between the different measures of inequality. Specifically, in this dissertation I investigated the concept of social inequality.

In Article 1, we explored the correlation between different measures of social inequality, based on the social category used to measure the inequality—in this case, the achievement gap between groups. Specifically, we contrasted the achievement gaps between high and low SES groups, between native and immigrant students, and between boys and girls. We found that there was no correlation between SES, immigrant, and gender achievement gaps; some countries had high inequalities under one social category but low inequality in another category. This indicates that there is not an 'umbrella' social inequality, but independent social inequalities. In addition, while SES gaps are present in almost all education systems and almost all follow the same direction, gender gaps and immigration gaps were not present in every context and the gap followed different directions between countries.

In Article 2, we explored the different ways of measuring socioeconomic inequality. We studied the correlation between SES measures used to estimate further SES inequality and different indicators of SES inequality. The results reflected a high correlation between most SES indicators, especially aggregated at the country-level, and a high correlation between SES inequalities based on different SES indicators. However, some differences remain, and some SES indicators are shown to work better in some countries than others. The results also indicated only a low correlation between the scores' dispersion and standardized SES gaps.

Different Measures of Inequality, Different Conclusions? The Case of Between-School Tracking

The second research question concerns how the conclusions of research into school differentiation change according to the measure of inequality studied. We prepared a series of articles on the effects of between-school tracking on different measures associated with inequality. Between-school tracking is a good example of how the educational structure affects student outcomes and school compositions. Its effect on measures besides socioeconomic inequalities, such as inadequacy, segregation, gender gaps, or to a lesser degree inequality, have not received enough attention. We used a similar methodology in each study, exploiting the change between primary and secondary school to perform a difference-in-differences analyses, though we had to adjust each study to each research question.

In Article 3, we studied the effect of between-school tracking on inequality, SES inequality, and inadequacy, using all available ILSAs data at the time the study was designed. The findings indicate different conclusions depending on the measurement of inequality studied. Between-school

tracking had a strong positive effect on SES inequality and was very consistent between replications, it had a positive effect on dispersion inequality but less consistent, and it had a significant but weak positive effect on inadequacy. The study also investigated the effect on mean performance, finding no effect. The study found very robust evidence concerning the effects of between-school tracking, but also highlighted the importance of the differences between conceptions and measures of inequality. Each effect on each concept of inequality was theoretically justified under different terms, and each different effect was evaluated under a different normative framework.

In Article 4, we estimated the effects of between-school tracking on socioeconomic segregation across schools. The findings indicate a strong effect of tracking increasing socioeconomic segregation across schools. The study highlighted the importance of investigating the effects of school structure on the composition of schools, as achievement is not the only good obtained from education. This was also the first such study to perform analyses with international data.

In Article 5, we estimated the effects of between-school tracking on gender segregation across schools and on the achievement gap. On the one hand, we found that tracking had a robust effect on gender segregation; tracking increased the disparity of the gender composition of schools. On the other hand, we found no effect from tracking on the achievement gap. This is interesting, as we would have expected that the bias on the track selection—observable in gender segregation—would also have affected the gap between boys and girls; in addition, this contrasts with what we found when studying the effects on socioeconomic segregation and the socioeconomic achievement gap in study 4 and 3, respectively.

Outreach and Limitations

Causality on ILSAs and Assumptions Behind Studies

Part of this dissertation aimed to determine the effect of between-school tracking on different indicators related to inequality. These articles exploited the variation at international level on the levels of inequalities, segregation, and school differentiation, the repeated iterations of each study from 1995 to 2018, and the availability of probabilistic data at different stages: grade 4 (TIMSS and PIRLS), grade 8 (TIMSS) and 15-year-old students (PISA). This enabled the building of longitudinal datasets aggregated at the country-level and the use of a difference-in-differences design to identify the effect of tracking on the inequality or segregation measure. The main advantage of this design is that it controls for unobserved heterogeneity without the need to include further control variables; the latter assumes that all the relevant variables are observed in the model, an assumption difficult to sustain.

