Review of
Educational Research

The Effect of School Tracking on Student Achievement and Inequality: A Meta-Analysis

Journal:	*RER*
Manuscript ID	RER-21-Feb-MS-103.R2
Manuscript Type:	Manuscript

The Effect of School Tracking on Student Achievement and Inequality: A Meta-Analysis

Abstract

This meta-analysis examines the effects of sorting secondary students into different tracks ("between-school" tracking) or classrooms ("within-school" tracking) on the efficiency and inequality levels of an educational system. Efficiency is related to the overall learning achievement of students, while inequality can refer to "inequality of achievement" (i.e., the dispersion of outcomes) or "inequality of opportunity" (i.e., the strength of the influence of family background on student achievement). The selected publications are 53 analyses performed in the period from 2000 to 2021, yielding 213 estimates on efficiency and 230 estimates on inequality. The results show that the mean effect size (Hedge's G) of tracking on efficiency is not statistically significant (*G = -.063*), whereas it is significantly positive (*G = .117*) on equality. We further set out to explain variation in effect sizes by (a) policy characteristics, (b) the operationalization of main variables, (c) the research design, (d) the set of control variables included in the statistical analyses, and (e) the quality of the study, year of publication, and publication status (peer-reviewed or not peer-reviewed).

Keywords: secondary school tracking, achievement, efficiency, inequality, meta-analysis

The Effect of School Tracking on Student Achievement and Inequality: A Meta-Analysis

Almost all industrialized societies provide unified initial schooling in which all students receive the same general education and are exposed to a relatively homogeneous school environment. Nevertheless, in many countries, students are allocated to different types of education at some point in their educational career (Blossfeld et al., 2016). Indeed, most school systems practice some form of tracking, through which they provide a differentiated learning environment to students (Dupriez et al., 2008).

In a broad sense, tracking (also known as streaming, sorting, or ability grouping) refers to the assignment of students to different types of education—kind of school, curricula, subjects, classes—according to their ability, interests, or attitudes (Betts, 2001; Brunello & Checchi, 2007; Woessmann, 2009). School tracking is considered by social scientists to be one of the most important features of school systems because it has important consequences for students' school performance, educational pathways, entrance into higher education, and subsequent labor market outcomes (Brunello & Checchi, 2007; van de Werfhorst & Mijs, 2010).

National educational systems differ in the way they organize secondary education. Some offer different pathways to students or group them based on their ability and interests relatively early in their educational career, while others opt for later student sorting or a more comprehensive model of education (Dupriez et al., 2008; Blossfeld et al., 2016). It has long been argued that the shape of secondary education has consequences for the efficiency of an educational system, that is, the overall achievement of students (Betts, 2001; Clifford & Heath, 1984; Gamoran & Mare, 1984).

The underlying argument supporting student tracking is that more homogeneous groups of students allow teachers to tailor their pedagogical strategies to students' abilities and interests, making the learning process more efficient. This is called the "specialization effect" and, following this argument, tracking can allow students to learn more. Nonetheless, the possibility that school tracking may affect educational inequality must also be considered (van de Werfhorst & Mijs, 2010). Some authors argue that tracking increases inequality among students (Gamoran & Mare, 1989; Lucas, 1999/2001; Oakes, 1985) not only in terms of dispersion of achievement (Galindo-Rueda & Vignoles, 2005; Hanushek & Woessmann, 2006; Manning & Pischke, 2006) but also in terms of inequality of opportunities, namely, the strength of the influence of family background on student outcomes. These studies usually quantify the extent to which economic, cultural, and social resources in the home environment affect student achievement across educational systems characterized by different institutional arrangements (Bol & van de Werfhorst, 2013; Brunello & Checchi, 2007; Jackson, 2013). From this perspective, gains in achievement in a tracked system concern students at more advanced levels or with higher socioeconomic status, whereas students from socio-economically disadvantaged families face deterioration in their learning progress (Hallinan, 1994).

Studies on the effects of tracking are important not only for the academic debate but also because of their informative potential in shaping educational policies. Three scenarios arise regarding the effect of tracking on both dimensions (efficiency and inequality). First, if tracking increases efficiency as it exacerbates inequality, policy-makers will be dealing with an "efficiency-equity trade-off" (Skopek et al., 2019, p. 224): is it more important to maximize student performance at all costs, or is it preferable to distribute education more equally within the system? Second, if the effects of tracking are negative or null in terms of efficiency, would

increasing inequality only for the benefit of more privileged students be legitimate? Third, if the effects on both efficiency and inequality are null, would tracking still be a valid policy choice for providing students with a more specialized education?

While a fully comprehensive assessment of the effects of tracking would also include its longer-term consequences for employability and occupational outcomes, in this analysis, we will focus on the consequences of tracking for short-term outcomes only, that is, educational achievement. This homogenizes tracking outcomes both in conceptual terms and in their measurements, thus posing fewer problems in the design of the meta-analytical review. Furthermore, thanks to the existence of several international pre-harmonized surveys using student test scores, a far greater amount of consolidated empirical literature is available regarding the effects of tracking on educational achievement than about its longer-term consequences for occupational prospects.

As suggested by Woessmann (2009), since there are plausible contrasting arguments at a theoretical level about the role of tracking in contemporary educational systems, it is of paramount importance to address this issue at the empirical level. However, there is no consensus in the empirical literature. Indeed, some studies report either positive (Ariga & Brunello, 2007; Galindo-Rueda & Vignoles, 2005; Horn, 2013; Lassibile & Gomes, 2010) or null effects of tracking on the overall level of student performance, and others report negative (Ammermüller, 2005; Bol & van de Werfhorst, 2013; Brunello & Checchi, 2007; Dunne, 2010; Hanushek & Woessmann, 2006; Piopiunik, 2014) or null effects (Jakubowski, 2010; Ruhose & Shwerdt, 2016; Waldinger, 2007) on social inequalities. Differences in the qualitative direction of the results and the magnitude of the effects might stem from the heterogeneity of samples, the

definition of tracking, the type of research design, and several other characteristics (e.g., Ammermüller, 2005; Jakubowski, 2010; Waldinger, 2007).

The last extensive meta-analysis of secondary tracking was performed in the 1980s by Kulik and Kulik (1982). It focused exclusively on within-school tracking and the problem of efficiency, with only side concerns for inequality. Their results pointed to small yet significant effects of ability grouping on overall achievement. Given that the empirical literature has since flourished, especially in the 2000s, due to growing attention to the topic and the increased availability of suitable data and analytical techniques, it is important to understand the findings of recent research. The few existing reviews have summarized the key elements of the recent literature. In their assessment, van de Werfhorst & Mijs (2010) suggest that tracking increases inequality and lowers efficiency. Meanwhile, Woessmann (2016) shows that tracking is a crucial institutional feature for explaining differential performances across countries in international standardized tests such as the Program for International Student Assessment (PISA) and that evidence points to a deterioration of the performance of students placed in low and middle tracks. Skopek et al. (2019) review the plurality of approaches employed to analyze the effects of tracking, suggesting that the heterogeneous results of empirical studies might depend partly on the diversity of analytical strategies. However, these reviews are non-systematic and cannot provide an exhaustive description of the phenomenon or explain the most important drivers of the variation in findings across empirical studies with proper statistical analysis. Older reviews include the best-evidence synthesis performed by Slavin (1990) and the summary of survey and ethnographic research compiled by Gamoran and Berends (1987), both published by this journal. The former found no effect of ability grouping on overall achievement in studies that treated tracking as a systemic variable and in studies that evaluated the effect of placement in different

ability classes. The latter stressed the difficulties of the literature in disentangling the influence of tracks from the effect of students' social-class, calling for more robust quantitative longitudinal research, which only flourished recently and will be reviewed in this meta-analysis.

Therefore, we believe that the field would benefit from a systematization of these research findings capable of answering two key questions in the current debate. First, to what extent does tracking increase or decrease efficiency and inequality in the educational system? Second, to what extent can the variation in results across studies be explained by the latter's characteristics?

We propose a meta-analysis of studies that operationalized tracking as a systemic/institutional macro-level variable (i.e., the sorting of students within the educational system) and measured its impact on efficiency and/or inequality for the educational system as a whole. The meta-analytic method offers some advantages. Through its concept of "mean effect size," it allows us to take the estimates reported in different statistical forms and produced based on different research designs and make them comparable through a standardization process that mainly takes into account sample size and variance (Borenstein et al., 2009; Lipsey & Wilson, 2001). Second, via the systematic coding of studies, it is possible to explain variation in effect sizes by study and policy features. We (meta-)regress effect sizes on five blocks of explanatory variables: (a) policy characteristics, (b) the operationalization of the tracking and outcome variables, (c) the research design, (d) the set of control variables included in the statistical analyses, and (e) the quality of the study, year of publication, and publication status (peer-reviewed or not peer-reviewed). All of these elements received diverse treatment in the studies selected for analysis, each possibly contributing to variation in effect sizes. We run separate analyses for efficiency and inequality.

Theoretical framework

In the following section, we discuss the main mechanisms highlighted in the literature that could explain the effects of tracking on efficiency and inequality in educational systems. They include (a) institutional characteristics (e.g., curricula), (b) characteristics of the learning environment (e.g., classroom composition and peer effects), and (c) the unequal distribution of educational resources (e.g., teachers' characteristics, class size). We will review the main theoretical arguments in favor of the tracking policy before considering the arguments that highlight its potentially adverse effects.

Potential benefits of tracking

The proponents of tracking argue that early tracking can have a positive effect on the overall level of learning and school performance because, in a tracked system, students attend a school environment that is tailored to their needs and receive instruction adequate for their skill level, which in turn maximizes each student's potential. The most important pedagogic rationale for tracking is that students differ in their academic potential and the environments in which they learn best.

A tracked school system is also helpful for teachers, who can develop their teaching approach based on the ability level of the classroom, which is on average more homogeneous than in a comprehensive school system because of the prior selection and streaming (Lassibile & Gomez, 2010; Manning & Pischke, 2006). From a systemic point of view, if the allocation is efficient—that is, students are appropriately assigned to the track that best fits their ability level—a tracked system should produce an overall higher level of student performance than a

non-tracked system. In other terms, by raising average outcomes, efficient tracking should enhance educational productivity (Betts, 2001).

According to Gamoran and Mare (1989), this argument can be traced back to old studies that analyzed the effects of classroom homogeneity on educational achievement (Cook, 1924; Keliher, 1931; Whipple, 1936) and concluded that a more homogeneous environment improves the efficiency of the learning process. An underlying assumption of this view is that the process of sorting students should be based solely, or at least largely, on students' academic abilities. In this perspective, an appropriately designed tracking system could not only positively affect the efficiency of the system but also compensate for inequalities in achievement between low- and high-performing students. If, after track allocation, the achievement of low-track students increases more than that of high-track students, inequality will decline (Gamoran & Mare, 1989).

More recent works suggest that, in many educational systems, the process of sorting students occurs not only on the basis of their abilities but also by taking into account their attitudes and aspirations, leaving space for families to choose their child's academic destination independently from their demonstrated academic competence (Jackson, 2013). Some students aim to enter the labor market early and develop occupation-specific and applied skills. On the contrary, others are oriented to continue to study at university either because they like studying and do not have well-defined occupational expectations or because their target job requires a university degree (e.g., doctor, lawyer). Many European educational systems differentiate between schools offering academic-oriented curricula and others providing vocationally oriented instruction and training (Shavit & Müller, 1998). A school with an efficient tracking system should provide study courses appropriate for different types of students to better address their needs. As suggested by Gamoran and Mare (1989), "ideally, a system of academic tracking

matches students' aptitudes with the objectives and learning environments to which they are best suited" (p. 1148). If student sorting can effectively assign students to the most appropriate track, it could lead to higher levels of learning and academic performance, greater student satisfaction, and lower dropout rates.

