



BRIDGE Project Deliverable 2.1

Causal effects of educational between-school selection mechanisms: A multilevel meta-analysis

BRIDGE is an impact-driven project that aims to build resilient individuals through effective educational transition by formulating robust policy recommendations on education and training policies, the efficient use of public resources and the support of equity and inclusion in education and training systems in the EU.

This work was supported by the European Commission Horizon Europe project BRIDGE (grant agreement No. 101177154). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Keywords: school choice; early tracking; elite tracking; voucher lottery
DOI: 10.5281/zenodo.18696882

Authors: Kaire Põder, Faisal Mohammed, Triin Lauri

Affiliation: Estonian Business School

Date: February 25, 2026

Deliverable: D2.1

Abstract

We synthesise evidence on three selection mechanisms: school choice by random allocation of students, formal (early) and elite tracking – focusing on both average effects and heterogeneity for low- and high-performing students’ outcomes. By concentrating only on causal studies which apply robust experimental or quasi-experimental evidence and using multilevel random-effects meta-regression with cluster-robust standard errors, we avoid the confounding biases of a limited number of datasets applied or multiple outcomes from the same author. Across early tracking systems, mean effects are often reported as negative, and for the random allocation mechanism and elite tracking, positive. Our results show that grand mean effects are not entirely consistent with the state of the art. In the case of elite tracking, we can report only local treatment effects on treated across the cut-off, which is a small positive. In the case of early tracking, the effect is small and negative, but the distributional patterns of the effects show that low and high-achieving students benefit from this selective mechanism. In case of random assignment, the grand mean effect is zero, but top-performers are hurt by lotteries. So, the applications of selective school admission mechanisms always contain political choices – who to grant preferable treatment.

1 Introduction

The Programme for International Student Assessment (PISA) 2022 results reveal a significant decline in student achievement, with the average performance across OECD countries falling by 15 points in mathematics and 10 points in reading—roughly equivalent to half to three-quarters of a school year of learning (OECD, 2023, p. 44). While part of this decline can be linked to the disruptions caused by the COVID-19 pandemic, evidence indicates that this does not account for the full extent of the downturn. Notably, data show that high-achieving students have experienced steeper declines than their lower-achieving peers (ibid. 2023, p. 189), suggesting that the erosion of performance affects the entire achievement distribution rather than only traditionally disadvantaged groups.

Against this backdrop, there is a growing policy focus on high achievers. This brings, among new innovative (digital) solutions, old debates about selectivity in education into focus. By selection mechanisms, we mean system-level policies such as governance of parental choice, early (formal) or elite tracking. All these selective mechanisms are often justified as ways to match instruction with student aptitude and promote the efficiency of teaching. So, we ask what the cumulative causal evidence is for selective practices on the achievement of students in mathematics, reading, and writing, and whether these effects systematically differ by students’ achievement level.

Selective mechanisms structure who learns with whom and where. We study the effects of three selective mechanisms at the basic school level. First is early tracking – a formalised policy by which students are assigned to distinct educational pathways – such as academic, vocational, or intermediate tracks (higher and lower tracks) – typically around the ages of 10 to 14. Second, elite tracking, by contrast, operates largely

through parental choice: students and families self-select into more prestigious or academically demanding (elite public or private) schools, based on competitive entrance criteria. Third, assigning students by lottery, by contrast, aims at equity and fair access under conditions of oversubscription of schools.

So, in the case of three selective mechanisms, the assignment rules differ – early tracking is system-wide and affects all students and typically relies on teachers’ recommendations; elite tracking is pushed by (a group of more ambitious) parents by aptitude tests; choice by lottery is also triggered by parents, but allocation (of the voucher) is random. Thus, while all three mechanisms allocate students across schools, they differ fundamentally along two dimensions: who makes the assignment decision (teachers, parents, or randomization) and whether selection applies system-wide or only at the margin.

Selective mechanisms inevitably affect peer composition, teacher assignment, and school environments. However, the object of analysis in this study is not peer effects per se, but the causal effect of selectivity as an assignment rule. That is, we estimate how different selection mechanisms – early tracking, elite tracking, and random allocation – affect student outcomes through the combined bundle of changes they induce, rather than isolating specific channels such as peer quality or teacher effects. Distinguishing selectivity effects from individual mechanisms is beyond the scope of this meta-analysis and would require different research designs.

Our contribution is related to estimating the heterogeneous effects of the selective mechanism. As well-established literature shows (Woessmann, 2007; Brunello & Checchi, 2007; Hanushek & Woessmann, 2006; Ammermüller, 2005), early tracking is generally detrimental to students on average. We show the same, but complement the finding with subject- and achievement-specific insights and report positive effects on high- and low achievers. In elite tracking (Clarke 2010; Pop-Eleches & Yrquiola, 2013; Estrada & Gignoux, 2017), the literature shows positive effects on treated and negative effects on non-treated, meaning those who were left behind (Schiltz et al. 2019). We report the same regarding the local effects on treated borderline students who crossed the threshold and can be defined as low achievers for elite schools. In school choice, the lottery is expected to equalise the chances but benefit the winners (Shakeel et al., 2021; Abdulkadiroglu et al., 2011; Angrist et al., 2013; Dobbie Fryer, 2013; Chabrier et al., 2016). We show that this is not so on average, meaning that the mean effect of winning is zero, and high achievers perform worse even if winning the lottery.

Our empirical strategy is to combine a database of causal studies published within the last 20 years and apply a multilevel meta-analytic method designed for dependent effect sizes. We start from more than 72,000 sources, but after filtering out non-causal, different outcome and age group studies, we end with 23 studies: 6 from early tracking, 12 from school choice and only 5 from elite tracking. Since many studies in our dataset report multiple estimates across academic domains, cohorts, or achievement groups, we employ three-level random-effects meta-regression models that explicitly partition sampling variance, within-study heterogeneity, and between-study heterogeneity. To ensure valid inference under potentially mis-specified dependence structures and a limited

number of studies, we use cluster-robust variance estimation with small-sample corrections. This framework allows us to estimate overall effects as well as mechanism-specific, domain-specific, and achievement-level effects while appropriately accounting for statistical dependence and heterogeneity across studies.

We continue as follows. First, we briefly discuss the insights from the theoretical literature and existing meta-studies. Then we concentrate on data, which in our case are statistics and estimates from quasi-experimental empirical studies and give our estimation strategy based on particularities originating from the data. Finally, we give the results, their robustness, and discuss the findings.

2 Underpinnings from the literature

Systems with early tracking (e.g., Germany, Austria, the Netherlands) separate students at a younger age, while comprehensive systems delay such differentiation until later stages of schooling. Mostly, these decisions are made by combining teacher recommendations with grades or test results. Though mechanisms of early and elite tracking differ in scope and structure, they share a common goal of aligning instruction with ability. The state of the art in the literature has been that early tracking is harmful for disadvantaged children (Terrin & Triventi, 2023; European Commission, 2022; Matthewes, 2021), leading many continental European countries to adopt de-tracking measures (i.e. delaying or reducing the separation of students). However, a similar approach is not evident in several comprehensive systems¹ regarding elite tracking, which relies on entrance examinations or prior achievement thresholds to track high-achieving students in separate, often high-status educational institutions or programs (Horn, 2013). The US school districts have experimented with centralised mechanisms and randomised allocation in the case of oversubscription to schools for years, advocating for more fair educational chances for disadvantaged students (Abdulkadiroğlu et al., 2005a; Abdulkadiroğlu et al., 2005b). So, to whom do selective mechanisms benefit?

2.1 System-level (formal) and school-level (elite) tracking

The argument for addressing students’ differing learning needs is based on the fact that children vary in their cognitive abilities, learning pace and mastery of content. Classic developmental theories support this idea. For example, Piaget’s theory of cognitive readiness (1972) suggests that effective learning only happens when instruction matches the child’s developmental stage. Likewise, Vygotsky’s concept of the zone of proximal development (1978) – the gap between what a learner can do alone and what they can achieve with guidance – shows why teaching should be tailored to students’ current abilities and potential for growth. Later, research in neuroscience also confirms that students differ in the way they process information, which implies that many benefit from specialised or adapted instruction (Neubauer & Fink, 2009; Casey et al., 2005).

¹Comprehensive systems often just did de-tracking reforms historically earlier, e.g. Finland implemented a detracking reform from 1972-1977 (Pekkarinen et al., 2006); UK in 2000s.

According to the aptitude–treatment interaction theory, teaching is most effective when instructional methods align with a student’s ability level. Grouping students by ability can therefore help teachers to appropriately adjust the pace, depth and style of learning. This logic has a long history in both theory (Cronbach & Snow, 1977; 1986) and practice, and is supported by empirical work showing that differentiated learning paths – such as grade-skipping, subject acceleration or enriched curricula – can yield strong academic benefits without harming students’ social or emotional development (Tomlinson, 2001; Kulik & Kulik, 1992). From an economic perspective, differentiation is also linked to the broader value of human capital formation: policies aimed at fostering high achievement contribute to national productivity and innovation by nurturing scientists, creatives and future leaders.

To further develop this “human capital argument”, it can be argued that it is grounded in an ideological belief in meritocracy: that individuals should advance on the basis of ability and effort rather than social background or privilege. From this perspective, selective educational arrangements can be viewed as a neutral mechanism for recognising talent and rewarding achievement, especially when access to selective tracks is based on standardised criteria. However, there is growing concern that many countries fail to deliver truly meritocratic education systems (Barone, 2019). Despite the rhetoric and formal equality of opportunity, structural inequalities such as social class, migration background and parental education continue to shape educational trajectories from an early age. This challenges the legitimacy of selective and tracking-based policies that are not accompanied by strong equity measures.

Within this framework, we distinguish between two forms of selective practices in education: (early) formal tracking and elite tracking. While both involve differentiation by academic ability, their allocation mechanism often differs – in the case of elite tracking, only grades (tests) matter. Unlike early tracking, which can indiscriminately sort large populations of students and often leads to entrenched disadvantage for low performers, elite tracking targets top performers (and their ambitious parents) and requires that schools have autonomy over student admissions. The negative consequences of early tracking for low-achieving students are well documented (European Commission, 2022; Bol et al., 2014; Barone et al., 2018; Van de Werfhorst & Hofstede, 2007), and similar concerns extend to elite tracking. Elite tracking exacerbates disparities by concentrating resources and opportunities on high achievers (Bygren & Rosenqvist, 2020). At the same time, a growing body of research suggests that peer composition can play a central role in shaping outcomes in elite settings. Studies exploiting quasi-experimental variation show that exposure to high-achieving peers can generate positive academic spillovers for top-performing students, particularly in high-stakes, exam-based systems (Ding & Lehrer, 2007; Lavy, Silva & Weinhardt, 2012; Mendolia, Paloyo & Walker, 2018). However, previous meta-review evidence indicates that such peer effects are typically modest on average and highly heterogeneous, with benefits concentrated among students near the cut-off point (Terrin & Triventi, 2023).

2.2 School choice: random assignment

School choice is frequently promoted as a strategy to enhance quality through competition or more dynamic and responsive governance practices; its positive impact on student outcomes appears contingent on local institutional and contextual conditions (Wilson & Bridge, 2019; Triventi et al., 2020). Historical trajectory of promoting school choice through market mechanism "a voucher system that would enable parents to choose freely the schools that their children attend is the most feasible way to improve elementary and secondary education in the US" (Friedman 1997) has given ground to "choice-autonomy-accountability platform". In more recent formulations, this market logic has evolved into what is often described as a choice-autonomy-accountability framework, where school choice is expected to improve outcomes only when combined with meaningful school autonomy, credible accountability mechanisms, and well-informed parental decision-making (Hanushek et al., 2013; Wöessmann et al. 2009). However, as research shows (DeAngelis & Erickson, 2018; Greaves et al., 2023) the choice-autonomy-accountability framework often fails in implementation.