The main limitation of our strategy in these articles is the assumption of parallel trends. These analyses assume that the only relevant variable that changes between primary and secondary school is the streaming of students in some of the countries. This assumption remains at a theoretical level, as it is not possible to test it with the current data we have, especially on such a large scale with the

heterogeneous sample of countries included in the analyses. However, it is a preferable assumption to make, as it is more clearly delimited than the assumptions in normal cross-sectional models; it is more manageable to discuss which other system-level features between primary and secondary school would need to change to invalidate the results, rather than trying to identify all relevant variables in a model, when many of them are not observable. It is also difficult to imagine which other system-level characteristic that could bias the estimates is systematically present in tracking vs. non-tracking countries.

Low-N Analyses and Replications

A natural limitation of all the articles in this dissertation is that the variable of interest was inequality at system-level. All the analyses had to be performed with variables aggregated at the country-level, implying that the analyses are done with a low N. Low-N analyses have their own complications; because of the smaller sample, the results are not as stable as analyses at student level. As observed in Article 3, Article 4, and Article 5, some effect estimates are rather inefficient in several replications, and there are large standard errors with wide confidence intervals. In addition, each sample of countries is not only small, but also varies greatly between studies and cycles, implying that results may change depending on the specific set of studies analyzed. Recognizing this limitation, in studies 3, 4, and 5 we followed a meta-analytical approach, and replicated the results multiple times with different sets of studies and cycles. This approach confirmed the limitations, as the variation of effect estimates was considerably high across replications, but it also meant we could circumvent the issue by synthetizing effects via estimation of mean effects. These studies show the importance of replication of analyses, especially when dealing with data where the N of the variable of interest (in this case, system-level inequalities) is limited.

Future research could further explore some remaining methodological challenges. The studies used two-step approaches, estimating achievement gaps and segregation levels in the first step and using point estimates as data for the second step. This was necessary due the lack of comparability between different studies and different populations (Contini & Cugnata, 2020), but does not include the efficiency of the first-step estimations in the modelling. Including the uncertainty of these point estimations would be even more relevant if similar studies are made using smaller subsamples, such as studies on immigration gaps or segregation.

Importance of ILSAs Data to Study Inequality and Policies

The previous points demonstrate the importance of using international data when studying inequalities and policies. By using ILSAs, we were able to count with variability on the key interest variables within this dissertation. To study the effects of between-school tracking, it is important to use comparable information on countries that do not have tracking (or at a later stage); the alternative is to study within-country policy reforms, as with van de Werfhorst (2018) and Piopiunik (2014), but that approach has other disadvantages, such as with the generalizability of results and the availability of data.

There are some limitations to working with ILSAs data. First, using ILSAs forced us to oversimplify some of the educational features; for example, we measured between-school tracking as a dichotomic variable (except in article 1, where age of tracking was used as the predictor), while educational paths available in each country are rather more complex and tracking is not implemented in the same way across countries. Similarly, we only measured between-school tracking, and counted countries with within-school tracking (such as the USA) as comprehensive or non-tracked systems, even though within-school tracking has been shown to affect student outcomes (Chmielewski, 2014; Chmielewski et al., 2013). This is a natural limitation of working with international data, involving a sacrifice of profundity for comparability across contexts.

Final Remarks

Choosing a Metric: Inequality, Social Inequality, Adequacy, and Segregation

This dissertation reinforces the importance of problematizing which metric of inequality or another related outcome is used. First, independent of the empirical differences, each measure of inequality corresponds to a different normative foundation, i.e., different views of what is fair and what is problematic. The differences between dispersion inequality and social inequality are not as apparent at first, but the rejection or acceptance of the concept of meritocracy is implied in each one; the differences are even more dramatic between inequalities and the related-concept of inadequacy, which focuses on the minimums that education should bring without finding problematic differences between low and high achievers or between socially disadvantaged and socially advantaged students, as long as the former have access to a certain level or quality of education. This dissertation also brings the concept of segregation back as an outcome; this also has a normative assumption (the desirability of students of different origins developing in the same space). This is even more important when discussing public policies, as these principles have to coexist with other sets of values independent to education.

Second, even when disregarding the theoretical discussion, the measures of inequalities are hardly correlated. As shown in this dissertation, the evaluation of system features (such as tracking) changes both theoretically and empirically depending on the metric of interest. In effect, the studies with tracking show that the effects are highly dependent on which concept of inequality is studied. This is reinforced by the correlational studies in this dissertation that show that the social inequalities based on different social indicators correspond to parallel dimensions of social inequality.

