Does the track matter? A comparison of students' achievement in different tracks

Dockx J., De Fraine B. & Vandecandelaere M.

KU Leuven

Author Note

This research was funded by the Flemish Policy Research Centre for Education

Correspondence concerning this article should be addressed to Jonas Dockx, Centre for

Educational Effectiveness and Evaluation, Dekenstraat 2 box 3773

3000 Leuven Belgium

Abstract

This study compared the effects of being in different tracks during the first three years of secondary education on student academic performance. A sample of a longitudinal cohort study in Flanders (3205 students in 46 schools) was used to describe the learning gains for mathematics and reading comprehension. Four tracks were distinguished, with a clear hierarchy in mean student academic ability. A comparison was made per pair of tracks that are hierarchically consecutive by matching students who are comparable across these tracks. Three matching methods were applied: propensity score matching, Mahalanobis distance matching and coarsened exact matching. Multilevel latent growth curves and generalized estimating equations were used to describe learning gains over the first three years of secondary education. The results reveal that hierarchically higher tracks benefit students' academic performance for mathematics and reading comprehension. However, effect sizes differ across tracks, time and outcomes. Moreover, differences in relative learning progress between compared tracks diminish over time.

Keywords: tracking, matching, longitudinal studies, mathematics, reading comprehension, growth curves

Educational Impact And Implications Statement

This study investigated whether the track students are allocated to during secondary education matters for their learning progress. We found that being in a more academically focused track benefits students' learning progress for mathematics and reading comprehension. This shows that a structural feature of an education system, having tracks, has tangible consequences for students who are in that education system. Hence, the track does matter.

Does the track matter? A comparison of students' achievement in different tracks

Most education systems track students during secondary school, placing them into educational environments tailored to their abilities and interests (e.g. Hanushek & Wößmann, 2006; OECD, 2012; Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006; Van de Werfhorst & Mijs, 2010). The goal of tracking is to advance academic performance, based on the assumption that an environment fitted to students will lead to efficient education (e.g. Hanushek & Wößmann, 2006; LeTendre, Hofer, & Shimizu, 2003). Hence, this assumed efficiency is a strong rationale for policy makers to either retain or implement some form of tracking.

However, there is a long-standing debate on the merits and detriments of tracking in both political and scientific spheres (e.g. Trautwein et al., 2006). Studies comparing education systems have shown that early tracking increases social inequality in academic performance, while not advancing academic performance of the education system (e.g. Van de Werfhorst & Mijs, 2010). The increasing inequality is usually attributed to lower socioeconomic status students being allocated to lower tracks (e.g. Jackson, Erikson, Goldthorpe, & Yaish, 2007). Due to less focus on performance in these tracks, the differences in student academic performance between social groups are widened over time (e.g. Hanushek & Wößmann, 2006).

Much research on assessing the effects of tracks on academic performance is based on comparing education systems (e.g. Hanushek & Wößmann, 2010; Lavrijsen & Nicaise, 2015; Van de Werfhorst & Mijs, 2010). Such studies describe relations between variables aggregated to the level of the education system, not showing the effects of different tracks within education systems. However, the differences between education systems with different tracking practices are usually attributed to the effects of tracks within education systems (e.g. Hanushek & Wößmann, 2006). Furthermore, there are many differences in tracking practices between

education systems, while comparisons of educational systems generally only distinguish between early and late tracking systems (e.g. OECD, 2012; Trautwein et al., 2006). Directly assessing the effects of tracks within education systems has so far happened only sparingly (e.g. Retelsdorf, Becker, Köller, & Möller, 2012), thus there is a need to ascertain these effects within education systems.

This study is a replication of previous studies on how tracks influence academic performance for different academic performance indicators. It extends former research by assessing how the effects change over time and differ across tracks and students. The current study also compared different matching methods to control for selection bias, relevant for any researcher using matching methods. Both generalized estimating equations and latent growth curves were applied, contrasting them as different approaches of assessing learning progress. In the following sections, we describe tracking literature and the research strategy in more detail.

What is tracking?

Tracking is usually understood as the ability-grouping of students into different educational programs called tracks (e.g. OECD, 2012; Trautwein et al., 2006; Van de Werfhorst & Mijs, 2010). Most education systems track students during secondary education, but they differ in implementation (OECD, 2012, p.57-58). These implementation differences encompass the age when tracking starts (e.g. Brunello & Checchi, 2007; OECD, 2012; Van de Werfhorst & Mijs, 2010), whether student track assignment is based on either ability, interest or even financial resources (e.g. Brunello & Checchi, 2007; Buser, Niederle, & Oosterbeek, 2014; Trautwein et al., 2006), whether track assignment depends on standardized tests (e.g. Bol, Witschge, Van de Werfhorst, & Dronkers, 2014; Tieben, de Graaf, & de Graaf, 2010; Trautwein et al., 2006), the number of tracks (e.g. Thijs Bol & van de Werfhorst, 2013; Brunello & Checchi, 2007), the

differentiation between tracks (e.g. do they have a strong vocational focus or do broad similarities in curriculum remain; e.g. Shavit & Müller, 2000) whether track changes happen often (e.g. Guill, Lüdtke, & Köller, 2016), and whether different tracks exist within classes, each class belongs to a separate track or each school belongs to a separate track (e.g. Trautwein et al., 2006; Van de Werfhorst & Mijs, 2010). Hence, while tracking is a ubiquitous practice of placing students into different educational programs, its specific implementation differs between education systems.