In this study, we define school choice narrowly as an allocation mechanism that allows families to express preferences over schools, with students assigned to oversubscribed schools by random lottery rather than by aptitude or prior achievement. However, some threshold can be a prerequisite to be able to join the program. Handling self-selecting features is crucial for causal identification, as most studies in this literature must first address the self-selection problem inherent in voucher programmes, whereby participation is disproportionately driven by families with higher socio-economic status. Lottery-based designs exploit random variation in access among applicants, thereby isolating the effect of gaining access to a preferred school, rather than the effect of choosing per se.

Winning a lottery typically grants students access to schools that differ from their fallback options, often including private schools with stronger peer environments, greater autonomy, or distinctive pedagogical practices. In this sense, lottery winners are expected to benefit from features commonly associated with elite or selective schooling, albeit without explicit ability-based sorting. However, a key limitation of randomised school choice as a policy instrument for supporting excellence lies precisely in this neutrality with respect to ability: because lotteries do not prioritise high-achieving students, they may fail to match advanced learners to more demanding educational environments and may even dilute peer effects at the top of the achievement distribution (Hastings et al., 2006; Cullen et al., 2006).

In systems where random assignment mitigates self-selection, high achievers might not benefit from differentiated environments that do not cater specifically to their needs, as shown also by many authors in Appendix 1. Without institutional (elite) tracking, enriched curricula or specialised instructional methods that are targeted at advanced learners, the potential academic benefits of school choice may be minimal.

Previous meta-studies summarising mostly US evidence (e.g. Fryer, 2017; Shakeel et al., 2021) report null or positive grand mean effects. Shakeel et al. (2021) also report discipline-specific effects, indicating larger positives for reading than math.

3 Methods and data

3.1 Selection of studies

We conducted a systematic literature search using the EBSCO database filtered through Scopus as the content provider. This choice was justified by Scopus’s comprehensive coverage and employment of multiple quality indicators, including Source Normalised Impact per Paper (SNIP) and SCImago Journal Rank (SJR), which enable better identification of high-quality research through citation-weighted measures of scholarly influence compared to databases relying solely on journal impact factors.

Our search strategy focused on identifying causal studies of three distinct educational selection mechanisms: formal tracking (particularly early tracking implemented before age 12), elite or selective (e.g. private) school admission based on academic merit, and school choice implemented through random allocation (lottery systems). The search encompassed studies published between January 2005 and September 2024, reflecting the period during which rigorous quasi-experimental methods became standard in educational research.

Studies were eligible for inclusion if they employed causal inference methods, including randomised controlled trials or quasi-experimental designs such as difference-in-differences, regression discontinuity design, or instrumental variable estimation. We required studies to focus on students aged 15 years or younger at treatment exposure, measure student achievement in at least one PISA discipline (mathematics, reading, or science), and provide sufficient statistical information to calculate standardised effect sizes. Following evidence standards established by the What Works Clearinghouse (2022), we required at least three independent causal studies per policy mechanism to ensure a minimum level of replication, as single or paired studies may be idiosyncratic, context-bound, or underpowered.

The study selection process followed PRISMA 2020 guidelines (Page et al., 2021) across four stages: identification, screening, eligibility assessment, and inclusion (Figure 1). Three independent reviewers (authors) screened titles and abstracts using predefined criteria focusing on causal inference methodology, student achievement outcomes, and relevance to one of the three selection mechanisms. Disagreements were resolved through discussion among the reviewers. Full-text articles of potentially eligible studies were retrieved and assessed against complete inclusion criteria. The final sample comprised studies on early tracking, elite/selective admission, and school choice via lottery.

The geographical distribution of included studies reflects historical patterns in policy implementation and data availability. Tracking studies predominantly originated from continental European countries, where early tracking systems have been institutionalised. Elite tracking studies came from comprehensive systems (e.g. UK and post-Soviet), while school choice studies were concentrated in the United States, where lottery-based admission provides natural experiments. Methodologically, tracking studies most commonly employed difference-in-differences designs exploiting policy reforms, elite tracking studies typically used regression discontinuity designs leveraging admission cut-offs, and school choice studies predominantly used instrumental variable estimation,

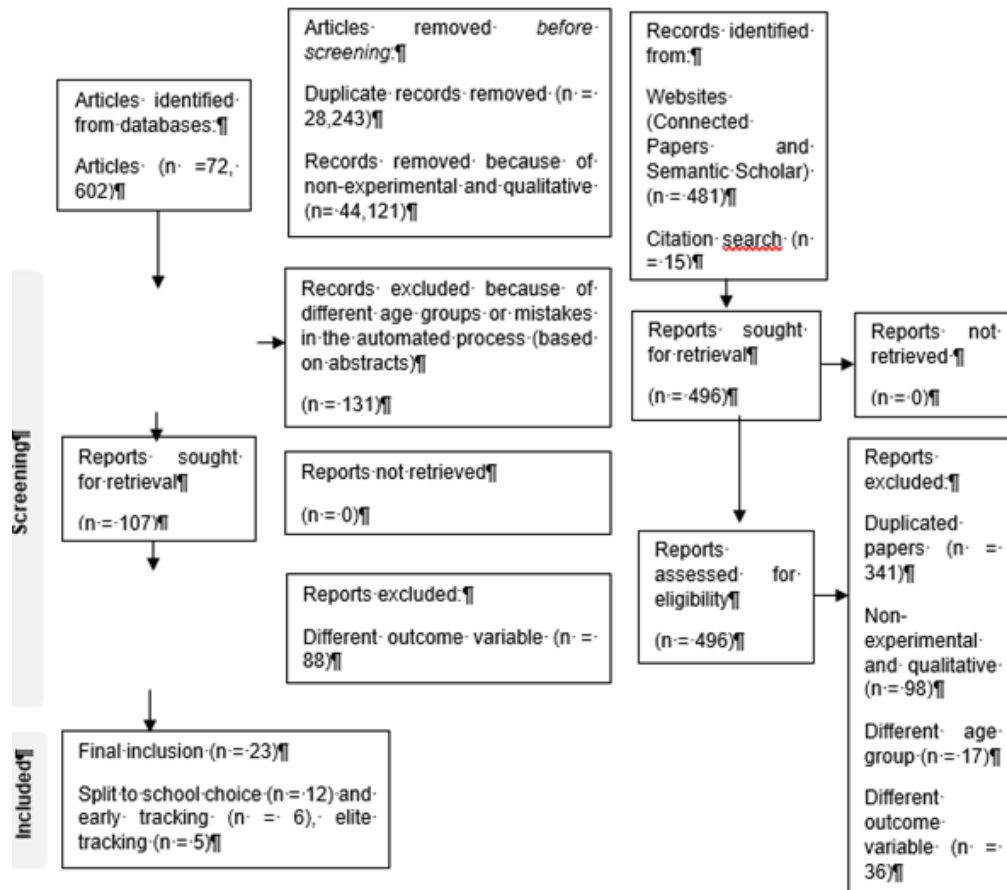


Figure 1: PRISMA flow diagram

Notes: This figure presents the standard structure of the PRISMA 2020 flow diagram, which documents the process of study identification, screening, eligibility and inclusion in systematic reviews. It distinguishes between records identified via databases/registers and those identified through other sources. It also tracks the reasons for exclusions and clarifies the number of studies retained at each stage. Source: Page et al., (2021).

treating lottery offers as instruments for actual attendance.

3.2 Data extraction and effect size calculation

Data records point estimates of treatment effects, standard errors, sample sizes, outcome measures, study design characteristics, treatment characteristics, and sample characteristics (see also Appendix 1 for a qualitative description of data). When studies reported differential effects by baseline achievement level, we extracted separate estimates for high-achieving students (defined as \geq PISA proficiency level 5 or \geq 75th percentile) and low-achieving students (defined as $<$ PISA proficiency level 2 or \leq 25th percentile).

To create a common metric across studies using different achievement tests and reporting formats, we converted all treatment effects into Hedges' g , a standardised mean difference correcting for small-sample bias (Hedges, 1981). Effect sizes were calculated manually for each study using analytic formulas following canonical meta-analytic methods (Lipsey & Wilson, 2001)². When studies reported effects in standard deviation units (e.g., standardised regression coefficients or Cohen's d), these were treated as unbiased estimates of Cohen's d and converted to Hedges' g using the appropriate small-sample correction based on the total analytic sample size. Corresponding sampling variances were computed analytically rather than inferred from reported confidence intervals. For studies reporting unstandardized coefficients in original test-score units, we reconstructed standardised mean differences using pooled outcome standard deviations computed directly from reported group-level statistics when available, or derived analytically using canonical variance formulas. For studies using PISA outcomes, we used population standard deviations from PISA technical documentation to ensure consistency across countries and waves.

The hierarchical structure of our data—with multiple effect sizes nested within studies—necessitated a three-level random-effects meta-regression model. This specification partitions variance across three levels: Level 1 captures sampling variance of individual effect sizes due to sampling error; Level 2 accounts for within-study heterogeneity arising from multiple effect sizes per study (different domains or cohorts); and Level 3 models between-study heterogeneity due to substantive differences across studies. The full model can be expressed as:

$$g_{ij} = \beta_0 + u_i + u_{ij} + e_{ij} \quad (1)$$

Where g_{ij} is the j th effect from i th study, β_0 is the overall mean effect, u_i represents between-study random effects, u_{ij} represent within-study random effects and e_{ij} represents the known sampling variance. The variance components partition the heterogeneity at within-study (ω^2) and between-study (τ^2) levels. Models were estimated using restricted maximum likelihood (REML) via the metafor package (Viechtbauer, 2009) in R. The three-level specification with REML as recommended by Pustejovsky and Tipton (2022) for data structure like ours accommodates both within-study correlation in

²Effect size and variance calculations follow canonical meta-analytic formulas and were implemented manually, with transformations cross-validated against Wilson's (2023) Effect Size Calculator (version 2023.11.27; Campbell Collaboration). Available at: [Effect Size Calculator – Campbell Collaboration](#).

sampling errors and hierarchical nesting of effects within studies. The use of fully inverse-variance weighting in our framework has been shown to yield more precise estimates than semi-efficient diagonal weights, particularly when moderators vary at the effect-size level (Pustejovsky & Tipton, 2022)³. To assess the robustness of our findings to alternative estimation approaches, we estimated all models using standard robust variance estimation (RVE) methods with method-of-moments variance estimation as implemented in the *robumeta* package (Fisher & Tipton, 2017).

To guard against misspecification of the assumed correlation structure, we employed cluster-robust variance estimation using the CR2 estimator with Satterthwaite small-sample corrections (Pustejovsky & Tipton, 2022) in three-level random effect models. This approach treats the study as the primary clustering unit and provides valid inference even when the assumed dependence structure is incorrect, which is particularly important given the modest number of studies typically available in meta-analysis (Hedges et al., 2010)⁴.

We quantified heterogeneity using the I^2 statistic, which represents the proportion of total variation attributable to true heterogeneity rather than sampling error, with values of 25%, 50%, and 75% interpreted as indicating low, moderate, and high heterogeneity, respectively (Higgins et al., 2003). To examine systematic variation in effects, we conducted separate meta-analyses stratified by selection mechanism (early tracking, elite tracking, school choice), academic domain (mathematics, reading, science, results shown in Appendix 2), and baseline achievement level (high-achieving, low-achieving students). This stratified approach allowed us to estimate selectivity mechanism-specific, domain-specific, and achievement-level-specific effects while maintaining the three-level hierarchical structure with cluster-robust variance estimation within each stratum.