Why tracking can be harmful

At the theoretical level, tracking can also have negative consequences for both efficiency and equity. First, it has been highlighted that the expected benefits outlined in the previous section depend on the extent to which institutions and families are able to place the students in the type of school or study program that is most appropriate for their skill and aptitude levels. In this respect, it is often claimed that the earlier the separation of students occurs, the greater the likelihood of making an inappropriate assignment (Betts, 2001; Brunello & Checchi, 2007; Dustmann, 2004). This is because the pace of cognitive development varies across students, as well as the degree of maturation and development of non-cognitive skills that are conducive to higher achievement (Cohn, 1991).

Several authors have pointed out possible negative outcomes of tracking in terms not only of dispersion of achievement levels but also inequality in educational opportunity (Triventi et al., 2019; van de Werfhorst & Mijs, 2010). First, it is likely that the stratification introduced by the practice of tracking particularly penalizes students from disadvantaged socio-economic backgrounds when the placement in lower-level study programs is affected by pupils' social origin, beyond school achievement. If enrolment in different types of schools is free and depends on families' decisions (as in Italy), the process of segregation in different types of education is mainly based on students' self-selection. In those school systems where allocation to different

tracks is based on teacher recommendations (as in some German *Länder*), the resulting social stratification is mainly due to institutional practices (Contini & Scagni, 2011).

In a free-choice system, for culturally embedded reasons (Bourdieu & Passeron, 1970) or because of risk aversion (Breen & Goldthorpe, 1997), lower-status parents may decide to enroll their children at a vocational school even if they have a relatively good school record. On the contrary, children from higher socio-economic status (SES) families with a similar level of achievement are more likely to attend the academic track to increase their chances of admission to university in the future (Gambetta, 1987).

In a system taking into account teacher recommendations, students of lower social origin are more likely to be directed toward vocational schools or study programs, for two reasons: (a) lower-class pupils may have lower school achievement in primary school because of differences between social classes in the development of cognitive abilities in early age (Brunello & Checchi, 2007); (b) teacher recommendations can be biased in favor of upper-class children due to teachers' beliefs about students' "natural" attitudes as well as pressures from upper-class parents to place their children in the higher track (Barg, 2013; Brantlinger, 2013; Laureau, 1987).

Tracking can exacerbate social inequality in school achievement and performance through several mechanisms. The first one is peer-group effects: if highly motivated and high-achieving pupils are grouped, the low-achieving students are segregated and cannot benefit from proximity to such peers (Betts & Shkolnik, 2000; Deci et al., 2001; Esser, 2016). Additionally, attending school with more motivated peers could create a favorable learning environment, allowing teachers to devote more time to effective teaching and less time to managing the classroom or dealing with students' misconduct. The second is teacher sorting (Bonesronning et al., 2005): it is possible that the ablest and most motivated teachers prefer to teach the brightest

students in the academic track (Brunello & Checchi, 2007, p. 795). If this is the case, teachers engage in positive self-selection into the higher-level tracks, resulting in an increased learning gap between students placed in high and low tracks. For instance, until a few years ago, teachers' official preparation for the academic track in Germany lasted longer and was more complete than for other tracks (Roloff et al., 2020, p. 3). The third mechanism refers to differences in the quality of curricula and teachers' expectations. In higher-level tracks, educational standards are stricter and teachers often have higher expectations regarding their students' academic potential. This could contribute to improving student performance in this type of school because positive teacher expectations (Rosenthal & Jacobson, 1968) and rigorous educational standards seem to foster school achievement (Betts & Grogger, 2003). The fourth potential mechanism concerns the educational resources invested in different tracks, such as average expenditure per student or student-teacher ratio (Brunello & Checchi, 2007). One example of potentially heterogeneous resources across tracks is class size. A student in a class with a small number of students is more likely to receive personal teacher support during lessons than students in more crowded classrooms. Some international evidence points to lower student-teacher ratios in general tracks than in vocational tracks (Brunello & Checchi, 2007, p. 795). However, the extent to which this can exacerbate inequalities or compensate for them seems to vary across countries (Betts, 2001).

In summary, since tracking policy can intensify inequalities among students in terms of the quality of their regular educational environment (educational resources, peers), we expect it to lead to heterogeneous outcomes across tracks or ability groups, canceling out any possible improvement in efficiency produced by specialization. Therefore, we hypothesize that *tracking does not increase the overall efficiency (achievement) of an educational system* (H1a) and that

tracking increases the level of inequality of achievement/opportunity within an educational system (H1b).

If tracking increases efficiency as it increases inequality, support is provided for the "equality-efficiency trade-off" (Skopek et al., 2019, p. 224). If its effect is null on efficiency but positive on inequality, support is provided for studies that are more critical of tracking. Finally, if its effects are null on both dimensions, support is provided for the organizational and transitional arguments, that is, the idea that tracking is still valid for attending to students' specific interests and facilitating their transition to the labor market.

Explaining variations across studies: hypotheses on the role of meta-regressors

As anticipated, research on the effect of tracking as an organizational feature of the educational system used a variety of tracking measures, research designs, datasets, and country or regional samples. We set out to explain variation in effect sizes based on five blocks of explanatory variables: (a) policy characteristics, (b) the operationalization of tracking and outcome variables, (c) the research design, (d) the set of control variables included in the statistical analyses, and (e) the quality of the study, year of publication, and publication status.

Policy characteristics

Type of tracking. A basic distinction in tracking policy is that between "within-school" and "between-school" tracking. While the first is more common in anglophone countries, the latter is usually observed in Europe (Betts, 2000; Blossfeld et al., 2016). Within-school tracking, frequently referred to in the literature as "ability grouping," consists in assigning students to different groups or classrooms (e.g., basic, intermediate, or advanced mathematics classes) within the same school. Between-school tracking assigns students to different school types (e.g., academic, technical, or vocational tracks). This distinction is important because the type of tracking adopted by an educational system will affect the change in school/class composition and the distribution of educational resources differently.

For instance, within-school tracking affects student group composition only at the classroom level, where the learning process mainly takes place, whereas between-school tracking alters it more drastically, re-shaping peer composition at both the classroom and school levels. Consequently, we believe that the effects of tracking should be different for each of these modalities. Students that are sorted into different schools will learn in a fully specialized environment, while students that are sorted only into different classes or courses will have a part-time specialization, still sharing the school environment and potentially other courses with less advanced peers. Thus, we expect that *between-school tracking can increase efficiency more than within-school tracking* (H2a) and that *between-school tracking can increase inequality more than within-school tracking* (H2b).

Outcome variables

Subject domain. Several studies operationalize the dependent variable by pooling scores from different domains in one single overall score (Ayalon, 2006; Kerr et al., 2013; Koerselman, 2013; Piopiunik, 2014). Nevertheless, numerous others maintain the separation between subject domains (Hanushek & Woessmann, 2006; Waldinger, 2007), assuming that tracking may have differentiated effects across school subjects (Dammrich & Triventi, 2018).

The bulk of reading and mathematical competencies is developed at school. However, social background seems to play a more important role in the development of reading skills than in that of mathematical competencies (Bol et al., 2014; Cooper et al., 1996). Verbal and reading skills are also acquired via interactions with parents and peers and reading at home, whereas mathematical competencies are mostly developed at school (Cooper et al., 1996). Following this reasoning, on theoretical grounds, we can expect student achievement in mathematics to be more sensitive to the characteristics of the educational system than student achievement in reading. Thus*, the effect of tracking should affect mathematics scores more markedly than reading scores* (H3a) and *create more inequality in mathematics than in reading* (H3b).

Type of outcome. As anticipated, we divide studies into two main groups. The first analyzes the effect of tracking on the average level of student achievement, which is thought to capture the consequences of school tracking for efficiency. The outcome is usually measured based on test scores constructed by applying item-response theory or related techniques to a set of student answers to a standardized test. The second group focuses on the effect of tracking on achievement inequality. Given that achievement inequality has been measured in two very distinct ways in tracking studies, a specific hypothesis for the operationalization of the dependent variable is needed for research focusing on this outcome.

The first subgroup of studies measures achievement inequality in terms of dispersion of student performance, that is, the standard deviations of scores, interquartile range, score gap between top and low performers, and ratio of low and top performers, inter alia (Bol & van de Werfhorst, 2013; Galindo-Rueda & Vignoles, 2005; Hanushek & Woessmann, 2006; Jakubowski, 2010; Vandenberghe, 2006). The second subgroup of studies, in their basic setting, measures inequality of opportunities by interacting an indicator of family background (e.g., number of books at home, parental education, parental occupation or income) with a macro-variable capturing the tracking regime (e.g., Ammermüller, 2005; Le Donné, 2014). A positive interaction indicates that the effect of social background on student achievement (i.e., inequality of opportunity) is larger in tracked than in untracked systems. While achievement dispersion is an inter-individual measure of inequality, the second is a group-based measure of inequality in which categories of students are identified on the basis of an ascriptive trait (Breen & Jonsson, 2005) that can lead to scholastic advantages or disadvantages for which students do not bear any responsibility (Roemer, 1998). Many authors agree that both measures of achievement inequality are needed to develop a more comprehensive view of the social consequences of tracking (Schütz et al., 2008; Waldinger, 2007; Pfeffer, 2008; Schlicht et al., 2010; Burger, 2016; Ruhose & Schwerdt, 2016).

It is not easy to formulate a hypothesis on whether we should expect a higher level of inequality using one measure or the other because empirical research has not engaged in this comparison explicitly. To this end, we should consider that since inequality of opportunity in these studies is measured in a more indirect way (in a moderation analysis) than achievement dispersion, we can tentatively expect that *measurements of inequality of opportunity tend to report smaller estimates than measurements of inequality in achievement* (H3b.2).

Tracking indicator

The independent variable (tracking) is operationalized diversely across the studies. A first basic distinction concerns whether tracking is treated as a dummy variable ("tracked" vs. "comprehensive" or "highly tracked" vs. "moderately tracked" systems) or as a continuous variable (e.g., age at first tracking, number of years exposed to tracking, or number of tracks offered). If tracking is operationalized as a dummy variable, the study reflects the full potential effect of a tracking policy. Conversely, if it is operationalized as a continuous variable, the study reports the effect of one additional unit of a certain tracking component.

Another important distinction in this sense is when tracking is operationalized based on reform dummies. Systems that have implemented educational reforms—such as Finland, the United Kingdom, Sweden, Poland, and Bavaria in Germany—offer interesting cases for measuring the impact of tracking because of its longitudinal nature. Yet, one important drawback of studies that analyze tracking reforms is that if the reform includes any institutional or organizational change other than tracking (e.g., interventions on time of instruction, curricula modifications), its effects are very difficult to isolate (Koerselman, 2013). Additionally, predicting the direction of the potential bias is tricky since it depends on the nature of concomitant interventions. In any case, we believe that by focusing on longitudinal variations within countries, the studies analyzing reforms are not affected by time-constant unobserved heterogeneity at the macro-level, which is more likely to plague estimates in international studies, especially those based on cross-sectional designs. Following this consideration, we hypothesize that *studies that analyze national educational reforms tend to report smaller effects of tracking both for efficiency* (H4a) *and inequality* (H4b).

Research design

Choice of counterfactual. One important feature of tracking studies is the type of reference category or counterfactual to which tracking is compared. Many studies compare tracking not to comprehensive systems but to tracking regimes with less strict rules. The most obvious example is studies comparing regimes that track at a very early age to regimes that track later. Bauer and Riphan (2013, p. 112), for instance, take advantage of the variation in tracking rules across Swiss cantons: the first tracking in Switzerland occurs mostly between grade 5 (approximately 10–11 years of age) and grade 7 (12–13 years). Since their outcome variable comes from a dataset that measured the ability of students at the age of 17, all students had already been tracked by this time. Thus, in this study, the estimated impact was the consequences of tracking one or two years earlier, not the impact of attending a tracked rather than a comprehensive system. Other examples of this type of comparison can be found in Dronkers et al. (2012) and Dunne (2010), who contrasted "between-school" tracking with the "within-school" modality.