Studying Inequalities in an International Context

There are also challenges to studying inequality in an international context. What works one way in one educational system does not necessarily work the same in another. We found great variability across countries in the immigration gap and the gender gap, with changes even in the directionality of the social inequality. SES inequality also showed variation across countries, and the use of a standardized SES measure across a heterogeneous set of countries deserves caution.

School Differentiation

This dissertation extensively studied the effects of between-school tracking on several outcomes of inequalities and segregation. We found important evidence regarding the consequences of tracking. First, we found strong evidence of the effects of tracking on socioeconomic and gender segregation. Second, tracking was shown to increase dispersion inequality and socioeconomic inequality on achievement, with a weak effect on the gender inequality of achievement in mathematics in favor of boys.

School differentiation can be evaluated under different frameworks. From the egalitarian perspective, tracking has noticeable effects on dispersion inequality and especially socioeconomic inequality, although only weak or non-significant effects on gender inequality. However, tracking is shown to not only increase socioeconomic segregation, but also gender segregation across schools. From an adequacy perspective, we find weak evidence for the negative effects of tracking, with no effect on the mean performance of early tracking countries. In modern societies which place great emphasis on specialization in the job market, education systems can be expected to separate worse and better performers, but attention should be paid to both social equality and adequacy, as there is no justification for differences that are not meritocratic or school systems that do not achieve a minimum standard of quality.

Conclusion: Measurement Matters

The measurement of inequality and related concepts matters. Researchers should rationalize and make explicit which frameworks underlie their studies of inequality and segregation.

References

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

Contini, D., & Cugnata, F. (2020). Does early tracking affect learning inequalities? Revisiting difference-in-differences modeling strategies with international assessments. *Large-Scale Assessments in Education*, *8*(1). https://doi.org/10.1186/s40536-020-00094-x

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

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

IV. Appendix

Publication Status of the Individual Contributions

Article 1:

Accepted for publication.

Strello, A., Strietholt, R., & Steinmann, I. (in press). Mind The Gap… But Which Gap? The Distinctions Between Social Inequalities in Student Achievement. *Social Indicators Research*.

Article 2:

Strietholt, R., & Strello, A. (2022). Socioeconomic Inequality in Achievement: Conceptual Foundations and Empirical Measurement. In Nilsen, T., Stancel-Piątak, A., Gustafsson, JE. (eds), *International Handbook of Comparative Large-Scale Studies in Education: Perspectives, Methods and Findings*. Springer. https://doi.org/10.1007/978-3-030-38298-8_11-1

Article 3:

Strello, A., Strietholt, R., Steinmann, I., & Siepmann, C. (2021). Early tracking and different types of inequalities in achievement: difference-in-differences evidence from 20 years of large-scale assessments. *Educational Assessment, Evaluation and Accountability*, *33*, 139-167. https://doi.org/10.1007/s11092-020-09346-4

Article 4:

Strello, A., Strietholt, R., & Steinmann, I. (2022). Does tracking increase segregation? International evidence on the effects of between-school tracking on social segregation across schools. *Research in Social Stratification and Mobility*, *78*. https://doi.org/10.1016/j.rssm.2022.100689

Article 5:

Steinmann, I., Strello, A., & Strietholt, R. (2023). The effects of early between-school tracking on gender segregation and gender gaps in achievement: a differences-in-differences study. *School Effectiveness and School Improvement,* *34*(2), 189-208. https://doi.org/10.1080/09243453.2023.2165510

Author Contributions to Articles

Article 1: Strello, A., Strietholt, R., & Steinmann, I. (in press). Mind The Gap… But Which Gap? The Distinctions Between Social Inequalities in Student Achievement. *Social Indicators Research*.

- Conception of the study and article: Strello designed the article and original idea, advised by Strietholt and Steinmann.
- Statistical analysis: Strello had the main responsibility in analysis, supervised by Strietholt and Steinmann.
- Conception of the written version of article: Strello was main responsible in writing, Strietholt and Steinmann advised and contributed with pieces of text.
- Written drafting and revision after peer reviewing: the article was written mainly by Strello, with revisions from Strietholt and Steinmann.