We also test robustness of our findings sensitivity to publication selection bias using multiple approaches, such as visual inspection of funnel plots, Egger’s regression test (Egger et al., 1997), and trim-and-fill analysis (Duval & Tweedie, 2000) to estimate the number and impact of potentially missing studies. Because these diagnostic tools cannot accommodate multilevel data structures, publication bias assessments used univariate random-effects models fit with the *rma()* function in *metafor*, rather than the three-level multivariate models (*rma.mv*) used for main analyses. This is standard practice in multilevel meta-analysis (Assink & Wibbelink, 2016). We considered publication bias likely when multiple diagnostic approaches converged on similar conclusions

³Pustejovsky and Tipton (2022) show that, for data structures like ours, a three-level specification with REML estimation and fully inverse-variance weighting can outperform standard RVE methods. First, they show through simulation that this approach can yield 10-50% more precise estimates than RVE’s semi-efficient diagonal weights when moderators vary at the effect-size level, as is the case in our domain and achievement-level analyses. The precision gains stem from fully inverse-variance weighting that properly accounts for the complete covariance structure within studies. Second, REML estimation of the three-level model provides estimates of both between-study (τ^2) and within-study (ω^2) variance components, offering richer descriptive information about heterogeneity sources than RVE’s single variance parameter—particularly valuable given the wide variation in outcome measures, follow-up times, and treatment conditions in our synthesis.

⁴The RVE approach uses different assumptions about the within-study correlation structure and employs residual degrees of freedom rather than Satterthwaite corrections.

and reported both unadjusted and bias-adjusted estimates when adjustment materially affected conclusions.

4 Results

4.1 Formal early tracking

Figure 2 shows the results of early tracking on students' performance. While many primary studies found statistically insignificant effects of early tracking on students' performance, the pooled effect revealed a small negative overall effect on students' achievement ($g = -0.10$, 95% CI $[-0.12, -0.08]$, $p < 0.0001$, $k = 21$ effect sizes from 6 studies). This finding indicates that students in early-tracking systems (tracking before age 12) achieve approximately 0.10 standard deviation lower in test scores than comparable students in delayed-tracking or comprehensive systems. Heterogeneity was low across studies ($Q(20) = 15.66$, $p = 0.74$, $I^2 = 0\%$), with negligible between-study ($\tau^2 < 0.001$) and within-study ($\omega^2 = 0.0006$) variance components, indicating no variation in tracking effects beyond sampling error.

The negative average effect of early tracking on students achievement is consistent across academic domains (see Appendix 2, Figure 2.1), with mathematics showing $g = -0.07$ (95% CI $[-0.10, -0.04]$, $p = 0.004$), reading $g = -0.15$ (95% CI $[-0.23, -0.17]$, $p = 0.0001$), and science $g = -0.08$ (95% CI $[-0.13, -0.03]$, $p = 0.015$). These results show that early tracking has the strongest negative impact on reading (0.15 standard deviation decline), approximately twice the magnitude observed for mathematics or science. This pattern aligns with evidence suggesting that language skills may be particularly sensitive to early peer effects and classroom interaction, as reading development relies more heavily on verbal interaction during critical language acquisition periods (Snow et al., 1998), though our data cannot directly test this mechanism.

Figure 3 presents effect sizes (Hedges' g) estimating the impact of early tracking on student achievement separately for high-achieving (upper panel) and low-achieving students (lower panel). Even though we rely only on 3 studies, the multiple estimates are shown for studies reporting effects across different cohorts or domains, which allows us to make generalisations from existing evidence. The effect of early tracking is statistically insignificant for both high- and low-achieving students. For high-achieving students (≥ 75 th percentile or PISA Level 5+), early tracking showed a small positive but statistically insignificant effect ($g = 0.06$, 95% CI $[-0.17, 0.28]$, $p = 0.40$, $k = 13$ estimates from 3 studies). Low-achieving students (≤ 25 th percentile or PISA Level < 2) showed a negligible positive effect that is statistically insignificant ($g = 0.02$, 95% CI $[-0.04, 0.07]$, $p = 0.31$, $k = 13$ estimates from 3 studies). Notably, these subgroup patterns differ from the overall negative effect ($g = -0.10$), likely because only three of the six tracking studies reported achievement-level heterogeneous effects, and the overall negative effect may be driven primarily by middle-achieving students, not examined separately in subgroup analyses. That is, high achievers may be buffered by their placement in top tracks with advantaged peers. Low achievers may receive targeted support. But mid-achieving

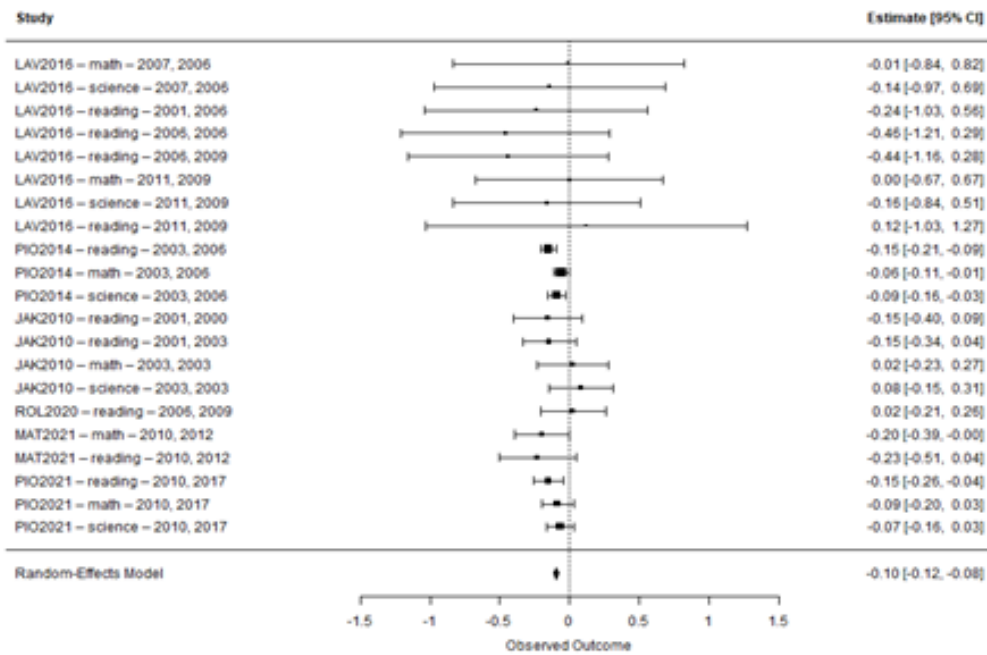


Figure 2: Grand mean effect of early tracking on students' achievement

Notes: Each study is uniquely labelled by study ID (first three letters of the author's name plus year of publication), domain (mathematics, science, or reading) from which the effect size is estimated, and the survey year(s) from which the primary study conducted its analysis. Each square represents an individual effect size (Hedges' g), with horizontal lines indicating 95% confidence intervals. The diamond represents the pooled random-effects estimate. Confidence intervals are cluster-robust to account for multiple estimates per study. Negative values indicate lower achievement in early-tracking systems.

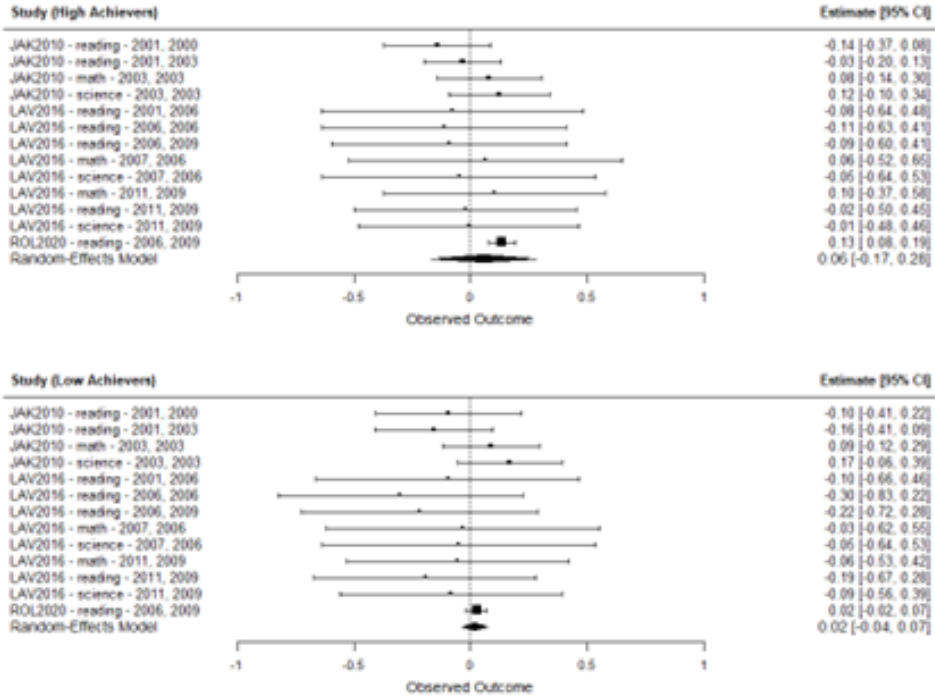


Figure 3: Heterogeneous grand mean effect of early tracking on students' achievement

Notes: Each study is uniquely labelled by study ID (first three letters of the author's name plus year of publication), domain (mathematics, science, or reading) from which the effect size is estimated, and the survey year(s) from which the primary study conducted its analysis. Each square represents an individual effect size (Hedges' g), with horizontal lines indicating 95% confidence intervals. The diamond represents the pooled random-effects estimate. Confidence intervals are cluster-robust to account for multiple estimates per study. Negative values indicate lower achievement in early-tracking systems.

students fall through the cracks. This finding challenges the common narrative that tracking primarily harms the lowest performers.

4.2 Elite tracking

Figure 4 shows the results of elite/selective school tracking on students' performance. The pooled effect revealed a small positive effect but a statistically insignificant effect on students' achievement ($g = 0.05$, 95% CI [-0.03, 0.13], $p = 0.14$, $k = 12$). This finding indicates that students attending elite or selective schools achieve approximately 0.05 standard deviation higher in test scores than comparable students in non-selective schools, though not statistically significant. Heterogeneity was high across studies ($Q(11) = 49.66$, $p < 0.0001$, $I^2 = 77.8\%$), with small between-study ($\tau^2 = 0.0006$) and moderate within-study ($\omega^2 = 0.0047$) variance components, suggesting considerable variation in selective school effects both across and within studies. Domain-specific analyses (see

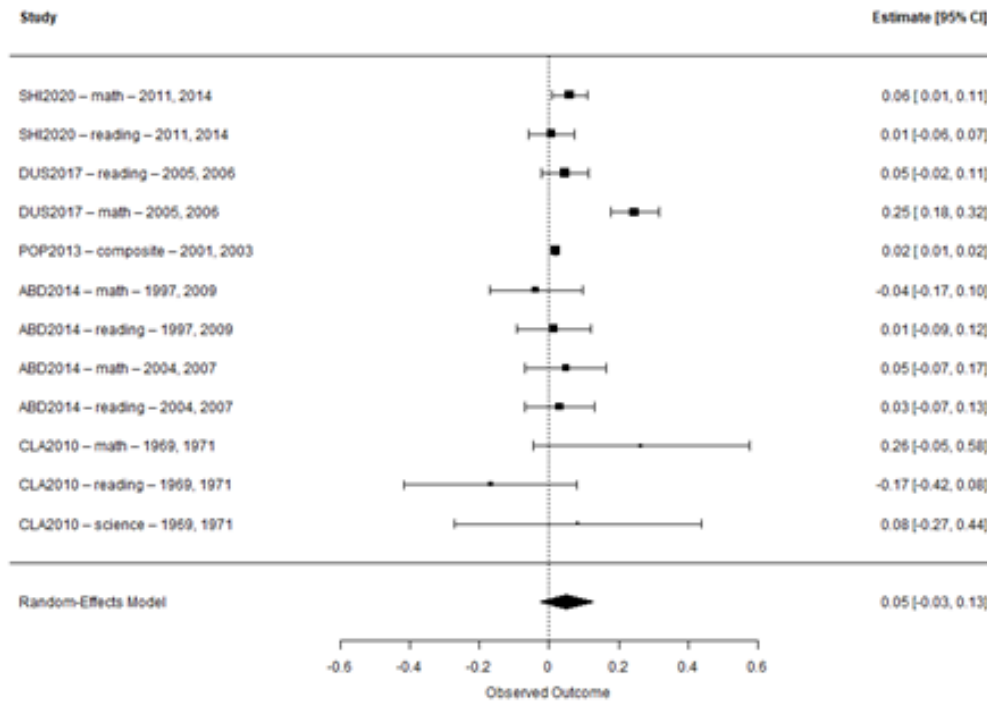


Figure 4: Grand mean effect of elite tracking on students' achievement

Notes: Each study is uniquely labelled by study ID (first three letters of the author's name plus year of publication), domain (mathematics, science, or reading) from which the effect size is estimated, and the survey year(s) from which the primary study conducted its analysis. Each square represents an individual effect size (Hedges' g), with horizontal lines indicating 95% confidence intervals. The diamond represents the pooled random-effects estimate. Confidence intervals are cluster-robust to account for multiple estimates per study. Negative values indicate lower achievement in early-tracking systems.