We set out to test whether studies that compare tracking with comprehensive systems (Felouzis & Charmillot, 2008; Hanushek & Woessmann, 2006; Hoffer, 1992; Horn, 2013; Jakubowski, 2010; Pfeffer, 2008) report larger estimates for tracking than those that use other types of reference categories. Since the contrast is starker when tracking regimes are compared to non-tracked ones, we hypothesize that *studies that use comprehensive education as their counterfactual tend to report larger effects of tracking on achievement* (H5a) *and inequality* (H5b) *than those whose counterfactuals represent less-tracked systems.*

Analytical design. When looking at comparative studies investigating the role of school tracking systems on inequalities in educational outcomes, two main types of research designs can be identified (Skopek et al., 2019): cross-sectional multilevel-analysis (e.g., Ayalon & Gamoran, 2000; Burger, 2016; Schütz et al., 2008) and difference-in-differences designs (e.g., Hoffer, 1992; Zimmer, 2003).

In the first stream of research, the increased availability of international large-scale student surveys – e.g., PISA, Progress in International Reading Literacy Study (PIRLS), Trends in International Mathematics and Science Study (TIMMS) – has promoted the development of comparative quantitative research on educational inequalities. Multilevel cross-sectional designs rely on the variation in educational policies across countries, taking into account the hierarchical nature of educational data and correcting for its dependencies (e.g., students nested in schools, nested in countries). Since individual achievement is explained by variation at the macro-level, it is rather unlikely that confounders at the micro-level are correlated with these macro-features. A problem with this approach, however, is that it does not rule out confounders at the macro-level (Hanushek & Woessmann, 2006). Variation in efficiency and inequality across national educational systems could be attributed not only to tracking but also to differences between countries regarding other features, such as GDP, educational investment, or even national culture. While some of these macro-level confounders can be taken into consideration, the limited number of countries available makes it impossible to account simultaneously for all possible confounders, leaving open the possibility that the effect of tracking estimated by these studies is biased by unobserved factors.

The second strategy, which was introduced in this field by the pioneering work of Hanushek and Woessmann (2006), intends to overcome some of the aforementioned limitations.

Applied to international data on student test scores, the difference-in-differences strategy aims to enhance the internal validity of the estimated effect of tracking by relying on (at least) two points in time. The key idea is to compare student outcomes (e.g., PISA scores measured around the age of 15) in countries that have already tracked students and countries that have not yet tracked them, comparing this difference with a previous point in time at which tracking had not yet occurred in all countries (e.g., PIRLS and TIMSS scores measured around the age of 10). By relating changes in outcomes to the changing policy (from untracked to tracked in a subsample of countries), this design can remove time-constant heterogeneity across countries. While this is a valuable improvement on the previous strategy, the causal interpretation of the tracking effect will also be undermined in this design if there are unaccounted changes in other institutional features between the two data points, which correlate with changes in tracking practice and influence student performance (e.g., expenditure on education, quality of teachers).

In any case, in line with the literature (Skopek et al., 2019; van de Werfhorst, 2019), we believe that longitudinal designs are better at granting internal validity and ruling out possible alternative explanations. Thus, we hypothesize that, *compared to the cross-sectional multilevel design*, *the difference-in-differences design tends to report smaller effects of tracking on overall student achievement* (H6a) *and inequality* (H6b).

Inclusion of educational mechanisms/resources. In discussing our theoretical framework, we have highlighted that tracking could affect student outcomes by altering the school/class composition and educational resources to which children are exposed in secondary education. Relying on the literature on counterfactual causal inference, we distinguish between studies that estimate the total effect of tracking and those that estimate an over-controlled model by including potential mediators as well, that is, variables that intervene in the relationship between tracking

and outcomes (Angrist & Pischke, 2008; Pearl, 2009). Regarding the set of controls, we classify studies into four groups: (a) studies that control for educational resources (e.g., student-teacher ratio, time of instruction, teacher qualifications), (b) studies that control for school/class composition (e.g., school average SES or proportion of minority status among students), (c) studies that account for both sets of controls (i.e., composition and resources), and (d) studies that do not control for either of these mechanisms. Consequently, we hypothesize that *when a study controls for educational resources and/or school/class composition indicators, the effect sizes of tracking tend to be smaller for efficiency* (H7a) *and inequality* (H7b).

Quality, year, and publication status

Quality score. In addition to the choice of an adequate research design, the quality and rigor with which they are applied can explain variation in effect sizes. We propose an evaluation matrix that takes into account general and more specific criteria of evaluation (adapted from Conn, 2017). The first two criteria are general and apply to all research designs. The first criterion (a) takes into account the quality of the language and whether basic quantitative information is missing (e.g., descriptive statistics, sample size, standard errors). The second criterion (b) considers how well presented the research design is and whether the authors show awareness of its limitations. The third and fourth criteria are design-specific and focus on the application of best practices to reduce threats to internal validity. The third criterion (c) only concerns studies that use a difference-in-differences design and reflects whether they included at least one previous point in time to check for pre-existing trends and whether they balance treatment and control groups on important observable variables (i.e., for which variables they control). The fourth criterion (d) only applies to studies that employ a cross-sectional design and verifies whether a study respects

the hierarchical structure of the educational data and controls for confounders at the macro-level. Table 1 presents the structure of the evaluation matrix.

Scores can range from 0 to a maximum of 6 per study. Here, we follow the general pattern found in other meta-analyses and hypothesize that *the higher the quality of the study, the smaller the tracking effects will be on efficiency* (H8a) *and inequality* (H8b). To test these hypotheses, we use the overall quality score but also design-specific items that focus on internal validity. We believe that if the first two criteria contain any sort of subjectivity, using only criteria (c) and (d) will largely eliminate it, offering a more objective test.

Year of publication. We aim to understand whether the size of the estimates reported by the studies under analysis follows any systematic trend over time, from 2000 to 2021.

Peer-review. Another potentially relevant feature is the publication status of the studies. In particular, we distinguish between peer-reviewed articles and all other types of products, such as policy reports and working papers. Although it can be assumed that peer-reviewed studies underwent more rigorous quality control, the literature highlights many issues in the current publication system in academic journals. Among the most serious concerns is the "file-drawer problem" (e.g., Rosenthal, 1979), namely, the fact that the system discourages the publication of studies that do not report statistically significant estimates. This occurs directly in the peer-review system, in which reviewers are more likely to approve papers with statistically significant findings, and indirectly via authors' reticence to submit to academic journals papers whose results provide no or weak support for their hypotheses. Thus, we anticipate that *peer-reviewed studies tend to report larger tracking estimates than policy and discussion reports both for efficiency* (H9a) *and inequality* (H9b). If supported, this will provide evidence of a publication bias in the current scientific publication system.

Method

Meta-analysis

Meta-analysis is characterized by clear-cut procedures and transparency. It offers an objective and mostly unbiased way to review a large body of literature on a certain topic and, especially in the social sciences, is a useful tool for standardizing results presented in the most diverse statistical forms to make them comparable. More specifically, we rely on random-effect meta-analysis, which is capable of systematizing study estimates by calculating effect sizes and can explain their variation based on study and policy features. This technique is particularly well suited when studies analyze heterogeneous populations and use different research designs, as is the case for the empirical literature on school tracking. Furthermore, estimates are weighted by the inverse of their squared variance, giving more weight to studies characterized by greater precision and statistical power (Lipson & Wilson, 2001; Siddaway et al., 2018). For our calculations, we used the "metafor" package (Viechtbauer, 2010), which is designed specifically to perform meta-analyses with the R software.

Search procedure

A systematic search in the Education Resources Information Center (ERIC) and Web of Science (WoS) databases was performed in September 2021 using three blocks of terms combined by the operator "AND": (1) educational-field terms ("educ*" OR "student*"); (2) tracking-related terms ("track*" OR "ability group*" OR "stream*" OR "sort*"); and (3) terms associated with identification strategies ("experiment" OR "quasi-experiment" OR "difference-in-difference*" OR "multilevel" OR "regression*" OR "propensity score" OR "instrumental variable"). This procedure yielded 3,655 results, which were filtered in a first-level screening by

scanning the publications' title and abstract to identify whether they concerned educational tracking and employed any sort of quantitative analysis. A total of 222 reports were assessed for eligibility, and 24 were retained. Figure 1 shows a PRISMA diagram describing the screening process and reasons for exclusion. To complement this initial sample, we snowballed Skopek et al. (2019), which was found to be the most recent non-systematic qualitative literature review on the topic. This procedure produced 168 additional reports for screening, of which 29 remained after the assessment based on the eligibility criteria. Therefore, the final sample is composed of 53 publications.

Selection criteria

All the selection criteria can be found in Appendix A. This set of criteria first aimed to detect a basic distinguishing feature: the central theme of the study should be school tracking in secondary education. As the independent variable, we considered "tracking" measured variously by, inter alia, the presence or absence of tracking (dummy, educational reforms), the age at first tracking, the number of tracks, and the length of tracking. One important reminder is that only studies that treated tracking as a systemic organizational variable were coded for the analysis. This excludes numerous studies on "within-school" tracking performed mainly in the United States, in which the individual-level effect of attending specific tracks is estimated, taking into account selection into different tracks or ability groups (Betts, 2001; Burks, 1994; Gamoran & Mare, 1989). These studies measure the individual impact of assignment to a certain track instead of the impact of tracking as a system-level policy, which is a linked but analytically distinct research question (Triventi et al., 2021).

As dependent variables, for educational efficiency, we selected a range of possible indicators of student performance: standardized proficiency tests (PISA, TIMSS, PIRLS), the

number of correct answers in a test, and non-standardized tests.[1] For achievement inequality, we considered measurements of dispersion (e.g., standard deviation, interquartile range, the share of top/bottom performers, etc.). For inequality of opportunity, we essentially considered the interaction between tracking and the family-background effect.

We reviewed only studies written in English and available online. The publications were mainly of two types: scientific journal articles and policy/discussion reports. At first, we chose to include all studies published since 1982, the year of publication of the last extensive meta-analysis on the topic, performed by Kulik and Kulik. Nonetheless, we eventually settled on a final sample of only one study from 1992; the rest was published from 2000 onward. We believe that this is due to the use of research designs such as multilevel analysis and difference-in-differences, which have gained popularity in the social sciences in the last two decades. Therefore, we adjusted the time span to include only studies from 2000 onward, excluding Hoffer (1992). Other examples of exclusion concern studies that do not cover the population of interest (secondary school students) but focus on primary school students (e.g., Duflo et al., 2011; Sterbinsky et al., 2006), studies that measure the impact of attending a specific track (e.g., Livingston, 2010; Pop-Eleches & Urquiola, 2011; Van Houtte et al., 2013) instead of the overall effect of tracking on the educational system, studies whose dependent variable was not cognitive achievement or student performance but, for instance, non-cognitive skills or educational transitions (e.g., Elfers, 2011; Guyon et al., 2012; Sampermans et al., 2021), and studies that

[1] IQ was reported as a dependent variable in only one study: Koerselman (2013). This represents only two statistically non-significant estimates for efficiency. Although IQ can express the ability of a student to learn, justifying its use only as an independent variable, Kenneth et al. (2015) contend that it can vary over time depending on the amount of education received and the social environment. We considered this debate to be unresolved and respected the author's choice to treat IQ as a dependent variable.

address different research questions and/or employ a different analytical design (Banerjee, 2017; Beattie, 2017; Mickelson & Everett, 2008).