The prevalence of tracking across education systems raises the question why students are tracked. Generally, the intention is to create learning environments tailored to different student groups. For these tracks create more homogeneous student groups, with the opportunity of focusing curricula and teachers on specific learning needs, benefitting academic performance (e.g. Hanushek & Wößmann, 2006). This differentiation of students also allows for skill specialization, which is valued by the labor market (e.g. Bol & van de Werfhorst, 2013; Kerckhoff, 2001; Van de Werfhorst & Mijs, 2010). However, students from different social backgrounds are thereby separated, providing a form of social closure. This institutionalizes social distance between student groups (for a discussion see Bol & van de Werfhorst, 2013, pp. 287-289). In sum, tracks are assumed to improve efficiency of education systems, at the cost of institutionalizing social distance.

The goal of tracks improving efficiency and consequently institutionalizing social distance has guided research into tracking. It ties into the notion that tracking practices matter for student academic performance within education systems. Hence, many studies have focused on how efficiency and equality in academic performance relate to tracking practices of education systems.

Efficiency and equality of education systems

Many studies have compared education systems to discern the effect of tracking on inequality and efficiency, distinguishing between selective systems and comprehensive systems. Generally, the former applies when tracking starts at an early age (10 or 12), whereas the latter applies when tracking starts at a later age (14 or 16). Inequality is operationalized in different ways in different studies (Van de Werfhorst & Mijs, 2010). A first stream of studies used test score dispersions, with some having found that selective systems have larger variances in student outcomes (Huang, van den Brink, & Groot, 2009; Jenkins, Micklewright, & Schnepf, 2008), while others did not (Brunello & Checchi, 2007; Duru-Bellat & Suchaut, 2005; Vandenberge, 2006). A second stream of studies investigated the relationship between socioeconomic status and academic performance across education systems (Van de Werfhorst & Mijs, 2010). Almost all of these studies concluded that selective systems are associated with more social inequality (Ammermüller, 2005; Bauer & Riphahn, 2006; Brunello & Checchi, 2007; Horn, 2009; Marks, 2005; Schütz, Ursprung, & Wößmann, 2008; Wößmann, 2008). A third stream focused on the change in relationship between socioeconomic status and academic performance over time across education systems. Most studies found that social inequality increases more in selective systems (Ammermüller, 2005; Hanushek & Wößmann, 2006; Lavrijsen & Nicaise, 2015), but not all (Waldinger, 2007). None of the three research streams have shown a clear relation between tracking and overall efficiency. Thus, the consensus tilts towards tracking increasing social inequality in education systems, without improving efficiency.

Other studies have investigated how changing tracking practices of education systems changes student outcomes. Most focused on the effects of de-tracking, changing from a selective system to a comprehensive system. Malamud and Pop-Eleches (2011) in Romania found that

students from disadvantaged areas and students from less educated parents more often finished an academic track after the reform. In Finland, Kerr, Pekkarinen and Uusitalo (2013) found small positive effects on verbal test scores, but not in arithmetic and logical reasoning. However, test scores of students whose parents did not receive a high school education did improve.

Furthermore, the effect of the father's income on student performance was reduced. Hall (2012) showed that the amount of upper secondary schooling increased among vocational students in Sweden. In Poland, Jakubowski, Patrinos, Porta and Wiśniewski (2016) found a positive impact on student performance for the whole sample. Piopiunik (2014) evaluated the reverse situation from the former authors, for in Bavaria tracking was hastened by two years. The results indicated a reduction in performance in middle to lower track schools while the number of low-performing students increased in lower track schools. Generally, previous research shows that comprehensive systems benefit disadvantaged students.