Article 2: Strietholt, R., & Strello, A. (2021). Socioeconomic inequality in achievement: conceptual foundations and empirical measurement. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International Handbook of Comparative Large-Scale Studies in Education: Perspectives, Methods and Findings*. Springer. https://doi.org/10.1007/978-3-030-38298-8_11-1

- Conception of the study and article: Strietholt designed main idea, with further comments of Strello.
- Statistical analysis: Strello had the main responsibility in analysis, supervised by Strietholt.
- Conception of the written version of article: Strietholt was the main responsible in writing, Strello advised and contributed with pieces of text.
- Written drafting and revision after peer reviewing: the article was written mainly by Strietholt, with revisions from Strello.

Article 3: Strello, A., Strietholt, R., Steinmann, I., & Siepmann, C. (2021). Early tracking and different types of inequalities in achievement: difference-in-differences evidence from 20 years of large-scale assessments. *Educational Assessment, Evaluation and Accountability, 33,* 139-167. https://doi.org/10.1007/s11092-020-09346-4

- Conception of the study and article: Strietholt, Strello, and Steinmann.
- Statistical analysis: Strello had the main responsibility in analysis, supervised by Strietholt and Steinmann. Siepmann and student assistants contributed with preliminary analyses. Siepmann did original categorization of countries, later revised by Strello.
- Conception of the written version of article: Strello was main responsible in writing, Strietholt and Steinmann advised and contributed with pieces of text
- Written drafting and revision after peer reviewing: the article was written mainly by Strello, with revisions from Strietholt and Steinmann.

Article 4: Strello, A., Strietholt, R., & Steinmann, I. (2022). Does tracking increase segregation? International evidence on the effects of between-school tracking on social segregation across schools. *Research in Social Stratification and Mobility*, *78*. https://doi.org/10.1016/j.rssm.2022.100689

- Conception of the study and article: Strello designed the article and original idea, advised by Strietholt and Steinmann.
- Statistical analysis: Strello had the main responsibility in analysis, supervised by Strietholt and Steinmann.
- Conception of the written version of article: Strello was main responsible in writing, Strietholt and Steinmann advised and contributed with pieces of text
- Written drafting and revision after peer reviewing: the article was written mainly by Strello, with revisions from Strietholt and Steinmann.

Article 5: Steinmann, I., Strello, A., & Strietholt, R. (2023). The effects of early between-school tracking on gender segregation and gender gaps in achievement: a differences-in-differences study. *School Effectiveness and School Improvement, 34*(2). 189-208. https://doi.org/10.1080/09243453.2023.2165510

- Conception of the study and article: Steinmann designed the idea, with further comments of Strello and Strietholt.
- Statistical analysis: Steinmann made data preparation, with analysis done by Strello and further revised by Steinmann.
- Conception of the written version of article: Steinman was main responsible in writing, with comments by Strello and Strietholt.
- Written drafting and revision after peer reviewing: the article was written mainly by Steinmann, Strello and Strietholt contributed with further revisions after peer reviewing.

Eidesstattliche Erklärung

Hiermit versichere ich schriftlich und eidesstattlich gemäß § 11 Abs. 2 PromO v. 08.02.2011:

(1) Die von mir vorgelegte Dissertation ist selbstständig verfasst und alle in Anspruch genommenen Quellen und Hilfen sind in der Dissertation vermerkt worden.

(2) Die von mir eingereichte Dissertation ist weder in der gegenwärtigen noch in einer anderen Fassung an der Technischen Universität Dortmund oder an einer anderen Hochschule im Zusammenhang mit einer staatlichen oder akademischen Prüfung vorgelegt worden.

_____ _____

Ort, Datum Unterschrift

(3) Weiterhin erkläre ich schriftlich und eidesstattlich, dass mir der „Ratgeber zur Verhinderung von Plagiaten“ und die „Regeln guter wissenschaftlicher Praxis der Technischen Universität Dortmund“ bekannt und von mir in der vorgelegten Dissertation befolgt worden sind.

_____ _____

Ort, Datum Unterschrift