Mean effect-size calculations

The studies that analyzed the effects of tracking, in general, applied various forms of regression models and reported β coefficients and standard errors for these estimates. However, the statistical theory for meta-analysis was mainly developed for experimental designs that presuppose independent treatment and control groups and the reporting of means and standard deviations for each. Even though Lipsey and Wilson (2001) consider that, when a non-experimental design uses a dummy variable as an independent variable, this could be interpreted as the mean difference between two groups (i.e., *tracked system = 1*), this might not be true when continuous variables such as the age at first tracking or the number of tracks are employed. Thus, the calculation of effect sizes here considers the effect of tracking for one unit, be it a dummy or a continuous variable. Consequently, to interpret the results, we must bear in mind that they measure the difference between tracked and comprehensive systems but also between tracked and less-tracked systems. All multivariate models presented in the results section are adjusted for whether tracking is operationalized as a dummy or a continuous variable.

The formulas used to calculate effect sizes are displayed in Appendix B and were taken mainly from Borenstein et al. (2009), who propose specific calculations for non-experimental designs. Two blocks of formulas for effect-size calculations are shown in two sections of Appendix B, (a) for studies that reported already standardized regression coefficients and (b) for studies that reported unstandardized regression coefficients.

A potential limitation of retrieving β estimates from different multivariate-regression models for meta-analysis is comparability. Borenstein et al. (2009) argue that because the set of controls might vary from one model to the next, the magnitude of the effect would depend on what is being controlled for. We decided to always take the fullest model specification presented by each study, that is, the model with the most controls. Furthermore, we also theorized the mechanisms through which tracking works to account for this problem. We classified studies according to whether they control for crucial intervening variables in the relationship between tracking and student outcomes, such as school/class composition and educational resources.

We elected Hedge's G as our measure to quantify the mean effect size, which is a corrected version of Cohen's D for samples that are too small ($n < 20$). Since some studies use small samples of countries (sometimes as little as eight countries), we correct for this upward Cohen's D bias (Borenstein et al., 2009). The signs of the retrieved β estimates were consistently coded for a positive sign to indicate that tracking improves student achievement and increases inequality. Hedge's G unit offers an intuitive interpretation as it measures effect sizes in the form of standardized mean differences between groups, which can be read as standard deviations.

Coding of studies

It is important to stress that in a meta-analysis, one publication does not always equal one study. For instance, a publication may run the same model for different outcome variables, such as "efficiency" and "inequality." In this case, we ran separate analyses for each of these dimensions. Furthermore, a publication can be split into two or more studies when the sample changes. Here, we follow the exact definition proposed by Lipsey and Wilson (2001, p. 76), which states that a study is "a set of data analyzed under a single research plan from a designated [specific] sample of respondents." This means that one study does not correspond to one

publication but to a unique combination of author and sample. Since meta-analysis also requires binding different populations together for generalizability, by doing this, we do not treat different populations as equal simply because they are included in the same publication. Furthermore, we gain in sample size at the macro-level (number of studies) and, consequently, in degrees of freedom, not restricting our main model excessively. Thus, our sample of 53 publications provides 213 estimates for efficiency that can be grouped into 55 studies, as well as 230 estimates for inequality that can be grouped into 46 studies. Table 2 provides the descriptive statistics.

Both authors scanned all the results retrieved during the search procedure, screened all studies identified based on the eligibility criteria, and reviewed all 53 articles to build the dataset. We reached 85% agreement on the coding process, favoring the main author's coding in cases of disagreement. Table S.1 in the online Supplemental Material contains the references of all 53 coded publications.

Results

Publication bias

A common concern in meta-analyses is publication bias (Lipson & Wilson, 2001; Borenstein et al., 2009). This results from the well-known trend whereby studies that report more statistically significant estimates have a higher probability of being published.

To check for publication bias, we plotted the effect sizes on a funnel plot where the x-axis represents Hedges' G effect size and the y-axis represents the standard error for each effect (see Figure 2 for efficiency and Figure 3 for inequality). Both outcomes revealed asymmetry. In general, this could indicate that the selection of studies for the meta-analysis concentrated on published articles and that cases from unpublished papers were missing (Borenstein et al., 2009, p. 277).

In the present meta-analysis, we rely on a more balanced sample because 100 of the 213 estimates for efficiency (46.9%) and 66 of the 230 estimates for inequality (28.7%) are taken from working and discussion papers, policy reports, book chapters, or doctoral dissertations, that is, works that are not peer-reviewed. Furthermore, we can address this concern later in our meta-regression models using the variable "peer-review" to gather additional insight into the strength of this bias.

Mean effect sizes

Educational efficiency: level of student achievement. Table 3 reports the mean effect size of tracking on the average student's achievement. The effect size is not statistically significant at the 95% confidence level, providing support for the hypothesis that tracking does not increase the overall efficiency of an educational system (H1a).[2] However, this result must be assessed together with the results on inequality, since a 0 value could be due to the fact that students in the higher tracks improve their results, but this improvement is canceled out by the deterioration of performance in the lower tracks. Another explanation for the null effect of tracking might be related to the fact that many studies use international standardized tests such as PISA, which restricts the analysis to student outcomes in grade 9 (15 years old), when tracking starts in some countries. In this case, for many country units in the sample, the effects of tracking have not had time to fully develop yet.

Heterogeneity tests (Cochran's Q) returned statistically significant results, indicating that the studies included in the analysis contain too much variation and, thus, cannot be treated as belonging to the same population (Hak et al., 2016; Borenstein et al., 2009). This test provides the first evidence that, for the studies included in this meta-analysis, a random-effect analysis would be more appropriate than a fixed-effect one.

The tau^2 parameter provides the magnitude of this variation, such that if we were to draw infinite samples of studies, their variance would be .116 on average. Tests for the detection of

[2] Although some of the hypotheses outlined in the theoretical section are directional, we opted to rely on nondirectional statistical tests of the null hypothesis in the empirical analysis, for several reasons: 1) for some issues, there are plausible theoretical predictions in both directions (Woessman, 2009); 2) for other aspects, there is little previous theoretical elaboration, and our hypotheses are more explorative; 3) given that the critical rejection values are higher in a nondirectional test than in a directional test, this is a more conservative approach and the most commonly used in practice (Pillemer, 1991; Salkind, 2010); 4) this choice allows us to maintain homogeneity in the type of statistical test used throughout the paper.

outliers revealed four out-of-bound estimates for student achievement (see Supplemental Table S.2, available in the online version of the journal, for a graphic representation of the outlier tests performed). We run a model without outliers to determine how much the magnitude of the effect size would change; it gets closer to zero and remains statistically non-significant. I^2 is the proportion of between-study variance that reflects real differences among effect sizes, namely, the amount of variation that cannot be explained by chance only. Hence, a high value of I^2 (>25%) means that there is enough non-random variation to be explored in a multivariate meta-regression at a second stage (Hak et al., 2016). Model 2 reports a high value of I^2 (99.6%), justifying the multivariate analysis in the results section.

Model 3 applies multilevel techniques to account for the hierarchical structure of the data and the non-independence of estimates taken from the same studies. The 213 estimates of the effect of tracking on educational efficiency (level 1) were nested within 55 studies (level 2). The estimate of the mean effect size remains statistically non-significant. Thus, all in all, the results of three different models indicate that the effect of tracking on educational efficiency (i.e., whether overall student achievement is greater in a tracked system), oscillates around 0. This does not mean that the variation in results cannot be explained by moderators but merely that tracking studies report neutral, positive, and negative estimates in a rather balanced way.

Inequality in educational achievement. The same strategy was used to assess the mean effect size of tracking on student-achievement inequality (Table 4). This estimate is positive in sign and statistically significant at the 99.9% confidence level (in all specifications), providing support for the hypothesis that tracking increases the level of inequality in the educational system (H1b).

Model 1 reports a statistically significant estimate of .097 standard deviation (SD) increase in educational inequality. After excluding three outliers (see Supplemental Table S.3

available online), model 2 reveals a smaller effect of .076 SD. When grouping the 230 estimates into 46 studies in model 3, the mean effect size of tracking on inequality grows to .117 SD (99.9% CI: .008; .225). The mean effect size increases after grouping estimates into studies because all studies now operate with equal, balanced weight in the calculations, which are, thus, not biased by the number of estimates that each study reports. Even though the effect size is statistically significant with a 99.9% CI, it can be considered small on a meta-analytic scale (Cohen, 1988).[3]

Still, it is worth considering whether its magnitude has any practical social significance (Bernardi et al., 2017). Woessmann (2016) reported that in the PISA scale, for instance, students are capable of progressing, on average, from 0.25 to 0.33 SD in a school year. If all studies in our sample were to use PISA scores (and the majority does), this would mean that tracked systems exacerbate existing inequalities in student achievement from a third to almost half of a school year. Whether this increase is remarkable or not will depend on the already existing levels of inequality in elementary and primary schools. The top four countries in inequality growth between primary (PIRLS 2001) and secondary education (PISA 2003) reported by Hanushek and Woessmann (2006, p. 69) are Germany (0.71 SD), Greece (0.3 SD), the Czech Republic (0.25 SD), and Italy (0.22 SD), all considered early trackers by the authors. Thus, an increase of .117 SD for the whole variety of tracking systems reported in the literature (not only early trackers) seems to be relevant and capable of explaining much of this increase. This magnitude can explain, for example, the whole increase in achievement inequality in the Netherlands after tracking occurs.

[3] Cohen (1988) has proposed that effect sizes should be considered small between .0 and .29, medium between .3 and .49, and large above .5.

Multivariate meta-regressions

The following multivariate meta-regressions set out to explain the variation in effect sizes observed in the population of studies by a set of explanatory variables related to the tracking policy and the features of the studies.

We specified four statistical models with stepwise inclusion of independent variables in addition to the "intercept-only" models provided in the previous section. The first specification (model 1) in Tables 5 (efficiency) and 6 (inequality) include the independent variables related to the policy characteristics and aspects of substantive interest (i.e., type of tracking, subject of student competencies, and reform analysis). Model 2 adds the variables related to the methodological aspects of the research designs (i.e., counterfactual reference category, identification strategy, and whether the study controls for potential mediators). Model 3 adds the overall score for publication quality, and model 4 tests for the year of publication, peer-review, and quality of the research design.

Educational efficiency. Model 1 in Table 5 reveals a statistically significant positive effect of between-school tracking compared to within-school tracking. This indicates that studies focusing exclusively on measuring the impact of between-school tracking tend to report estimates .023 SD larger than studies of within-school tracking (see Table S.4 in the online Supplemental Material for an additional analysis considering only between-school tracking estimates). This confirms our expectation that between-school tracking can increase overall student achievement more than within-school tracking (H2a).

Model 1 also checks whether achievement in one specific cognitive domain (reading, mathematics, or science) is more sensitive to tracking. The conditional differences across subject

domains are not statistically significant and are small in size. This does not support the hypothesis that mathematics performance is more sensitive to the effects of tracking (H3a). Even though not statistically significant, studies that use pooled scores across domains tend to report effects much smaller in magnitude (-.047) than studies focusing exclusively on mathematics. The estimates related to reading, on the contrary, are slightly larger (.016) than those related to mathematics.

In Model 1, we also investigate whether studies concentrating on the impact of national reforms tend to report smaller estimates of the effects of tracking on student achievement. The effect size is not statistically significant, which does not confirm the hypothesis that studies of national educational reforms tend to report smaller effects of tracking (H4a).

Model 2 includes methodological and research-design characteristics. The choice of the counterfactual situation (reference category) in a study appears to influence the reported effect of tracking but in the opposite direction than expected. Studies that use "comprehensive" systems as a reference (instead of "less-tracked" systems) tend to report estimates .071 SD smaller, contradicting hypothesis H5a. Looking at the analytical strategy, when a study applies a difference-in-differences design, it tends to report smaller estimates (-.041 SD) than those using cross-sectional multilevel strategies, thus providing support for hypothesis H6a. This result could suggest that this analytical design is more effective in ruling out confounders that may inflate the size of tracking estimates. This estimate remains significant until model 4, increasing its magnitude to -.154.