Appendix 2, Figure 2.3) reveal a similar pattern across subjects. For reading, the effect was negligible ($g = 0.02$, 95% CI $[-0.02, 0.06]$, $p = 0.19$, $k = 5$), with no heterogeneity across studies ($Q(4) = 2.94$, $p = 0.57$). Mathematics showed a larger positive but statistically insignificant effect ($g = 0.12$, 95% CI $[-0.08, 0.32]$, $p = 0.15$, $k = 5$), with substantial between-study heterogeneity ($\tau^2 = 0.013$, $Q(4) = 25.07$, $p < 0.0001$). The larger and more variable effects in mathematics compared to reading may reflect that mathematics instruction may be more sensitive to peer composition and specialised teaching resources that vary considerably across elite school contexts, while reading skills may be more influenced by factors outside the school environment. However, the small sample size (only 4-5 studies per domain) limits our ability to draw firm conclusions, and these findings should be interpreted with caution.

4.3 School choice by winning an admission voucher through a lottery

From Figure 4, winning a school admission voucher shows no statistically significant overall effect on student achievement ($g = 0.09$, 95% CI $[-0.06, 0.24]$, $p = 0.21$, $k = 24$). However, heterogeneity was substantial across studies ($Q(23) = 216.54$, $p < 0.0001$, $I^2 = 97\%$), with considerable between-study ($\tau^2 = 0.041$) and within-study ($\omega^2 = 0.008$) variance components, indicating marked variation in lottery effects across different contexts, school types, and student populations.

The null overall effect of winning school admission voucher was consistent across academic domains (see Appendix 2, Figure 2.2), with mathematics showing negligible effect ($g = -0.03$, 95% CI $[-0.18, 0.23]$, $p = 0.78$, $k = 10$) and reading showing a small positive but insignificant effect ($g = 0.11$, 95% CI $[-0.01, 0.23]$, $p = 0.06$, $k = 10$). Both domains exhibited substantial heterogeneity (mathematics: $Q(9) = 78.74$, $p < 0.0001$, $I^2 = 94.56\%$; reading: $Q(9) = 106.94$, $p < 0.0001$, $I^2 = 91.68\%$), further underscoring the context-dependent nature of school choice effects. This high heterogeneity contrasts sharply with the homogeneous effects observed for early tracking ($I^2 = 0\%$), suggesting that lottery-based school choice produces highly variable outcomes that depend critically on local educational contexts, school characteristics, and student populations. However, winning a school admission voucher through a lottery shows divergent effects across achievement levels (see Figure 6). High-achieving students (top third of baseline performance distribution) experience small negative but statistically insignificant effects ($g = -0.06$, 95% CI $[-0.27, 0.15]$, $p = 0.53$, $k = 15$) on their test scores. In contrast, low-achieving students (≤ 25 th percentile or PISA Level < 2) gain a moderate positive statistically insignificant effect on their test scores ($g = 0.21$, 95% CI $[-0.025, 0.45]$, $p = 0.07$, $k = 8$). The diverging impact of winning admission vouchers through lottery on students' test outcomes by achievement level suggests that lottery-based school admission voucher functions more as an equity-enhancing mechanism than as a tool to foster academic excellence, benefiting struggling students while showing no benefits or potentially small negative effects for high-achieving students. Our findings further support the view that, in lottery-based systems, students at the upper end of academic achievement (and whose parents reveal a strong preference for academic quality) do not necessarily experience significant achievement gains. Consistent with the interpretation offered by

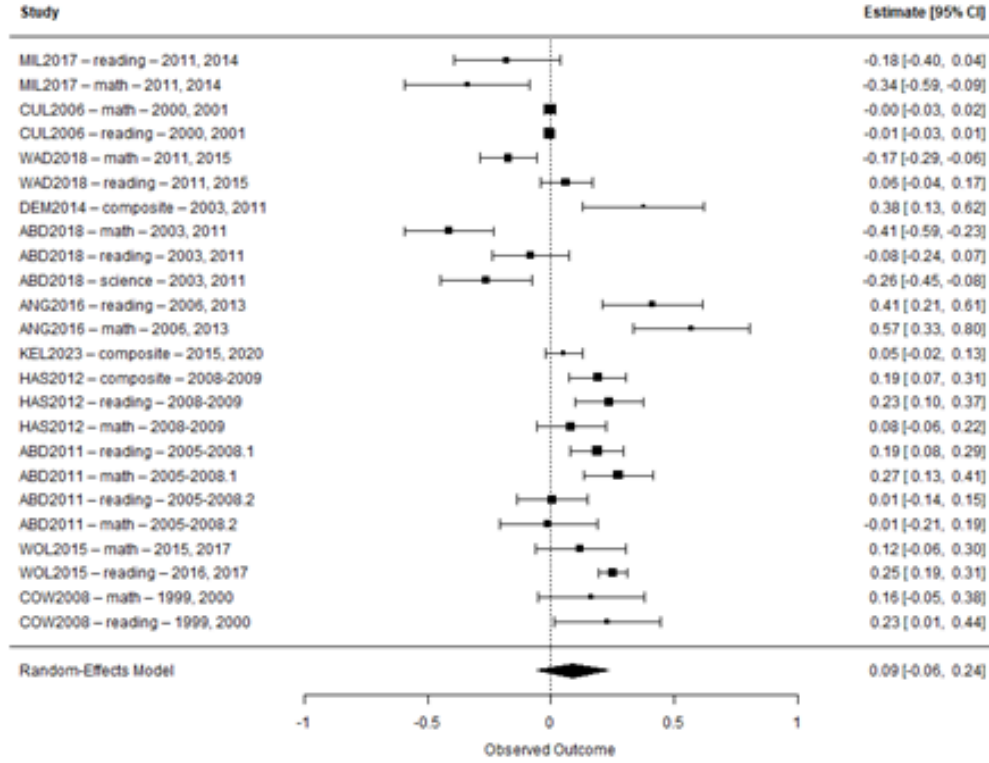


Figure 5: Grand mean effect of winning an admission lottery on students' achievement

Notes: Each study is uniquely labelled by study ID (first three letters of the author's name plus year of publication), domain (mathematics, science, or reading) from which the effect size is estimated, and the survey year(s) from which the primary study conducted its analysis. Each square represents an individual effect size (Hedges' g), with horizontal lines indicating 95% confidence intervals. The diamond represents the pooled random-effects estimate. Confidence intervals are cluster robust to account for multiple estimates per study. Negative values indicate lower achievement in early-tracking systems.

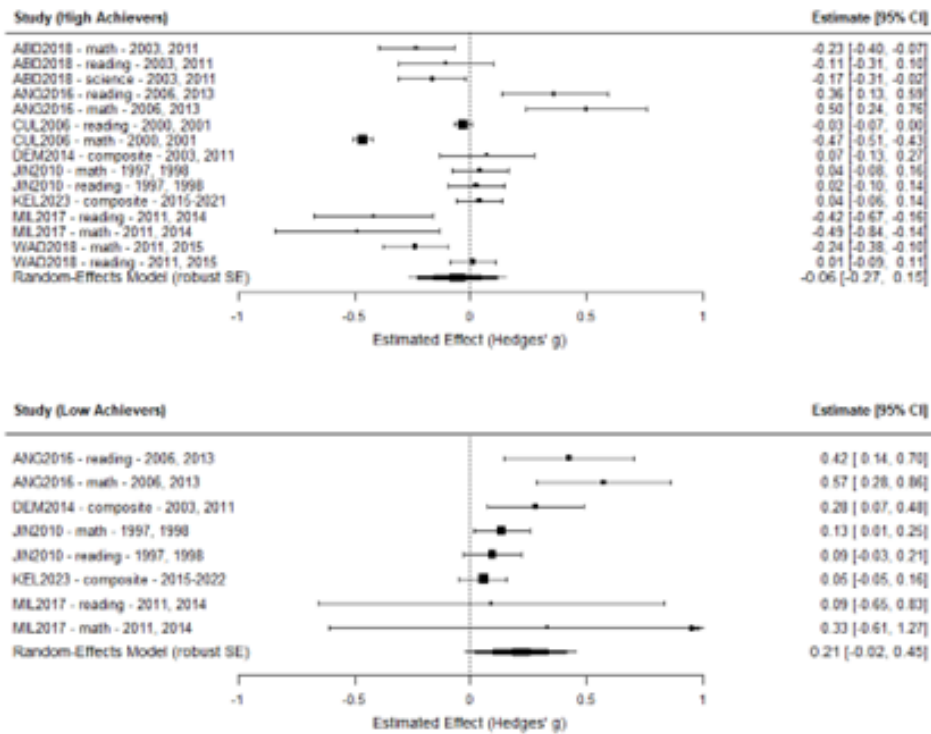


Figure 6: Heterogeneous grand mean effect of winning an admission lottery on students' achievement

Notes: Each study is uniquely labelled by study ID (first three letters of the author's name plus year of publication), domain (mathematics, science, or reading) from which the effect size is estimated, and the survey year(s) from which the primary study conducted its analysis. Each square represents an individual effect size (Hedges' g), with horizontal lines indicating 95% confidence intervals. The diamond represents the pooled random-effects estimate. Confidence intervals are cluster-robust to account for multiple estimates per study. Negative values indicate lower achievement in early-tracking systems.

Hastings et al. (2006), that the lottery assignment expands access for students with weaker prior motivation or, alternatively, that winning a lottery reduces incentives to exert effort once admission to a preferred school is secured.

4.4 Robustness of the results

We report our baseline finding of grand mean effects in Table 1, complemented with RVE estimates (Model 2). In general, RVE estimates show the same pattern of non-significant effects for both high achievers ($g = 0.100$, $p = 0.295$) and low achievers ($g = 0.023$, $p = 0.245$), though with somewhat wider confidence intervals reflecting the more conservative variance estimation approach. The substantive conclusions are unchanged: early tracking produces small negative average effects, with no clear evidence of harm to students at either tail of the achievement distribution. The results are also robust to RVE methods in the case of elite tracking. The overall effect remains small and statistically insignificant ($g = 0.053$, 95% CI $[-0.037, 0.144]$, $p = 0.163$), with nearly identical heterogeneity ($I^2 = 83.47\%$). Domain-specific RVE estimates likewise show no significant effects: reading ($g = 0.021$, $p = 0.349$) and mathematics ($g = 0.127$, $p = 0.159$). The consistency of point estimates and inference across estimation methods strengthens confidence that elite tracking produces small, insignificant effects for students near admission cutoffs, though substantial heterogeneity across contexts remains.

And finally, also in the case of randomised school choice, the robustness checks using RVE yield substantively identical conclusions. The overall effect remains null ($g = 0.080$, 95% CI $[-0.052, 0.213]$, $p = 0.207$) with comparably high heterogeneity ($I^2 = 86.36\%$). Domain-specific estimates are consistent: mathematics ($g = 0.010$, $p = 0.914$, $I^2 = 88.32\%$) and reading ($g = 0.111$, $p = 0.085$, $I^2 = 92.12\%$). Critically, the divergent pattern by achievement level persists: high achievers show small negative effects ($g = -0.052$, $p = 0.566$, $I^2 = 96.35\%$) while low achievers show positive effects approaching significance ($g = 0.189$, $p = 0.0997$, $I^2 = 63.20\%$). The convergence of findings across estimation methods, despite different variance assumptions and degrees-of-freedom corrections, provides strong evidence that lottery-based school choice produces heterogeneous effects that favour low achievers while offering no clear benefits to high achievers.

4.5 Publication selection bias and sensitivity analysis

We assessed the potential influence of publication selection using funnel plots and regression-based tests of funnel plot asymmetry across all three school selectivity mechanisms (Figure 7). Visual inspection of the funnel plots did not reveal systematic asymmetry for any of the interventions, and formal regression tests were consistent with this impression.

For school choice via admission lottery vouchers, the distribution of study estimates was broadly symmetric around the pooled effect. The regression test showed no evidence of funnel plot asymmetry. Trim-and-fill analysis suggested two potentially missing studies on the left side, resulting in a slightly smaller adjusted estimate, though the direction remained the same. Between-study heterogeneity was high, indicating substantial variation in effects across studies. These results suggest that the estimated effects of school

choice by winning admission vouchers on students' test scores are unlikely to be substantially distorted by selective publication.

A similar pattern emerged for early tracking. The funnel plot showed no clear evidence of asymmetry, and the regression test confirmed no significant funnel plot asymmetry. Trim-and-fill analysis imputed two studies on the right side, yielding a nearly identical adjusted estimate. Between-study heterogeneity was low, suggesting high consistency across studies. These findings indicate that the main conclusions regarding early tracking are robust to plausible forms of publication selection.

To further probe potential heterogeneity in publication bias, we conducted separate analyses by achievement group and selectivity mechanism (Figure 8). For the lottery-based admission voucher, the funnel plot for high-achieving students did not reveal systematic asymmetry and trim-and-fill analysis imputed zero missing studies on both sides. For low-achieving students, the funnel plot showed some visual asymmetry, and the regression test revealed statistically significant funnel plot asymmetry. However, trim-and-fill procedures imputed zero missing studies on both sides, with the pooled estimate remaining unchanged. Despite the detected asymmetry among low achievers, the substantive finding of positive effects for this group persisted across sensitivity analyses.

For elite tracking, the funnel plot showed reasonable symmetry around the pooled effect, and the regression test revealed no evidence of funnel plot asymmetry. Trim-and-fill analysis imputed three potentially missing studies on the right side, resulting in a slightly larger adjusted estimate, though the direction and general magnitude remained similar. Between-study heterogeneity was high, reflecting considerable variation across studies. While the adjusted estimate shifted slightly upward, the overall pattern suggests that publication selection is unlikely to fundamentally alter conclusions about elite tracking effects.

For early tracking, the subgroup analyses suggested slightly more scope for asymmetry, particularly among studies focusing on high-achieving students. Trim-and-fill procedures imputed some potentially missing studies, primarily on the right side of the funnel. However, even after these adjustments, the pooled estimates for both high- and low-achieving students remained small in magnitude and close to the original estimates. Importantly, the substantive interpretation of the results—namely, that early tracking effects are modest and do not reverse sign after accounting for possible publication selection—was unchanged.

Table 1: Main Effects and Robustness Checks: Three-Level Random-Effects Models (Model 1) and Robust Variance Estimation (Model 2)

Outcome		Model 1						Model 2						
		Est.	Lower	Upper	τ^2	ω^2	I^2 (%)	Est.	Lower	Upper	τ^2	I^2 (%)	k	m
ALL	Early tracking	-0.100*** (0.007)	-0.118	-0.082	0.000	0.001	11.54	-0.099** (0.010)	-0.148	-0.049	0.000	4.04	21	6
	Winning a lottery	0.088 (0.066)	-0.060	0.236	0.041	0.008	97.00	0.080 (0.059)	-0.052	0.213	0.017	86.36	24	11
	Elite tracking	0.051 (0.028)	-0.026	0.128	0.001	0.005	83.33	0.053 (0.030)	-0.037	0.144	0.005	83.47	12	5
READING	Early tracking	-0.150*** (0.010)	-0.172	-0.126	0.000	0.000	0.00	-0.147** (0.010)	-0.207	-0.087	0.000	0.00	10	6
	Winning a lottery	0.110 (0.051)	-0.006	0.227	0.000	0.022	91.68	0.111* (0.056)	-0.019	0.241	0.023	92.12	10	9
	Elite tracking	0.021 (0.012)	-0.019	0.060	0.000	0.000	0.00	0.021 (0.016)	-0.052	0.093	0.000	0.00	5	4
MATH	Early tracking	-0.068** (0.012)	-0.100	-0.035	0.000	0.000	0.00	-0.068 (0.016)	-0.185	0.049	0.000	0.00	6	5
	Winning a lottery	0.026 (0.087)	-0.176	0.227	0.000	0.069	94.56	0.010 (0.089)	-0.197	0.216	0.031	88.32	10	9
	Elite tracking	0.124 (0.063)	-0.077	0.324	0.013	0.000	84.63	0.127 (0.065)	-0.097	0.35	0.014	86.54	5	4
SCIENCE	Early tracking	-0.079* (0.016)	-0.128	-0.029	0.000	0.000	0.00	-0.078 (0.021)	-0.251	0.095	0.000	0.00	5	4
HIGH ACHIEVERS	Early tracking	0.056 (0.052)	-0.169	0.281	0.006	0.000	27.20	0.100 (0.055)	-0.409	0.608	0.003	23.92	13	3
	Winning a lottery	-0.058 (0.089)	-0.269	0.153	0.046	0.022	96.41	-0.052 (0.085)	-0.255	0.152	0.097	96.45	15	8
LOW ACHIEVERS	Early tracking	0.017 (0.013)	-0.037	0.070	0.000	0.000	0.00	0.023 (0.009)	-0.090	0.135	0.000	0.00	13	3
	Winning a lottery	0.214 (0.085)	-0.025	0.452	0.027	0.000	75.93	0.189* (0.080)	-0.067	0.446	0.014	63.20	8	5

Notes: Model 1 reports meta-regression results from a three-level random-effects model estimated using restricted maximum likelihood (REML) via the metafor package. Model 2 reports meta-regression results using robust variance estimation (RVE) with method-of-moments variance estimation, implemented in the robumeta package. τ^2 denotes between-study variation in true effect sizes, ω^2 denotes within-study variation in true effect sizes, and I^2 denotes the proportion of total variation attributable to true heterogeneity rather than sampling error. k is the number of effect sizes and m is the number of studies. Both models employ cluster-robust variance estimation (Model 1: CR2 estimator with Satterthwaite degrees of freedom; Model 2: CR1 estimator with residual degrees of freedom). Robust standard errors are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Overall, across mechanisms and achievement groups, the combined evidence from funnel plots, regression diagnostics, and trim-and-fill sensitivity analyses provides little support for the notion that publication selection bias materially influences the results. The central conclusions regarding the effects of lottery-based school choice and early tracking, as well as their heterogeneity by achievement level, therefore appear robust. Beyond publication selection bias, we assessed robustness to alternative meta-analytic estimation approaches. Results from robust variance estimation using method-of-moments variance estimation (robumeta package; Model 2 in Appendix 1, Table A2) are substantively identical to our primary three-level random-effects models estimated with REML (metafor package; Model 1). Point estimates differ by no more than 0.02 standard deviations across all comparisons, confidence intervals show substantial overlap, and statistical inference (significant vs. non-significant) is consistent for all main effects and most subgroup effects. This convergence across modelling approaches, combined with limited evidence of publication bias, strengthens confidence in the robustness of our findings.

5 Conclusions

Motivated by dropping students’ outcomes globally and especially in the case of high-achievers, we studied the meta-effects of the three selective mechanisms in education – early (formal) tracking, elite (school) tracking and assignment by lottery. We asked whether the heterogeneous effects by achievement levels of these mechanisms differ, so, unintentionally, by aligning policy to the mean student, we might hurt some achievement groups of students? The first mechanism – formal and early tracking – is applied by many continental European countries at the system level. From the early 2000s, the negative mean effects of tracking have been well reported, and many countries have postponed the age of first tracking. As tracking is a system-level phenomenon, most causal studies apply DiD as a methodology with synthetic controls and report average treatment effects on treated (e.g. tracked to lower track). Relying on 7 studies, we show negative mean effects of tracking, while tracking can be beneficial for both high- and low-performing students. So, neither high-achieving nor low-achieving students showed clear harms from early tracking, suggesting that the overall negative effect is concentrated among middle-achieving students. – a group often overlooked in tracking debates that focus on the extremes of the achievement distribution. So the policy advice is not straightforward – with de-tracking policy, the academic outcomes of low- and high-achieving students can diminish. These findings should not be interpreted as an argument for early tracking as a policy instrument per se. Rather, they indicate that instructional differentiation—when credibly implemented—can generate gains at both ends of the achievement distribution. Whether early tracking delivers such differentiation in practice remains an institutional and political question.

The second selective mechanism, elite tracking, defined as admission to selective public or private schools based on prior achievement or entrance examinations, presents a markedly different pattern. In contrast to early tracking, average treatment effects cannot be estimated; instead, we rely on (local) treatment effects on treated, that is,

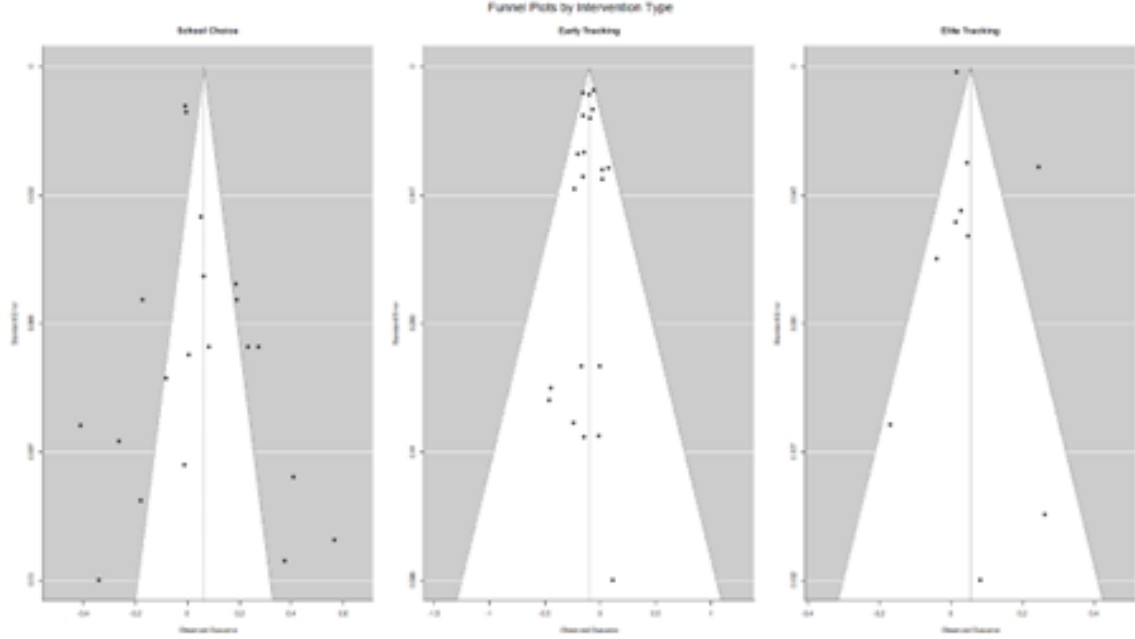


Figure 7: Funnel plots and publication bias assessment by selection mechanism

Notes: Funnel plots display observed effect sizes (Hedges' g) against their standard errors for studies of lottery voucher-based school choice ($k = 240$), early tracking ($k = 21$) and elite tracking ($k = 12$). Vertical dashed lines indicate pooled random-effects estimates; diagonal lines represent 95% pseudo-confidence limits. Regression tests for funnel plot asymmetry (Egger-type tests with standard error as predictor) showed no evidence of asymmetry for lottery voucher-based choice ($z = 0.23$, $p = 0.82$), early tracking ($z = -0.36$, $p = 0.72$) or elite tracking ($z = -0.11$, $p = 0.91$). Trim-and-fill analyses suggested modest adjustments: for school choice by winning admission voucher through lottery, two potentially missing studies on the left yielded $g = 0.037$; for early tracking, two missing studies on the right yielded $g = 0.10$; for elite tracking, three potentially missing studies on the right yielded $g = 0.074$. Heterogeneity patterns varied substantially: high for school choice ($I^2 = 96.61\%$) and elite tracking ($I^2 = 84.38\%$), but low for early tracking ($I^2 = 10.58\%$).