Model 2 further indicates that studies adopting an extensive model specification that controls for educational resources report on average -.022 SD smaller effects on the level of student achievement. Studies that control both for educational resources and school/class

composition also produce smaller estimates than studies that control for neither of these mechanisms. However, studies that control only for school/class composition do not report significantly smaller effect sizes, suggesting that a major mechanism of the tracking policy works through the unequal distribution of educational resources across different tracks and/or classrooms. These estimates provide support for hypothesis H7a.

Model 3 tests whether the reported effects of tracking on efficiency depend on the overall quality of the studies (i.e., the scores given to each study according to the evaluation matrix and presented in Table 1). Model 4 instead tests whether the effects of tracking are associated with the year of publication (from 2000 to 2021), the status of the study (peer-reviewed or not), and whether tracking effects depend on the quality of the research design (a sub-component of the overall quality score), which we believe to be the most objective item in the evaluation matrix and could isolate possible subjectivity contained in the other items. The results indicate that more recent studies report smaller estimates for the gain in achievement associated with tracking. This is in line with the literature that questions the gain in achievement due to specialization of tracked educational systems. Thus, the hypothesis that the higher the quality of the study, the smaller the effects will be (H8a) is contradicted as higher-quality studies report larger estimates. Furthermore, the hypothesis that peer-reviewed studies tend to report larger estimates (H9a) is not supported.

Educational inequality. The same models were applied almost identically to explain the variation in the effect sizes of tracking on student achievement inequality (Table 6). The only difference is the inclusion of a dummy variable that accounts for the fact that studies measured inequality in two ways, through the dispersion of student performance ("achievement inequality," used as reference category) or through the effect of family background on achievement ("inequality of opportunity").

Model 1 reports a significant negative effect of -.40 SD of between-school tracking on inequality. Thus, studies on the consequences of between-school tracking for inequality tend to report smaller estimates than studies focusing on within-school tracking, contradicting hypothesis H2b (see Table S.5 in the online Supplemental Material for an additional analysis considering only between-school tracking estimates). Nonetheless, the importance of informal and within-school forms of tracking for inequality of opportunities in education has been highlighted in some recent cross-national studies (Blossfeld et al., 2016; Triventi et al., 2020). A possible explanation for this result can be found in the stigmatization and self-fulfilling prophecy theory (Merton, 1968). Students that are placed into less-advanced classes still have to share the school environment with students in the more advanced classes, possibly creating a feeling of inferiority. Furthermore, teachers within the school deal with students who are labeled as less advanced beforehand, sometimes via decisions taken by the teachers themselves, possibly lowering teaching standards. Additionally, between-school tracking, in principle, is mostly related to the optimal allocation of students according to vocational, technical, and academic profiles to facilitate the school-to-work transition. Instead, ability grouping is mostly related to dividing students based on achievement, concentrating high-performers in more advanced courses.

Still, in model 1, similarly to the efficiency results, the effects of tracking on inequality appear to be similar across subject domains. This provides no support for the hypothesis that inequality in mathematics scores is more sensitive to the effects of tracking (H3b.1). Model 1 still reports a non-statistically significant estimate for studies that operationalized their measurement of inequality through the influence of family background on achievement, compared to measures of achievement dispersion. This does not confirm the hypothesis that the effects of tracking tend to be weaker on inequality of opportunity than on measurements of achievement dispersion (H3b.2).

Reform studies tend to report more conservative estimates for inequality in the magnitude of .05 SD, keeping its magnitude and significance until model 4. Therefore, the hypothesis that studies of national educational reforms tend to report smaller effects of tracking on inequality (H4b) is supported.

Model 2 contradicts the expectation that the effects of tracking tend to be larger when studies use a comprehensive educational system as their counterfactual (H5b). Moreover, this model does not corroborate the hypothesis that difference-in-differences designs tend to produce smaller estimates than cross-sectional multilevel designs (H6b) because the estimates are not statistically significant, even though their signs point in the hypothesized direction. Nevertheless, it provides evidence that the few studies that employed very simple identification strategies (i.e., one-level regression models) tend to report larger estimates for tracking effects, ranging from .242 (model 2) to .305 SD (model 4). Also, no support is provided for the hypothesis that when a study controls for educational resources and/or school/class composition indicators, the effect sizes of tracking tend to be smaller for inequality (H7b).

Model 3 presents a significant positive estimate for the overall quality of the study. Better-evaluated studies tend to report estimates .084 SD larger, contradicting our expectation that the higher the quality of the study, the lower the estimates. In model 4, when we test this hypothesis, substituting the overall evaluation score for its sub-component focused on the quality of the research design, the estimate increases to .094 and is still significant at the 95% CI. An explanation is that the quality of the study and the peer-review process could be intertwined. Articles that are published tend to be evaluated as higher quality. This would once again lend credence to the file-drawer problem, contradicting hypothesis H8b (i.e., the higher the quality of the study, the smaller the effects) and not supporting hypothesis H9b (i.e., peer-reviewed studies tend to report larger estimates). Table S.5 in the online Supplemental Material gives an overview of the testing of all the hypotheses.

Discussion

We analyzed the effect of tracking on two important outcomes of contemporary educational systems: efficiency and inequality. By "efficiency," we mean the overall (average) level of student achievement, whereas "inequality" refers to achievement dispersion (an indicator of interpersonal inequality) and inequality of opportunity (measuring the strength of the relationship between social background and achievement). Our main finding, which we obtained by applying meta-analytical techniques to studies from the last two decades, is that the mean effect size of tracking on efficiency is null, whereas it is positive for inequality. This evidence provides no support for the existence of an "equality-efficiency trade-off" (Skopek et al., 2019, p. 224), that is, the need to sacrifice equality to improve the overall performance of the educational system. Instead, this finding suggests that the stream of literature that emphasizes the role of tracking in enhancing both student achievement dispersion and inequality of opportunity relies on more solid empirical evidence than the theoretical arguments suggesting that tracking increases efficiency. Thus, our results indicate that de-tracking reforms, which postpone tracking, reduce the number of tracks, or smooth out the distinctions across tracks, have the potential to reduce inequality in educational opportunities based on social background without harming overall student achievement. Of course, our finding only considers cognitive-related achievement, and it should be complemented with considerations about the labor-market effects of school tracking and the institutional specificities of each country (Shavit & Muller, 1998). Moreover, future studies should start considering the possible effects of tracking on students' socioemotional development, a clear gap in the tracking literature.

As regards the choice of tracking policy, the between-school modality seems to be more effective in terms of specialization while producing less inequality than the within-school type,

but a caveat is warranted. Van de Werfhorst (2019) analyzed educational reforms in 21 European countries and found that de-tracking educational reforms reduce inequality of opportunity at the expense of lowering the achievement of more advantaged students. This means that there could be a risk that children from higher socioeconomic backgrounds will face deterioration in their learning progress if a de-tracking reform is implemented. Alternative policies should be put in place to avoid this drawback, such as stimulating students to take on a more proactive role, for example, by making them tutor low-performing students so that they may not be affected negatively (Hattie, 2009).

Important methodological remarks concern the fact that longitudinal designs, in general, report more conservative estimates than cross-sectional multilevel designs. This was confirmed for studies that analyzed the effect of educational reforms on inequality and studies that used a difference-in-differences design to identify the effects of tracking on efficiency. Multilevel techniques, for their part, are capable of correcting for part of the bias when compared to standard one-level multivariate regressions. Additionally, we provided evidence from two different tests that publication bias is present in the tracking literature and needs to be addressed by the scientific publication system to minimize the file-drawer problem.

Another contribution of this meta-analysis is theoretical. The debate in the field of sociology of education has moved from an input-output black-box type of research to one that investigates the mechanisms through which macro-systemic variables work. We found evidence that the presence of a specialization effect of tracking on efficiency depends on the type of controls included in the analysis (i.e., educational resources, school/class composition, or both). This reveals a need for more exchanges between the literature on tracking and the literature discussing the distribution of educational resources between schools as well as the effects of

school/class composition and peers on student achievement. Tracking likely works partly through these factors, and its effects should not be discussed without taking into account these mechanisms.

Lastly, our analysis showed no evidence of a differential effect of tracking across subjects. This probably would require more statistical power. Another limitation of this meta-analysis is that it focuses on the short-term effects of tracking on cognitive outcomes only, that is, how tracking affects educational achievement during the school phase, ignoring the transition to higher education and the job market. Future research should also concentrate on systematizing the effects of tracking on longer-term outcomes when more studies are available. This would expand the debate to other relevant questions, such as whether it is worth tolerating some level of short-term inequality if the professional placement of students is assured later, and whether the inequalities reproduced and intensified by tracked educational systems remain in access to higher education and higher-status positions in the labor market. Moreover, future investigations should look at the effects of various forms of tracking on students' socio-emotional competencies and civic engagement, which have been covered by a limited number of studies (e.g., Witschge & van de Werfhorst, 2020; Korthals et al., 2021; Österman, 2021) even though they are increasingly considered important for individuals' success in life (OECD 2015). Lastly, the literature seems to have somehow overlooked the economic aspects related to tracking policies. The financial cost of maintaining a tracked system instead of a comprehensive one is rarely discussed. Taking into consideration all these aspects is crucial to improving our overall understanding of how educational arrangements work and helping policymakers make informed decisions based on solid empirical evidence.

REFERENCES

References

Ammermüller, A. (2005). *Educational opportunities and the role of institutions* (ZEW Discussion Paper No. 05-44). Retrieved from the Econstor website: https://www.econstor.eu/bitstream/10419/24135/1/dp0544.pdf

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press. doi:10.2307/j.ctvcm4j72

Ariga, K., & Brunello, G. (2007). *Does Secondary School Tracking Affect Performance? Evidence from IALS* (Discussion paper n. 630). Retrieved from the Research Gate network: https://www.researchgate.net/publication/5136980_Does_Secondary_School_Tracking_Affect_Performance_Evidence_from_IALS

Ayalon, H. (2006). Nonhierarchical curriculum differentiation and inequality in achievement: A different story or more of the same? *Teachers College Record, 108*(6), 1186-1213. doi:10.1111/j.1467-9620.2006.00690.x

Ayalon, H., & Gamoran, A. (2000). Stratification in academic secondary programs and educational inequality in Israel and the United States. *Comparative Education Review 44*(1), 54-80. doi:10.1086/447591

Banerjee, N. (2017). Student–Teacher Ethno-Racial Matching and Reading Ability Group Placement in Early Grades. *Education and Urban Society*. V. 51 (3), pp.: 395-422. doi:10.1177/0013124517721948

REFERENCES

Barg, K. (2013). The influence of students' social background and parental involvement on teachers' school track choices: Reasons and consequences. *European Sociological Review, 29*(3), 565–579. doi:10.1093/esr/jcr104

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

Beattie, I. R. (2014). Tracking Women's Transitions to Adulthood: Race, Curricular Tracking, and Young Adult Outcomes. *Youth and Society*. First Published March 20, 2014. doi:10.1177/0044118X14527467

Bernardi, F., Chakhaia, L., & Leopold, L. (2017). 'Sing me a song with social significance': The (mis)use of statistical significance testing in European sociological research. *European Sociological Review, 33*(1), 1-15. doi:10.1093/esr/jcw047

Betts, J. R. & Grogger, J. (2003). The impact of grading standards on student achievement, educational attainment, and entry-level earnings. Economics *of Education Review, 22*(4), 343-352. doi:10.1016/S0272-7757(02)00059-6

Betts, J. R. (2001). The economics of tracking in education. In Hanushek, Eric A., Machin, S. & Woessmann, L. (Eds.), *Handbook of the economics of education,* Vol. 3 (pp. 341-381*),* Amsterdam, The Netherlands: North-Holland.

Betts, J. R., & Shkolnik, J. L. (2000). The effects of ability grouping on student achievement and resource allocation in secondary schools. *Econ. Educ. Rev. 19*(1), 1-15. doi:10.1016/S0272-7757(98)00044-2

REFERENCES

Blossfeld, H. P., Bucholz, S., Skopek, J., & Triventi, M. (2016). *Models of secondary education and social inequality: An international comparison.* Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing. doi:10.4337/9781785367267

Bol, T., & Van de Werfhorst, H. (2013). The measurement of tracking, vocational orientation, and standardization of educational systems. Comparative Education Review, 57(2), 285-308.