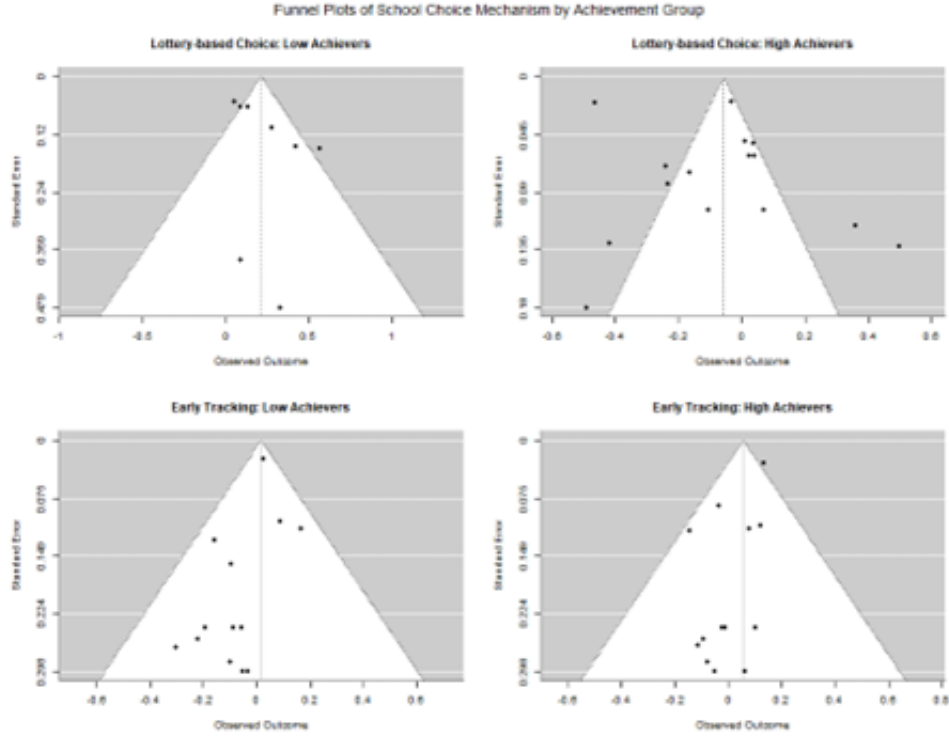


Figure 8: Funnel plots and publication bias assessment by selection mechanism and achievement level

Notes: Funnel plots display observed effect sizes (Hedges' g) against standard errors separately by achievement group and selectivity mechanism. For lottery-based choice, regression tests indicated no funnel plot asymmetry for high achievers ($z = 0.28$, $p = 0.78$; limit estimate $b = -0.11$, 95% CI $[-0.41, 0.18]$; $k = 15$) but statistically significant asymmetry for low achievers ($z = 2.63$, $p = 0.009$; $b = 0.01$, 95% CI $[-0.10, 0.12]$; $k = 8$). Trim-and-fill analyses imputed zero missing studies for both subgroups, with adjusted pooled effects identical to the original estimates (high achievers: $g = -0.07$, 95% CI $[-0.24, 0.06]$; low achievers: $g = 0.22$, 95% CI $[0.08, 0.35]$). For early tracking, regression tests showed no statistically significant asymmetry (high achievers: $z = -1.73$, $p = 0.08$; $b = 0.13$, 95% CI $[0.04, 0.23]$; low achievers: $z = -1.48$, $p = 0.14$; $b = 0.05$, 95% CI $[-0.03, 0.12]$; $k = 13$ in both groups). Trim-and-fill procedures suggested potential missing studies on the right side of the funnel, imputing seven studies for high achievers (adjusted $g = 0.12$, 95% CI $[0.03, 0.20]$; $k = 20$) and six studies for low achievers (adjusted $g = 0.03$, 95% CI $[-0.01, 0.07]$; $k = 19$). Original pooled effects prior to adjustment were small for both groups (high achievers: $g = 0.04$, 95% CI $[-0.04, 0.13]$; low achievers: $g = 0.02$, 95% CI $[-0.02, 0.06]$). Heterogeneity remained low to moderate across all subgroup analyses.

for students who marginally cross the admission threshold. The estimation strategy in this case is exclusively RDD design. Elite tracking can occur at different transition points, including from primary to lower-secondary education or from lower-secondary to upper-secondary education, but in all cases the identified effects are local by construction. Across studies, these local effects are on average small and positive for treated students, indicating modest gains from access to elite schools for marginal admitted students. At the same time, the literature consistently documents zero or negative effects for non-treated students who remain in non-elite schools, implying that elite tracking redistributes educational opportunities rather than uniformly raising achievement. As a result, elite tracking does not generate system-wide efficiency gains comparable to those hypothesized for early tracking. Instead, its consequences are inherently distributional, benefiting a narrow group of students at the admission margin while potentially disadvantaging those who are left behind. From a policy perspective, elite tracking therefore represents a selective expansion of opportunities rather than a Pareto-improving reform.

Third, school choice policies are often promoted to enhance educational outcomes through competition and alignment with student preferences. To increase access to a preferred school, lotteries (in the USA) are increasingly used. We identified nine studies, mostly applying instrumental variable regressions, and show that random allocation is neutral to mean outcomes and does not inherently support high achievement. We report the average causal effect for compliers – LATE. Similarly to our general conclusion, Abdulkadiroğlu et al. (2018) and Terrin & Triventi (2023) highlight that the introduction of school choice mechanisms (e.g. lotteries, vouchers or open enrolment) can redistribute students across schools, but do not consistently lead to better academic outcomes for the average or high-performing student when the choice process is equity-driven and randomised.

However, the execution and broader context of school choice policies can radically alter their effect on high achievers. In quasi-markets, where both public and private schools compete for students, and parents act as "consumers", school choice tends to amplify stratification. Evidence (Zimmer et al., 2010, Rosenqvist & Brandén, 2025) shows that schools increasingly differentiate themselves in terms of their performance, discipline or selectivity, and high-achieving students – often from more advantaged backgrounds – self-select into elite institutions. In such contexts, school choice turns from equity-driven to a selective mechanism (elite tracking) where information asymmetries and parental engagement can produce highly unequal outcomes. What if the allocation principle is not lottery, but deferred acceptance algorithm (see the methodology debates on complexities of estimation in Abdulkadiroğlu et al. 2017), *ibid* (2017) show that such a central allocation benefits student in all three “PISA disciplines” on average (no heterogeneous effects estimated). So, it can be argued that the negative effect of school random allocation can be explained by peer effects (Ding & Lehrer, 2007; more literature).

Several limitations should be acknowledged when interpreting our findings. First, although we apply multilevel meta-analytic methods, the available causal evidence remains limited in scope, particularly for elite tracking and for heterogeneous effects by

achievement level, where the number of contributing studies is small. This constrains statistical power and limits the precision of subgroup estimates, especially for low-achieving students in lottery-based school choice. Second, the studies included often span from homogeneous institutional contexts, which may limit the external validity of pooled estimates for any specific education system, especially in the case of randomised lotteries. Third, our analysis focuses on short- to medium-term achievement outcomes in mathematics, reading, and science; longer-term outcomes such as educational attainment, labour-market performance, or non-cognitive skills are not systematically captured. Finally, while our meta-analytic framework identifies average and distributional effects, it cannot fully disentangle underlying mechanisms – such as peer effects, instructional differentiation, or behavioural responses to selection – whose relevance likely varies across settings.

References

- Abdulkadiroğlu, A., Pathak, P. A., & Roth, A. E. (2005a). The New York City high school match. *American Economic Review*, 95(2), 364–367.
- Abdulkadiroğlu, A., Pathak, P. A., Roth, A. E., & Sönmez, T. (2005b). The Boston public school match. *American Economic Review*, 95(2), 368–371.
- Abdulkadiroğlu, A., Angrist, J., & Pathak, P. (2014). The elite illusion: Achievement effects at Boston and New York exam schools. *Econometrica*, 82(1), 137–196. <https://doi.org/10.3982/ECTA10266>
- Abdulkadiroğlu, A., Pathak, P. A., & Walters, C. R. (2018). Free to choose: Can school choice reduce student achievement? *American Economic Journal: Applied Economics*, 10(1), 175–206. <https://doi.org/10.1257/app.20160634>
- Abdulkadiroğlu, A., Angrist, J. D., Narita, Y., & Pathak, P. A. (2017). Research design meets market design: Using centralized assignment for impact evaluation. *Econometrica*, 85(5), 1373–1432.
- Abdulkadiroğlu, A., Angrist, J., Dynarski, S., Kane, T., & Pathak, P. (2011). Accountability and flexibility in public schools: Evidence from Boston’s charters and pilots. *Quarterly Journal of Economics*, 126(2), 699–748.
- Ammermüller, A. (2005). Educational opportunities and the role of institutions (ZEW Discussion Paper No. 05-44). Centre for European Economic Research.
- Angrist, J. D., Pathak, P. A., & Walters, C. R. (2013). Explaining charter school effectiveness. *American Economic Journal: Applied Economics*, 5(4), 1–27.
- Angrist, J. D., Cohodes, S. R., Dynarski, S. M., Pathak, P. A., & Walters, C. R. (2016). Stand and deliver: Effects of Boston’s charter high schools on college preparation, entry, and choice. *Journal of Labor Economics*, 34(2), 275–318.
- Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R:

- A step-by-step tutorial. *The Quantitative Methods for Psychology*, 12(3), 154–174. <https://doi.org/10.20982/tqmp.12.3.p154>
- Barone, C. (2019). Towards an education-based meritocracy? Why modernisation and social reproduction theories cannot explain trends in educational inequalities. *ISA eSymposium for Sociology*, 9(1).
- Barone, C., Assirelli, G., Abbiati, G., Argentin, G., & De Luca, D. (2018). Social origins, relative risk aversion and track choice: A field experiment on the role of information biases. *Acta Sociologica*, 61(4), 441–459. <https://doi.org/10.1177/0001699317729872>
- Bol, T., Witschge, J., Van de Werfhorst, H., & Dronkers, J. (2014). Curricular tracking and central examinations: Counterbalancing the impact of social background on student achievement in 36 countries. *Social Forces*, 92(4), 1545–1572. <https://doi.org/10.1093/sf/sou003>
- Brunello, G., & Checchi, D. (2007). Does school tracking affect equality of opportunity? New international evidence. *Economic Policy*, 22(52), 782–861.
- Bygren, M., & Rosenqvist, E. (2020). Elite schools, elite ambitions? The consequences of secondary-level school choice sorting for tertiary-level educational choices. *European Sociological Review*, 36(4), 594–609. <https://doi.org/10.1093/esr/jcaa008>
- Chabrier, J., Cohodes, S., & Oreopoulos, P. (2016). What can we learn from charter school lotteries? *Journal of Economic Perspectives*, 30(3), 57–84.
- Clark, D. (2010). Selective schools and academic achievement. *The B.E. Journal of Economic Analysis & Policy*, 10(1), 1–40.
- Cowen, J. M. (2008). School choice as a latent variable: Estimating the “complier average causal effect” of vouchers in Charlotte. *Policy Studies Journal*, 36(2), 301–315.
- Cullen, J. B., Jacob, B. A., & Levitt, S. (2006). The effect of school choice on participants: Evidence from randomized lotteries. *Econometrica*, 74(5), 1191–1230.
- Deming, D. J., Hastings, J. S., Kane, T. J., & Staiger, D. O. (2014). School choice, school quality, and postsecondary attainment. *American Economic Review*, 104(3), 991–1013. <https://doi.org/10.1257/aer.104.3.991>
- Ding, W., & Lehrer, S. F. (2007). Do peers affect student achievement in China’s secondary schools? *Review of Economics and Statistics*, 89(2), 300–312.
- Dobbie, W., & Fryer, R. G., Jr. (2013). Getting beneath the veil of effective schools: Evidence from New York City. *American Economic Journal: Applied Economics*, 5(4), 28–60.
- Dustan, A., de Janvry, A., & Sadoulet, E. (2017). Flourish or fail? *Journal of Human Resources*, 52, 756–799.
- Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method