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), 1545-1573. doi:10.1093/sf/sou003

Bonesronning, Hans & Falch, Torberg & Strom, Bjarne, 2005. Teacher sorting, teacher quality, and student composition. *European Economic Review*, *49*(2), 457-483. doi:10.1016/S0014-2921(03)00052-7

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein H. R. (2009). *Introduction to Meta-Analysis.* West Sussex, UK: John Wiley & Sons Ltd. doi:10.1002/9780470743386

Bourdieu, P., & Passeron, J. C. (1970). *La Reproduction.* Paris : Éditions de Minuit.

Brantlinger, A. (2013). Between politics and equations: Teaching critical mathematics in a remedial secondary classroom. *American Educational Research Journal, 50(5), 1050-1080.* doi:10.3102/0002831213487195

Breen, R., & Goldthorpe, J. H. (1997). Explaining educational differentials: Towards a formal rational action theory. *Rationality and Society, 9*(3), 275-305. doi:10.1177/104346397009003002

REFERENCES

Brunello, G., & Checchi, D. (2007). Does school tracking affect equality of opportunity? New international evidence. *Economic Policy, 22*(52), 781–861. doi:10.1111/j.14680327.2007.00189.x

Burger, K. (2016). Intergenerational transmission of education in Europe: Do more comprehensive education systems reduce social gradients in student achievement? *Research in Social Stratification and Mobility*, *44*, 54–67. doi:10.1016/j.rssm.2016.02.002

Burks, L. C. (1994). Ability group level and achievement. *School Community Journal, 4*(1), 11-24.

Clifford, P., & Heath, A. (1984). Selection does make a difference. *Oxford Review of Education, 10*(1), 85-97. doi:10.1080/0305498840100108

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cohn, L. D. (1991). Sex differences in the course of personality development: A meta-analysis. *Psychological bulletin*, *109*(2), 252-266. doi:10.1037/0033-2909.109.2.252

Conn, K. M. (2017). Identifying effective education interventions in Sub-Saharan Africa: A meta-analysis of impact evaluations. *Review of Educational Research,87*(5), 863-898. doi:10.3102/0034654317712025

Contini, D., & Scagni, A. (2011). Inequality of opportunity in secondary school enrolment in Italy, Germany and the Netherlands. *Quality & Quantity: International Journal of Methodology, 45*(2), 441-464. doi:10.1007/s11135-009-9307-y

Cook, R. R. (1924). A study of the results of homogeneous grouping of abilities in high-school classes. In Whipple, G. M. (Ed.) *The education of gifted children* (pp. 302-12). Bloomington, Ill.

REFERENCES

Cooper, H., Nye, B., Charlton, K., & Lindsay, J. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research, 66*(3), 227–268. doi:10.2307/1170523

Deci, E. L, Koestner, R., & Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research, 71*(1), 1-27. doi:10.3102/00346543071001001

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

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review*, *101*(5), 1739–1774. Doi: 0.1257/aer.101.5.1739

Dunne, A. (2010). *Dividing lines: Examining the relative importance of between- and within-school differentiation during lower secondary education* (Doctoral dissertation). Retrieved from the European University Institute website: https://cadmus.eui.eu/handle/1814/14497.

Dupriez, V., Dumay, X., &Vause, A. How do school systems manage pupils heterogeneity? *Comparative Education Review 52*(2), 245-273. doi:10.1086/528764

Dustmann, C. (2004). Parental background, secondary school track choice, and wages. *Oxford Economic Papers*, *56*(2), 209-230. doi:10.1093/oep/gpf048

Esser, H. (2016). The model of ability tracking–Theoretical expectations and empirical findings on how educational systems impact on educational success and inequality. In *Models of secondary education and social inequality*. Edward Elgar Publishing.

REFERENCES

Felouzis, G., & Charmillot, S. (2013). School tracking and educational inequality: A comparison of 12 education systems in Switzerland. *Comparative Education, 49*(2), 181-205. doi:10.1080/03050068.2012.706032

Galindo-Rueda, F., & Vignoles, A. F. (2005). The heterogeneous effect of selection in secondary schools: Understanding the changing role of ability (IZA Discussion Paper No. 1245). Retrieved from the Ideas website: https://ideas.repec.org/p/cep/ceedps/0052.html

Gambetta, D. (1987). *Studies in rationality and social change. Were they pushed or did they jump? Individual decision mechanisms in education.* Cambridge University Press.

Gamoran, A., & Mare, R. D. (1989). Secondary school tracking and educational inequality: Compensation, reinforcement, or neutrality? *American Journal of Sociology 94*(5), 1146-1183. doi:10.1086/229114

Gamoran, A., Berends, M. The Effects of Stratification in Secondary Schools: Synthesis of Survey and Ethnographic *Research Review of Educational Research*, vol. 57, 4: pp. 415-435. , First Published Dec 1, 1987. doi:10.3102/00346543057004415

Guyon, N., Maurin, E., & Mcnally, S. (2012). *The Effect of Tracking Students by Ability into Different Schools A Natural Experiment*. doi:http://statline.cbs.nl/Statweb/publication

Hak, T., Van Rhee, H. J., & Suurmond, R. (2016). *How to interpret results of meta-analysis* (Version 1.0) Rotterdam, The Netherlands: Erasmus Rotterdam Institute of Management. doi:10.2139/ssrn.3241367

Hallinan, M. T. (1994). School differences in tracking effects on achievement. *Social Forces, 72*(3), 799–820. doi:10.1093/sf/72.3.799

REFERENCES

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

Hoffer, T. B. (1992). Middle school ability grouping and student achievement in science and mathematics. *Educational Evaluation Policy Analysis, 14*(3), 205-227. doi:10.3102/01623737014003205

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*, 25-43. doi:10.1016/j.rssm.2013.01.002

Horn, D., Balázsi, I., Takács, S., & Zhang, Y. (2007). Tracking and inequality of learning outcomes in Hungarian secondary schools. *PROSPECTS. 36, 433-446.* doi:10.1007/s11125-006-9003-9

Huang, M-H. (2009). Classroom homogeneity and the distribution of student math performance: A country-level fixed-effects analysis. *Social Science Research*, *38*,81–791. doi:10.1016/j.ssresearch.2009.05.001

Jackson, M. (2013). *Determined to succeed? Performance versus choice in educational attainment*. Stanford: Stanford University Press. doi:10.1515/9780804784481

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.* (pp. 41-81). Springer Science. doi:10.1007/978-90-481-3993-4

Keliher, A. V. (1931). *A critical study of homogeneous grouping, with a critique of measurement as the basis for measurement.* New York, NY: Teachers College Press.

REFERENCES

Kendler K. S., Turkheimer, E., Ohlsson, H., Sundquist, J., & Sundquist, K. (2015). Family environment and the malleability of cognitive ability: A Swedish national home-reared and adopted-away co-sibling control study. *PNAS* 112 (15) pp. 4612-4. doi:10.1073/pnas.1417106112

Kerr, S. P., Pekkarinen, T., & Uusitalo, R. (2013). School tracking and development of cognitive skills. *Journal of Labour Economics*, *31*(3), 577–602. doi:10.1086/669493

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

Korthals, R., Schils, T., & Borghans, L. (2021). Track placement and the development of cognitive and non-cognitive skills. *Education Economics*, 1-20.

Kulik, C.-L. C., & Kulik, J. A. (1982). Effects of ability grouping on secondary school students: A meta-analysis of evaluation findings. *American Educational Research Journal, 19*(3), 415–428. doi:10.3102/00028312019003415

Lareau, A. (1987). Social class differences in family-school relationships: The importance of cultural capital. *Sociology of Education, 60*(2), pp. 73-85. doi:10.2307/2112583

Lassibile, G., & Gómez, L. N. (2000). Organization and efficiency of education systems: Some empirical findings. *Comparative Education, 36*(1), 7-19. doi:10.1080/03050060027737

Le Donné, N. (2014). European variations in socioeconomic inequalities in students' cognitive achievement: The role of educational policies. *European Sociological Review, 30*(3), 1–15. doi:10.1093/esr/jcu040

Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-Analysis.* Thousand Oaks, CA: Sage Publications, Inc.

REFERENCES

Lucas, S. R. (1999). *Tracking inequality: stratification and mobility in American high schools*. New York: Teachers College Press.

Lucas, S. R. (2001) Effectively maintained inequality: Education transitions, track mobility, and social background effects. *American Journal of Sociology 106*(6):1642-90. doi:10.1086/321300

Manning, A., & Pischke, J.S. (2006). *Comprehensive versus selective schooling in England in Wales: What do we know?'* (IZA Discussion Paper No. 2072). doi:10.3386/w12176

Marín-Martínez, F., & Sánchez-Meca, J. (2010). Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educational and Psychological Measurement, 70*(1), 56–73. doi:10.1177/0013164409344534

Merton, R. (1968). The self-fulfilling prophecy. In *Social Theory and Social Structure.* New York: The Free Press.

Mickelson, R. A., & Everett, B. J. (2008). Neotracking in North Carolina: How high school courses of study reproduce race and class-based stratification. *Teachers College Record, 110*(3), 535–570. Doi: https://psycnet.apa.org/record/2009-00714-001

Oakes, J. (1985). *Keeping track: How schools structure inequality.* New Haven: Yale University Press.

OECD (2015), *Skills for Social Progress: The Power of Social and Emotional Skills*, Paris: OECD Publishing.

Österman, M. (2021). Can We Trust Education for Fostering Trust? Quasi-experimental Evidence on the Effect of Education and Tracking on Social Trust. *Social Indicators Research*, *154*(1), 211-233.

Pearl, J. (2009). *Causality*. Cambridge University Press. doi:10.1017/CBO9780511803161

REFERENCES

Pfeffer, F. T. (2008). Persistent inequality in educational attainment and its institutional context. *European Sociological Review*, *24*(5): 543-565. doi:10.1093/esr/jcn026

Pillemer, D. B. (1991). One-versus two-tailed hypothesis tests in contemporary educational research. Educational Researcher, 20(9), 13-17.

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

Roloff, J., Klusmann, U., Lüdtke, O. &Trautwein U. (2020) The Predictive Validity of Teachers' Personality, Cognitive and Academic Abilities at the End of High School on Instructional Quality in Germany: A Longitudinal Study. *AERA Open*, vol. 6, 1, First Published January 8. doi: 10.1177/2332858419897884

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin, 86*(3), 638-641. doi:10.1037/0033-2909.86.3.638

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The urban review*, *3*(1), 16-20. doi:10.1007/BF02322211

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

Salkind, N. J. (2010). *Encyclopedia of research design* (Vols. 1-0). Thousand Oaks, CA: SAGE Publications, Inc. doi: 10.4135/9781412961288

Sampermans D., Claes, E. & Janmaat J. G. (2021). Back on track? How civic learning opportunities widen the political knowledge gap in a tracked education system. *School Effectiveness and*

REFERENCES

School Improvement An International Journal of Research, Policy and Practice V. 32 (2). doi:10.1080/09243453.2020.1830125

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

Schütz, G., Ursprung, H., & Woessmann, L. (2008). Education policy and equality of opportunity. *Kyklos*, *61*(2), 279–308. doi:10.1111/j.1467-6435.2008.00402.x

Shavit, Y., & Müller, W. (1998). *From school to work: A comparative study of educational qualifications and occupational destinations.* Oxford, UK: Oxford: University Press.

Siddaway, A., Wood, A., & Hedges, L. (2019). How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology. 70*, 747-770. doi:10.1146/annurev-psych-010418-102803

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 Becker, R. (Ed.), *Research Handbook on the Sociology of Education* (pp. 214-232). Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing. 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, vol. 60, 3: pp. 471-499. , First Published Sep 1, 1990. doi:10.3102/00346543060003471

Snijder, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel Modeling*. SAGE Publications London Thousand Oaks New Delhi.

REFERENCES

Triventi, M., Skopek, J, Kulic, N., Buchholz, S., & Blossfeld, H. P. (2020). Advantage 'finds its way': How privileged families exploit opportunities in different systems of secondary education. *Sociology, 54*(2), 237–257. doi:10.1177/0038038519874984

Van de Werfhorst, H. G. (2019). Early tracking and social inequality in educational attainment: Educational reforms in 21 European countries. *American Journal of Education 126*(1), 65-99. doi:10.1086/705500

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., Demanet, J. & Stevens, P.A.J. (2013). Curriculum tracking and teacher evaluations of individual students: selection, adjustment or labeling?. Soc Psychol Educ 16, 329–352 (2013). doi:/10.1007/s11218-013-9216-8

Vandenberge, V. (2006). Achievement effectiveness and equity: the role of tracking, grade repetition and inter-school segregation. *Applied Economics Letters,*13(11), 685-693. doi:10.1080/13504850500404944

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1-48. doi:10.18637/jss.v036.i03

Waldinger, F. (2007). Does ability tracking exacerbate the role of family background for students' test scores? (mimeo).

Whipple, G. M. (1936). *The Grouping of Pupils* (35, pt. 1). Bloomington, Ill.: Public School Publishing, National Society for the Study of Education Yearbook,

REFERENCES

Witschge, J., & van de Werfhorst, H. G. (2020). Curricular tracking and civic and political engagement: Comparing adolescents and young adults across education systems. *Acta Sociologica*, *63*(3), 284-302.

Woessmann, L. (2009). *International evidence on school tracking: A review*. CESifo DICE Research Report 1. Retrieved from https://www.ifo.de/DocDL/dicereport109-rr1.pdf

Woessmann, L. (2009). International evidence on school tracking: A review. CESifo DICE Report, 7(1), 26-34.

Woessmann, L. (2016). The importance of school systems: Evidence from international differences in student achievement. *Journal of Economic Perspectives. 30*(3)*,* 3-32. doi:10.1257/jep.30.3.3

Zimmer, R. (2003). A new twist in the educational tracking debate. *Economics of Education Review.* *22*(3)*,* 307-315. doi:10.1016/S0272-7757(02)00055-9

Table 1

Evaluation matrix [a]

	Criteria	Research design	0 point	1 point	2 points
1	Presentation of results	All	Poor: unclear language, incomplete tables, crucial information missing (sample size, SE or SD, etc.).	Adequate: although not very well textually presented, results are clear in the tables, or vice-versa.	Professional: results presented in clear language with complete and self-explaining tables, graphs, figures, etc.
2	Presentation of research design	All	Vague explanation of procedures and limitations not discussed.	Fair explanation of procedures and/or some limitations discussed.	Adequate explanation of procedures and limitations discussed.
3	Trend/ balances	Difference-in-differences	Does not check for previous trend and does not control for balancing variables.	Does not check for previous trend but controls for balancing variables.	Checks for previous trend and controls for balancing variables.
4	Macro-level heterogeneity	Cross-sectional models	Does not account for macro-level unobserved heterogeneity or applies one-level regression.	Applies multilevel regression and controls for at least one macro-level confounder (e.g., GDP).	Applies multilevel and controls for more than one macro-level confounder (e.g., GPD and investment per student).

[a] *Adapted from Conn (2017).*

TABLES

Table 2 *Descriptive Statistics*

Variables	Efficiency					Inequality				
	N.	%	Mean	Min.	Max.	N.	%	Mean	Min.	Max.
Policy characteristics										
Within school (ref.)	35	16.4%	0.16	0	1	25	10.9%	0.11	0	1
Between school	178	83.6%	0.84	0	1	205	89.1%	0.89	0	1
Achiev. Dispersion (ref.)	-	-	-	-	-	86	37.4%	0.37	0	1
Opportunity	-	-	-	-	-	144	62.6%	0.63	0	1
Dep. variable oper.										
Mathematics (ref.)	94	44.1%	0.44	0	1	75	32.6%	0.33	0	1
Reading	61	28.6%	0.29	0	1	62	27.0%	0.27	0	1
Science	25	11.7%	0.12	0	1	22	9.6%	0.10	0	1
Pooled	33	15.5%	0.15	0	1	71	30.9%	0.31	0	1
Indep. variable oper.										
Non-reform (ref.)	174	81.7%	0.82	0	1	211	91.7%	0.92	0	1
Reform	39	18.3%	0.18	0	1	19	8.3%	0.08	0	1
Dummy	171	80.30%	0.8	0	1	97	45.50%	0.46	0	1
Continuous	42	19.70%	0.2	0	1	133	62.40%	0.62	0	1
Study characteristics										
Less tracked (ref.)	103	48.4%	0.48	0	1	127	55.2%	0.55	0	1
Comprehensive	110	51.6%	0.52	0	1	103	44.8%	0.45	0	1
Multilevel (ref.)	88	41.3%	0.41	0	1	102	44.3%	0.44	0	1
Diff-in-diff	111	52.1%	0.52	0	1	110	47.8%	0.48	0	1
One level regression	14	6.6%	0.07	0	1	18	7.8%	0.08	0	1
Controls: neither (ref.)	24	11.3%	0.11	0	1	8	3.5%	0.03	0	1
Resources	100	46.9%	0.47	0	1	124	53.9%	0.54	0	1
School/class composition	69	32.4%	0.32	0	1	30	13.0%	0.13	0	1
Both	20	9.4%	0.09	0	1	68	29.6%	0.30	0	1
Publication quality										
Publication year	213	100%	12.07	0	21	230	100%	10.45	0	21
Study overall quality	213	100%	4.67	1	6	192	100%	4.74	1	6
Not peer-reviewed (ref.)	100	46.9%	0.46	0	1	66	28.7%	0.29	0	1
Peer-reviewed	113	53.1%	0.39	0	1	164	71.3%	0.71	0	1
Quality of research design	213	100%	0.93	0	2	230	100%	1.17	0	2

TABLES

Table 3 *Tracking mean effect size on student achievement level (educational efficiency)*

Dep. Variable: Hedge's G	Random-effect models								
	1. Full sample			2. No outliers			3. Nested		
	β	*SE*	*p-val*	*β*	*SE*	*p-val*	*β*	*SE*	*p-val*
Mean effect-size	-.024	.025	.320	-.002	.019	.903	-.063	.042	.138
Heterogeneity test	1,492.86 ***			1,412.00 ***			1,492.18 ***		
Tau^2	.116	.014		.065	.009		.087	.024	
I^2	99.76%			99.59%			99.69%		
N. of estimates (studies)	213			209			213 (55)		
Degrees of freedom (df)	212			208			212		

*† p < 0.10, *p < .05, **p < .01, ***p < 0.001*

Note. Model 1: estimation by "Hedges and Vevea (HE)" method according to recommendations from Marín-Martinez and Sánchez-Meca (2010) for standardized mean difference effect sizes. Model 2: estimation by "Restricted Maximum Likelihood (REML)". Estimates are nested within studies.

TABLES

Table 4

Tracking mean effect sizes on achievement inequality

Dep. Variable: Hedge's G	Random-effect models											
	1. Full sample				2. No outliers				3. No outliers / Nested			
	β		*SE*	*p-val*	*β*		*SE*	*p-val*	*β*		*SE*	*p-val*
Mean effect-size	.097	***	.022	<.0001	.076	***	.018	<.0001	.117	***	.033	<.001
Heterogeneity test	1782.12	***			1647.51	***			1782.12	***		
Tau^2	.100		.011		.063		.008		.045			
I^2	99.95%				99.93%				99.89%			
N. of estimates (studies)	230				227				230 (46)			
Degrees of freedom (df)	229				226				229			

*† p < 0.10, *p < .05, **p < .01, ***p < 0.001*

Note. Models 1 and 2: estimation by "Hedges and Vevea (HE)" method according to recommendations from Marín-Martinez and Sánchez-Meca (2010) for standardized mean difference effect sizes. Model 3: estimation by "Restricted Maximum Likelihood (REML)". Estimates are nested within studies

Table 5

Meta-regressions on the effects of tracking on educational efficiency with robust standard errors (nested within studies)

Dep. Variable: Hedge's G	*Model 1*			*Model 2*			*Model 3*			*Model 4*		
	β	*SE*	*p-val*	*β*	*SE*	*p-val*	*β*	*SE*	*p-val*	*β*	*SE*	*p-val*
Intercept	-.086 †	.048	.076	-.011	.065	.867	-.196	.124	.115	.214 †	.125	.087
Policy characteristics												
Between VS within school	.023 ***	.003	<.0001	.022 ***	.003	<.0001	.022 ***	.003	<.0001	.022 ***	.003	<.0001
Outcome subject												
Reading VS mathematics	.016	.014	.254	.015	.014	.270	.014	.014	.301	.014	.014	.319
Science VS mathematics	.002	.003	.384	.002	.003	.388	.002	.003	.384	.002	.003	.383
Pooled VS mathematics	-.047	.079	.553	-.046	.079	.565	-.046	.079	.562	-.047	.079	.556
Institutional variation												
Reform VS other	.097	.102	.341	.126	.104	.227	.104	.105	.322	.125	.101	.215
Research design												
Comprehensive VS less tracked				-.071 ***	.010	<.0001	-.071 ***	.010	<.0001	-.072 ***	.010	<.0001
Method: Diff-in-diff VS multilevel				-.041 ***	.011	.000	-.154 *	.066	.019	-.154 **	.052	.003
One level regression VS multilevel				-.082	.167	.626	-.054	.168	.748	-.137	.163	.402
Controls: Resources VS neither				-.222 *	.104	.032	-.247 *	.105	.018	-.314 **	.105	.003
Composition VS neither				-.063	.119	.597	-.107	.122	.382	-.055	.127	.666
Both VS neither				-.199 †	.103	.054	-.225 *	.104	.031	-.293 **	.105	.005
Publication quality												
Study overall quality							.057 †	.033	.081			
Publication year										-.020 *	.009	.021
Peer-reviewed										-.040	.097	.678
Quality of research design										.115 *	.052	.026
N. estimates (n. studies)	213 (55)			213 (55)			213 (55)			213 (55)		

*† p < 0.10, *p < .05, **p < .01, ***p < 0.001.*

Note: all models control for whether tracking was measured as a dummy or a continuous variable.