- of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2), 455–463. <https://doi.org/10.1111/j.0006-341x.2000.00455.x>
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629–634. <https://doi.org/10.1136/bmj.315.7109.629>
- Estrada, R., & Gignoux, J. (2017). Benefits to elite schools and the expected returns to education: Evidence from Mexico City. *European Economic Review*, 95, 168–194.
- European Commission, Directorate-General for Education, Youth, Sport and Culture. (2022). *Investing in our future: Quality investment in education and training*. Publications Office of the European Union. <https://doi.org/10.2766/45896>
- Friedman, M. (1997). Public schools: Make them private. *Education Economics*, 5(3), 341–344.
- Greaves, E., Wilson, D., & Nairn, A. (2023). Marketing and school choice: A systematic literature review. *Review of Educational Research*, 93(6), 825–861. <https://doi.org/10.3102/00346543221141658>
- 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.
- Hanushek, E. A., West, M. R., & Woessmann, L. (2013). *School accountability, autonomy, and choice around the world*. University of Chicago Press.
- Hastings, J. S., Kane, T. J., & Staiger, D. O. (2006). Gender and performance: Evidence from school assignment by randomized lottery. *American Economic Review*, 96(2), 232–236.
- Hastings, J. S., Neilson, C. A., & Zimmerman, S. D. (2012). The effect of school choice on intrinsic motivation and academic outcomes (NBER Working Paper No. 18324). National Bureau of Economic Research.
- Hedges, L. V. (1981). Distribution theory for Glass’s estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2), 107–128. <https://doi.org/10.3102/10769986006002107>
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1(1), 39–65. <https://doi.org/10.1002/jrsm.5>
- 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.
- 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. <https://doi.org/>

10.1007/978-90-481-3993-4-3

- Jin, H., Barnard, J., & Rubin, D. B. (2010). A modified general location model for noncompliance with missing data. *Journal of Educational and Behavioral Statistics*, 35(2), 154–173. <https://doi.org/10.3102/1076998609346968>
- Ketel, N., Oosterbeek, H., Sóvágó, S., & van der Klaauw, B. (2023). The (un)importance of school assignment (Tinbergen Institute Discussion Paper No. TI 2023-076/V).
- Lavrijsen, J., & Nicaise, I. (2016). Educational tracking, inequality and performance: New evidence from a differences-in-differences technique. *Research in Comparative and International Education*, 11(3), 334–349. <https://doi.org/10.1177/1745499916664818>
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Sage Publications.
- Matthewes, S. H. (2021). Better together? Heterogeneous effects of tracking on student achievement. *Economic Journal*, 131(635), 1269–1307.
- Mills, J. N., & Wolf, P. J. (2017). Vouchers in the Bayou: The effects of the Louisiana Scholarship Program on student achievement after 2 years. *Educational Evaluation and Policy Analysis*, 39(3), 464–484. <https://doi.org/10.3102/0162373717693108>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Pekkarinen, T., Uusitalo, R., & Pekkala, S. (2006). Education policy and intergenerational income mobility: Evidence from the Finnish comprehensive school reform (IZA Discussion Paper No. 2204). Institute for the Study of Labor.
- Piopiunik, M. (2014). The effects of early tracking on student performance: Evidence from a school reform in Bavaria. *Economics of Education Review*, 42, 12–33. <https://doi.org/10.1016/j.econedurev.2014.06.002>
- Piopiunik, M. (2021). How does reducing the intensity of tracking affect student achievement? Evidence from German state reforms. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3896149>
- Pop-Eleches, C., & Urquiola, M. (2013). Going to a better school: Effects and behavioral responses. *American Economic Review*, 103(4), 1289–1324. <https://doi.org/10.1257/aer.103.4.1289>
- Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, 23(3), 425–438. <https://doi.org/10.1007/s11121-021-01246-3>
- Roller, M., & Steinberg, D. (2020). The distributional effects of early school stratification: Non-parametric evidence from Germany. *European Economic Review*, 125,

103422. <https://doi.org/10.1016/j.euroecorev.2020.103422>
- Rosenqvist, E., & Brandén, M. (2025). School composition and academic decisions. *European Sociological Review*, 41(2), 232–247. <https://doi.org/10.1093/esr/jcae031>
- Schiltz, F., Mazrekaj, D., Horn, D., & De Witte, K. (2019). Does it matter when your smartest peers leave your class? Evidence from Hungary. *Labour Economics*, 59, 79–91. <https://doi.org/10.1016/j.labeco2019.04.001>
- Shakeel, M. D., Anderson, K. P., & Wolf, P. J. (2021). The participant effects of private school vouchers around the globe: A meta-analytic and systematic review. *School Effectiveness and School Improvement*, 32(4), 509–542.
- Shi, Y. (2020). Who benefits from selective education? Evidence from elite boarding school admissions. *Economics of Education Review*, 74, 101907. <https://doi.org/10.1016/j.econedurev.2019.07.001>
- Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. National Academy Press.
- Terrin, É., & Triventi, M. (2023). The effect of school tracking on student achievement and inequality: A meta-analysis. *Review of Educational Research*, 93(2), 236–274.
- Van de Werfhorst, H., & Hofstede, S. (2007). Cultural capital or relative risk aversion? Two mechanisms for educational inequality compared. *British Journal of Sociology*, 58(3), 391–415. <https://doi.org/10.1111/j.1468-4446.2007.00157.x>
- Viechtbauer, W. (2009). *metafor: Meta-analysis package for R* (Version 4.8-0). <https://doi.org/10.32614/CRAN.package.metafor>
- Waddington, R. J., & Berends, M. (2018). Impact of the Indiana Choice Scholarship Program: Achievement effects for students in upper elementary and middle school. *Journal of Policy Analysis and Management*, 37(4), 783–808. <https://doi.org/10.1002/pam.22086>
- Wilson, D., & Bridge, G. (2019). School choice and the city: Geographies of allocation and segregation. *Urban Studies*, 56(15), 3198–3215. <https://doi.org/10.1177/0042098019843481>
- Woessmann, L. (2007). Fundamental determinants of school efficiency and equity: German states as a microcosm for OECD countries (CESifo Working Paper No. 1981). CESifo.
- Wolf, P. J., Egalite, A. J., & Dixon, P. (2015). Private school choice in developing countries: Experimental results from Delhi, India. In S. McGrath & Q. Gu (Eds.), *Routledge handbook of international education and development* (pp. 456–471). Routledge.

Appendix 1 : Description of the studies included in the meta-analysis

1A. Early Tracking

Author	Treatment	Main Findings	Effect Sizes	Country/Age
Matthewes (2021)	Delayed tracking	Delaying tracking improves average test scores; gains concentrated among students in the lower tail of baseline achievement.	2	Germany, Grades 5–7
Piopiunik (2021)	Reduction of tracking intensity (DiD)	Reducing track intensity improves student performance, particularly for lower-performing boys and disadvantaged students.	3	Germany, Grades 5–10
Roller & Steinberg (2020)	Preponement of tracking (DiD, CiC)	Average effects are statistically insignificant. Students in the bottom 20% are negatively affected, while students in the top 20% benefit from early tracking.	5	Germany, Grade 9 (age ~15)
Lavrijsen & Nicaise (2016)	Early tracking before age 15 (DiD)	Early tracking negatively affects average reading performance. Low-achieving students experience clear negative effects, while evidence for high-achievers is inconclusive.	40	Cross-national, ages 10–15
Piopiunik (2014)	Preponement reform (DDD)	Earlier tracking reduced performance in low- and middle-track schools and increased the share of low-performing students in lower tracks.	3	Germany, Grades 4–6
Jakubowski (2010)	Tracking before age 15 (DDD)	No clear negative impact of early tracking on mean achievement or inequality once survey differences and Eastern European effects are controlled for.	12	Cross-national (15–23 countries), ages 9–15

1B. Elite Tracking

Author	Treatment	Main Findings	Effect Sizes	Country/Age
Shi (2020)	Selective boarding school admission (RDD)	Selective public boarding schools generate modest gains in SAT math scores, concentrated among disadvantaged and lower-achieving students; no significant gains for high-achievers.	2	USA, Grades 11–12
Dustan et al. (2017)	Admission to elite public high schools through a centralized exam-based allocation system / fuzzy RDD exploiting admission cut-offs	Admission to an elite high school increases end-of-high-school mathematics test scores for marginally admitted students but substantially raises the probability of high school dropout. The negative effects are strongest for students with weaker prior achievement and for those facing longer commuting distances. Overall, elite school admission involves a trade-off between modest academic gains and a significantly higher risk of dropout.	2	Mexico (Mexico City), Upper secondary school entry (age ~15–18 years)
Abdulkadiroğlu et al. (2014)	Eligibility for and attendance at elite public exam schools based on admission cut-offs / fuzzy RDD (exam score thresholds used as an instrument for attendance)	Attending an elite exam school has no statistically significant effect on students' achievement on state tests, PSAT/SAT participation and scores, or AP outcomes for marginal applicants. Small positive effects are found for specific subgroups, but overall achievement gains are negligible, suggesting limited value added for borderline students.	4	USA (Boston and New York City), Grades 7–10 entry cohorts (age ~11–15 years)
Pop-Eleches & Urquiola (2013)	Gaining access to a better-ranked secondary school / RDD based on centralised admission score cut-offs	Attending a higher-achievement secondary school improves students' graduation exam scores. Some offsetting short-run behavioural responses occur: parents reduce support, students feel academically marginalised, and more qualified teachers sort toward higher-achieving peers. Top-tercile students benefit more than lower-performing peers.	1	Romania, age 14–19 years
Clark (2010)	Admission to selective (grammar) schools based on entrance exam thresholds / fuzzy RDD exploiting test score cut-offs	Attending a selective grammar school has at most small and statistically insignificant effects on students' test scores, despite large peer differences. However, selective school attendance increases academic course-taking and university enrolment, suggesting effects on longer-term educational trajectories rather than short-term test score performance.	3	United Kingdom (England), Secondary school entry (grammar schools), age ~11–16 years

1C. School choice: random assignment

Author	Treatment	Main Findings	Effect Sizes	Country/Age
Ketel et al. (2023)	Lottery-based assignment (regression)	Lottery losers show no negative academic or behavioural effects despite assignment to less preferred schools. No significant heterogeneous effects by baseline achievement.	3	Netherlands, ages 12–18
Abdulkadiroğlu et al. (2018)	Voucher offer (RCT, IV)	Small positive overall effects of vouchers; negative effects for high-achieving students, suggesting insufficient differentiation for top performers.	6	USA, Grades 6–8
Mills & Wolf (2017)	Voucher lottery (RCT)	Voucher use led to significant negative effects in math and ELA after two years. Negative effects were strongest for initially higher-achieving students.	8	USA (Louisiana), Grades 3–8
Angrist et al. (2016)	Charter lottery (2SLS)	Charter attendance increases pass rates on high-stakes exit exams, improves SAT and AP participation, and shifts students toward four-year colleges. Positive heterogeneous effects are observed.	6	USA (Boston), ages 15–16
Wolf et al. (2015)	Voucher lottery (field experiment)	Voucher effects are positive only in English after two years and concentrated among girls; no significant effects in mathematics or by achievement level.	2	India (Delhi), ages 6–14
Deming et al. (2014)	Lottery to preferred public school (RCT)	Winning a lottery increases college degree completion for girls and improves GPA and college-prep coursework. Gains are concentrated among students accessing higher-quality schools.	3	USA (North Carolina), ages 11–16
Hastings et al. (2012)	Charter/magnet lottery (2SLS)	Substantial test score gains from charter attendance; some positive effects for high-value-added magnet schools.	3	USA, ages 14–16
Abdulkadiroğlu et al. (2011)	Charter/pilot lottery (2SLS)	Large and significant test score gains for charter lottery winners; pilot school effects are small and mostly insignificant.	4	USA (Boston), ages 14–16
Jin et al. (2010)	Voucher lottery (CACE model)	New methodology to address noncompliance and missing data; moderate positive effects for initially low-scoring students, no significant effects for high achievers.	4	USA (NYC), ages 6–14
Cowen (2008)	Voucher lottery (CACE)	CACE estimates produce smaller treatment effects than IV estimates, highlighting the importance of compliance adjustment in voucher evaluations.	2	USA (Charlotte), age 8
Cullen et al. (2006)	Randomised lottery admission (ITT)	Lottery winners attend higher-quality schools but show little consistent improvement in academic outcomes; some gains in non-academic outcomes. Negative effects are observed for high achievers.	4	USA (Chicago), high school students

Appendix 2: Subject-specific effects of selective mechanisms in education

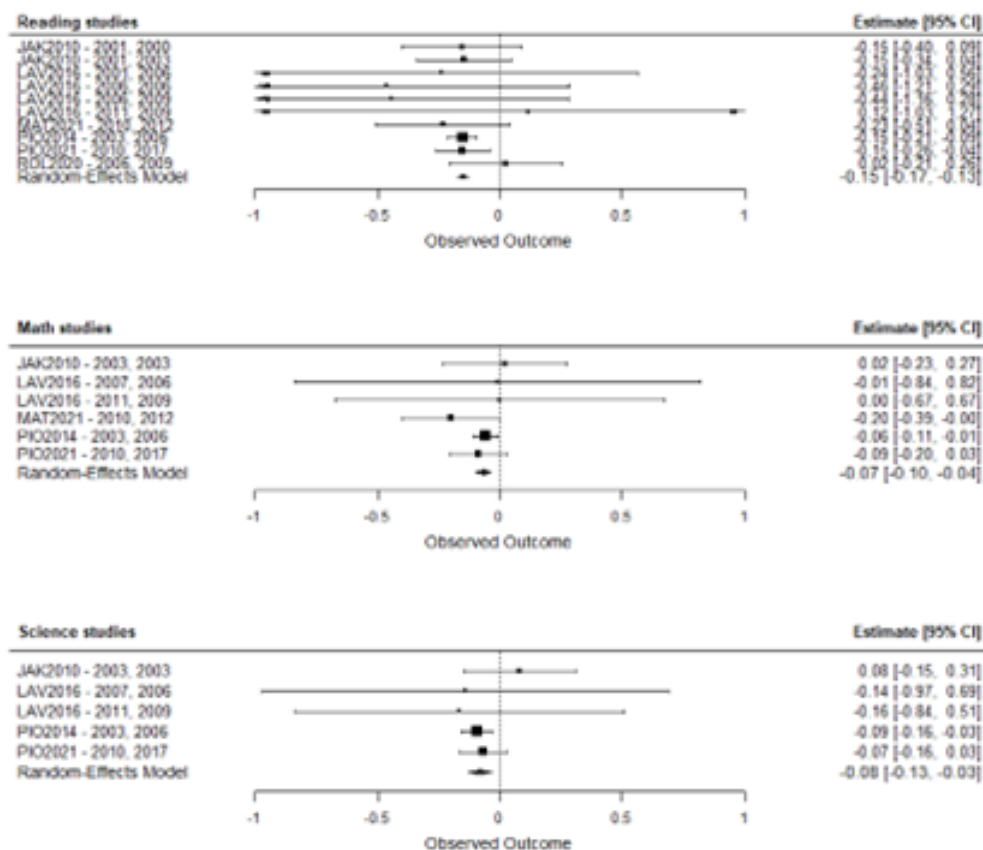


Figure 9: 2A. Grand mean effect of early school tracking on students' performance by domain

Notes: Studies are uniquely labelled by study identifier (first three letters of the first author's surname and year of publication), academic domain (reading, mathematics, or science), and the survey year(s) used in the original analysis. Panels report results separately for reading, mathematics, and science outcomes. Each square represents an individual standardised effect size (Hedges' g), with horizontal lines indicating 95% confidence intervals. The diamond denotes the pooled random-effects estimate within each domain. Confidence intervals are cluster-robust to account for statistical dependence arising from multiple estimates drawn from the same study. Negative values indicate lower achievement in early-tracking systems relative to comprehensive systems.

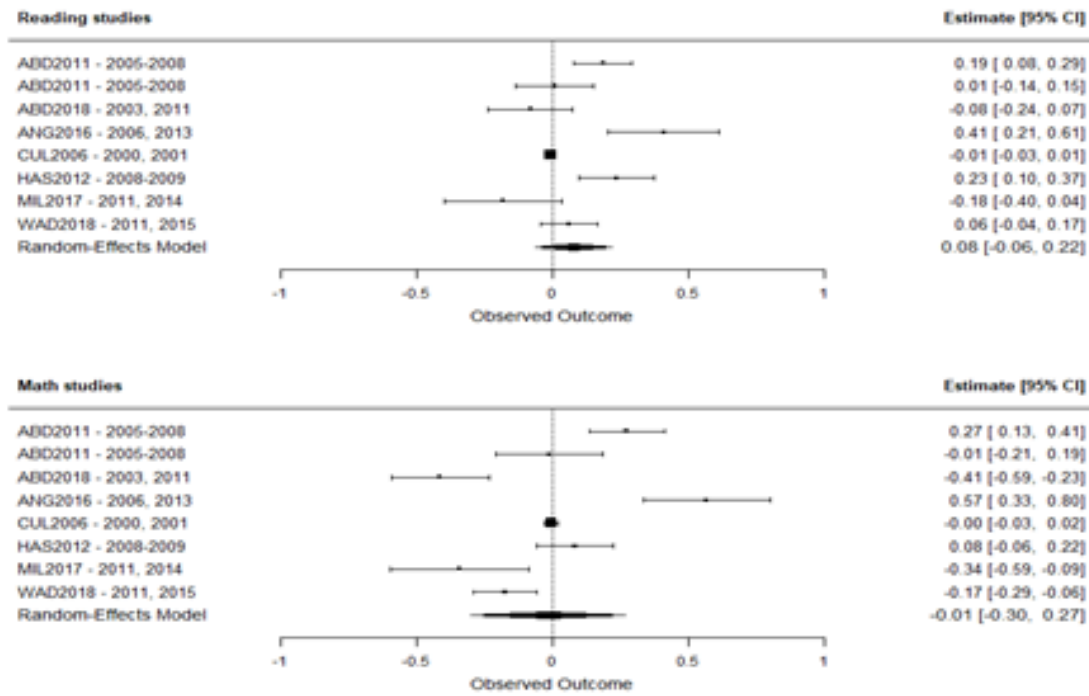


Figure 10: 2B. Grand mean effect of winning the school admission lottery on students' performance by domain

Notes: Studies are uniquely labelled by study identifier (first three letters of the first author's surname and year of publication) and the survey year(s) used in the original analysis. Panels report results separately for reading and mathematics outcomes. Each square represents an individual standardised effect size (Hedges' g), with horizontal lines indicating 95% confidence intervals. The vertical dashed line denotes no effect. The diamond represents the pooled random-effects estimate within each domain. Confidence intervals are cluster-robust to account for statistical dependence arising from multiple estimates drawn from the same study. Positive values indicate higher achievement associated with winning a school admission lottery.

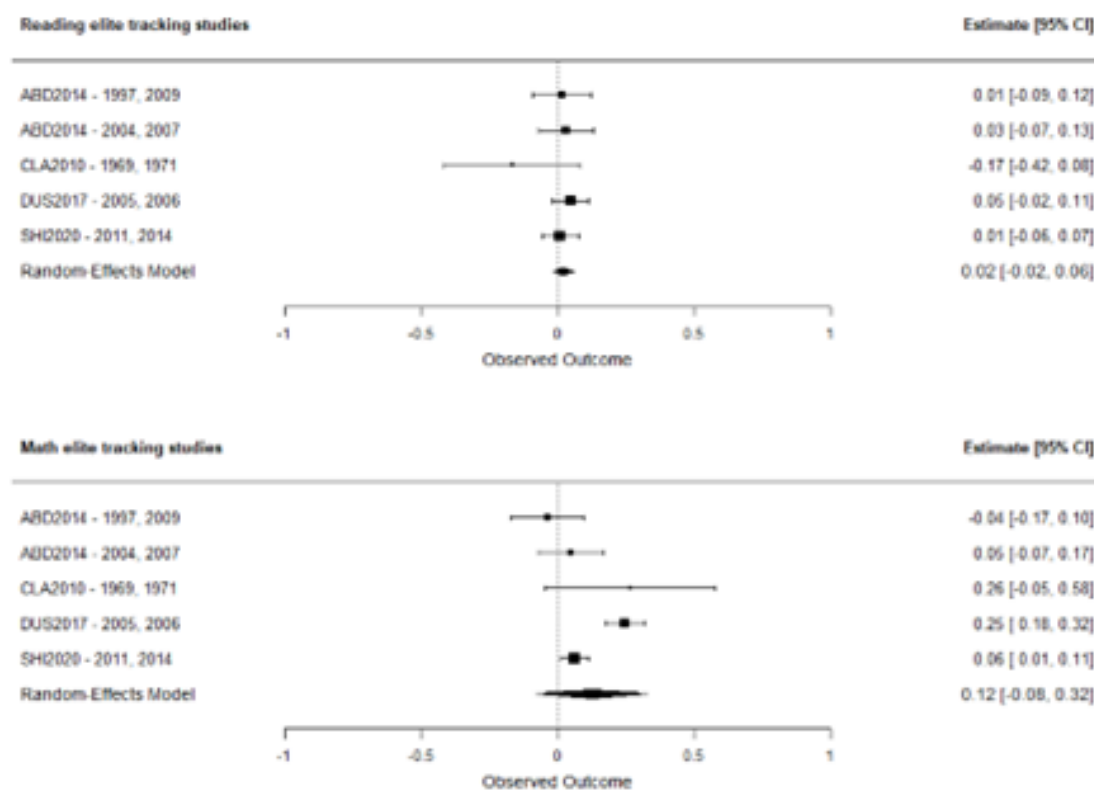


Figure 11: 2C. Grand mean effect of elite tracking on students' performance by domain

Notes: Studies are uniquely labelled by study identifier (first three letters of the first author's surname and year of publication) and the survey year(s) used in the original analysis. Panels report results separately for reading and mathematics outcomes. Each square represents an individual standardised effect size (Hedges' g), with horizontal lines indicating 95% confidence intervals. The vertical dashed line denotes no effect. The diamond represents the pooled random-effects estimate within each domain. Confidence intervals are cluster-robust to account for statistical dependence arising from multiple estimates drawn from the same study. Positive values indicate higher achievement associated with winning a school admission lottery.