TABLES 60

Table 6

Meta-regressions on the effects of tracking on educational inequality with robust standard errors (nested within studies)

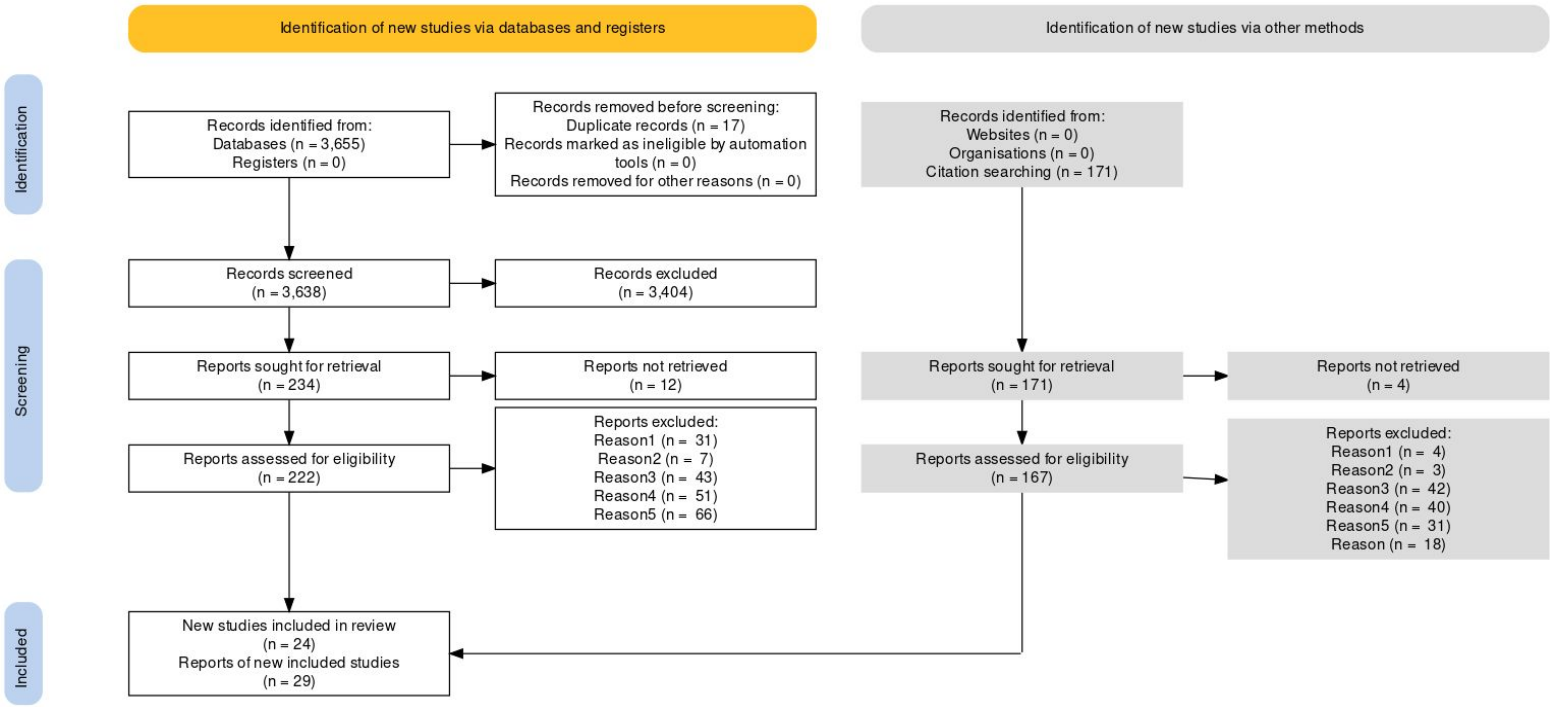
Dep. Variable: Hedge's G	*Model 1*				*Model 2*				*Model 3*				*Model 4*			
	β		*SE*	*p-val*	β		*SE*	*p-val*	β		*SE*	*p-val*	β		*SE*	*p-val*
Intercept	.138	**	.045	.002	.148	*	.060	.014	-.219		.142	.123	-.023		.102	.822
Policy characteristics																
Between VS within school	-.040	***	.007	<.0001	-.038	***	.007	<.0001	-.038	***	.007	<.0001	-.038	***	.007	<.0001
Outcome																
Opportunity VS achv	.020		.047	.664	.046		.050	.365	.026		.050	.609	.007		.052	.891
Reading VS mathematics	-.001		.002	.683	-.001		.002	.681	-.001		.002	.691	-.001		.002	.686
Science VS mathematics	-.001		.003	.678	-.001		.003	.674	-.001		.003	.679	-.001		.003	.680
Pooled VS mathematics	.026		.019	.157	.026		.019	.159	.021		.019	.261	.022		.019	.250
Institutional variation																
Reform VS other	-.050	***	.013	<.0001	-.044	***	.013	.001	-.045	***	.013	.000	-.045	***	.013	.001
Research design																
Comprehensive VS less tracked					-.004	*	.002	.045	-.004	*	.002	.047	-.004	*	.002	.048
Method: Diff-in-diff VS multilevel					-.016		.034	.646	-.025		.034	.461	-.014		.034	.688
One level regression VS multilevel					.244	*	.107	.023	.382	**	.116	.001	.309	**	.111	.005
Controls: Resources VS neither					-.055		.092	.550	-.120		.091	.184	-.081		.102	.425
Composition VS neither					-.098		.131	.454	-.046		.126	.713	-.165		.129	.202
Both VS neither					-.070		.106	.506	-.047		.101	.642	-.069		.105	.513
Publication quality																
Study overall quality									.084	**	.030	.005				
Publication year													.008		.008	.306
Peer-reviewed													.030		.071	.673
Quality of research design													.094	†	.053	.078
N. estimates (n. studies)	230 (46)				230 (46)				230 (46)				230 (46)			

*† p < 0.10, *p < .05, **p < .01, ***p < 0.001*

Note: all models control for whether tracking was measured as a dummy or a continuous variable.

Figure 1

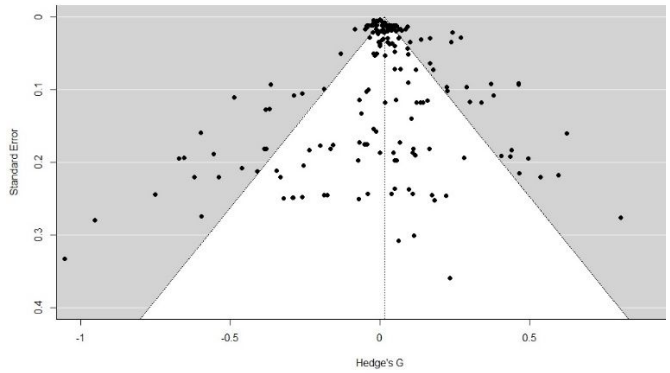
PRISMA diagram



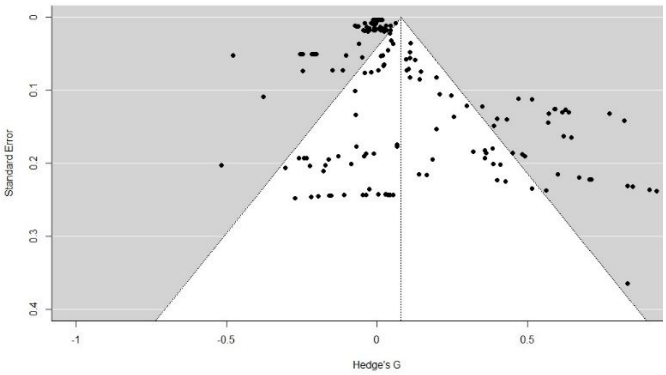
Reason 1: *educational level considered is not secondary education (e.g. primary education or higher education).*
Reason 2: *out of publication year range (2000-2021).*
Reason 3: *treatment is not tracking as a macro/institutional characteristic (in most cases it refers to studies focused on comparing students' outcomes across tracks).*
Reason 4: *design (it addresses a different research question and employs a different analytical design).*
Reason 5: *outcome is not achievement/student performance (it might be non-cognitive skills or educational transitions).*
Reason 6: *duplicated.*

Figure 2

Funnel plot for educational efficiency (left) and inequality (right)



Note. Regression Test for Funnel Plot Asymmetry: z = -4.9638, p < .0001
Limit Estimate (as sei -> 0): b = 0.0854 (CI: 0.0254, 0.1454)



Test for Funnel Plot Asymmetry: z = 6.9189, p < .0001
Limit Estimate (as sei -> 0): b = -0.0330 (CI: -0.0852, 0.0192)

APPENDICES

Appendix A

Eligibility criteria

***Independent variable*:** tracking (dummy, educational reforms, age of tracking, n. of tracks, length of tracking, share of vocational track, interaction tracking x family background, etc.)
Dependent variable (efficiency): achievement: standardized students proficiency (PISA, TIMSS, PIRLS), IQ tests, n. of correct questions, non-standardized tests, etc.
Dependent variable (inequality of achievement): inequality in achievement: gap between top/bottom performers, standard deviation, interquartile range, the share of top/bottom performers.
Dependent variable (inequality of opportunity): effect of family background on achievement.
Measurement: β coefficients.
Research respondents (population): Secondary students
Research methods: Experimental and quasi-experimental designs.
Language: English
Time frame: 1980 – 2020 (obs.: a meta-analysis on the topic was published in 1982).
Publication type: published journals, books, dissertations, technical reports, conference presentations, policy reports, etc.
Availability: online.

Search terminology

Terms for educational field:

"educ*" OR "student*"

AND

Terms for tracking:

track OR "ability group" OR stream OR sort

AND

Methodological terms:

experiment OR quasi-experiment OR difference-in-difference OR multilevel OR regression OR "propensity score" OR "instrumental variable"

Appendix B

Formulae: From Borenstein et at. (2009), chapter 4 (p. 21).

For regression coefficients reported in a standardized form:

Effect size (Cohen's *D*):

$$D = \overline{X}_1 - \overline{X}_2 \qquad Eq. (B. 1)$$

Where D is the mean difference between two independent groups, $\overline{X}_1$is the standardized mean of group 1 and $\overline{X}_2$ is the standardized mean of group 2, i.e., $\overline{X}_1 - \overline{X}_2$ is equal to the standardized regression coefficient.

Standard deviation of the difference:

$$S_{diff} = SE \times \sqrt{n-1} \qquad Eq. (B.2)$$

Where *SE* is the standard error of the regression coefficient and *n* is the sample size at the considered level of the analysis, i.e., n. of macro-units for macro-level variables and n. of micro-level units for interactions between micro and macro-level variables.

Variance of effect size:

$$V_D = \frac{S_{diff}^2}{n} \qquad Eq. (B.3)$$

Where S_{diff}^2 is the standard deviation of the difference squared and *n* is the sample size.

Standard error of the effect size:

$$SE_D = \sqrt{V_D} \qquad Eq. (B.4)$$

Where $\sqrt{V_D}$ is the squared root of the variance of the effect size.

For regression coefficients reported in an unstandardized form:

Effect size (Cohen's *D*):

$$D = \frac{\overline{Y}_{diff}}{S_{within}} = \frac{\overline{Y}_1 - \overline{Y}_2}{S_{within}} \qquad Eq. (B.5)$$

Where $\overline{Y}_1 - \overline{Y}_2$ is the unstandardized regression coefficient and S_{within} is the pooled standard deviation.

Standard deviation of the difference:

$$S_{diff} = SE \times \sqrt{n-1} \qquad Eq. (B.6)$$

Where *SE* is the standard error of the regression coefficient and *n* is the sample size for the respective level of analysis.

Pooled standard deviation:

$$S_{within} = \frac{S_{diff}}{\sqrt{2(1-r)}} \quad \textit{Eq. (B.7)}$$

Where S_{diff} is the standard deviation of the difference and *r* is the correlation of the dependent variable in times 1 and 2. No study reported *r,* so it was set to .5 in order to make the divisor neutral = *1*.

Variance of effect size:

$$V_D = \left(\frac{1}{n} + \frac{D^2}{2n}\right) \times 2(1-r) \quad \textit{Eq. (B.8)}$$

Where *D* is the regression coefficient, *n* is the sample size for the respective level of analysis and *r* is the correlation of the dependent variable between times 1 and 2 and was set to .5.

Standard error of the effect size:

$$SE_D = \sqrt{V_D} \quad \textit{Eq. (B.9)}$$

Bias correction (transformations of Cohen's *D* to Hedge's *G*):

Hedge's bias correction factor:

$$J = 1 - \frac{3}{4df - 1} \quad \textit{Eq. (B.10)}$$

Where *df* is equal to *n-2*.

Effect size (Hedge's *G*):

$$G = J \times D \quad \textit{Eq. (B.11)}$$

Where *J* is Hedge's correction factor and *D* is Cohen's *D*.

Variance of effect size (Hedge's *G*):

$$V_G = J^2 \times V_D \quad \textit{Eq. (B.12)}$$

Where *J* is Hedge's correction factor and $V_D$is the variance of Cohen's *D*.

Standard error of effect size (Hedge's *G*):

$$SE_G = \sqrt{V_G} \quad \textit{Eq. (B.13)}$$

Where V_G is the variance of Hedge's *G*