However, the distinction between selective and comprehensive systems is somewhat arbitrary, with the age when tracking starts typically used as criterion. Such a distinction does not account for the other aforementioned aspects that characterize the implementation of tracking. Accordingly, studies have found very specific aspects of tracking practices mattering for student outcomes, outside of the comprehensive-selective dichotomy. These encompass whether track assignment is dependent on test scores or is a free choice (Ayalon & Gamoran, 2000; Bol et al., 2014; Tieben et al., 2010), the vocational focus of tracks and if between-school tracking or within-school tracking applies (Van de Werfhorst & Mijs, 2010). Furthermore, how tracks are implemented cannot be completely separated from other characteristics in education systems. For example, Mons (2007) considers tracking as just one method of managing student heterogeneity, next to methods such as ability grouping, grade retention or differentiated teaching (Dupriez,

Dumay, & Vause, 2008). In sum, each education system provides a unique educational context and researchers should be cautious when drawing conclusions from cross-national comparisons. The many dissimilarities between education systems also provide an argument to investigate track effects within education systems.

How does tracking increase social inequality?

Inequality in academic performance due to tracking is typically explained using Boudon's (1974) theory on primary and secondary effects (Hanushek & Wößmann, 2006; Van de Werfhorst & Mijs, 2010). The primary effect entails that higher performing students are more often allocated to higher tracks compared to lower performing students. The secondary effect entails that higher socioeconomic status students are more often allocated to higher tracks than lower socioeconomic status students, even with identical academic performance. There is broad empirical evidence of the existence of both primary and secondary effects (Erikson, Goldthorpe, Jackson, Yaish, & Cox, 2005; Jackson et al., 2007; Kloosterman, Ruiter, De Graaf, & Kraaykamp, 2009). Rational action theory explains secondary effects due to students avoiding intergenerational downward mobility (Breen & Goldthorpe, 1997) and preferring a track with equal or higher status than their parents. Hence, students whose parents followed a lower track are satisfied with a lower track, while students whose parents followed a higher track are not. Moreover, low socioeconomic status students more often get teacher recommendations that orient them towards lower tracks (Boone & Van Houtte, 2013; Ditton, Krüsken, & Schauenberg, 2005).

When students are allocated to different tracks, they are subjected to different learning environments, both by pedagogical design and by peer-group composition. These differences are wide-ranging, concerning student behavior in peer relations, academic focus and aggression

(Barth, Dunlap, Dane, Lochman, & Wells, 2004), an anti-school culture in lower tracks (Van de gaer, Pustjens, Van Damme, & De Munter, 2006), teacher beliefs about their classrooms (Hallam & Ireson, 2003) and teachers' (pedagogical) content knowledge (Baumert et al., 2010).

Furthermore, research has shown that in higher tracks higher levels of problem solving and cognitive activating instructions are given, whereas in lower tracks memorization and disciplining students are emphasized (Kunter & Baumert, 2006; Retelsdorf, Butler, Streblow, & Schiefele, 2010; Van Houtte, 2004). However, we stress that the effects of pedagogical design and peer-group composition are typically inseparable, with pedagogical response resulting from peer-group composition and vice versa (e.g. Gamoran, 1992; Ireson, Hallam, & Plewis, 2001).

Concluding, tracks increase social inequality due to different student-groups being allocated into different tracks and the differences in educational environments provided by those tracks.

Track effects within education systems

Studies addressing the effect of being in a higher or lower track within one education system are relatively rare. The few earlier studies mainly used (multilevel) regression models to gauge track effects, adding confounders (i.e. variables which predict both track allocation and the outcome of interest) as covariates to control for selection bias. With this approach Retelsdorf and Möller (2008) found less gains in reading comprehension in a lower track than in a higher track in Germany. Accordingly, Gustafsson, (2008) used structural equations to test track effects on latent intelligence factors in a Norwegian sample. This study showed that being in a higher track benefits both visual and crystalized intelligence. Also in Germany, Becker (2009) found that students in a higher track gained intelligence compared to students in a lower track.

Recently, several authors critiqued regression-based models in discerning track effects for not satisfactory reducing confounding and potential extrapolation. Instead, these authors promoted quasi-experimental methods to control for confounders. Hence, several authors have used propensity score matching (e.g. Schafer & Kang, 2008) to test track effects on student groups with comparable confounder distributions in different tracks. This method is only applicable in education systems with comparable students across tracks. Three German studies have applied this approach. Becker, Lüdtke, Trautwein, Köller and Baumert (2012) found that students in the higher track have a larger increase in intelligence. Retelsdorf, Becker, Köller, and Möller (2012) found that students in a higher track had larger growth rates for reading decoding speed, although reading comprehension did not differ across tracks. Furthermore, Guill, Lüdtke and Köller (2016) found that students in a higher track showed higher intelligence after four years than students in a lower track. These studies only looked at average effects at a specific timepoint and didn't assess whether track effects may change over time or across student characteristics.

In education systems where there is (almost) no overlap in covariate distributions between students in different tracks, matching cannot be used. This is typical when student allocation into tracks is dependent on test scores, resulting in a threshold value to be assigned to a specific track. In such education systems, (fuzzy) regression discontinuity (e.g. Hahn, Todd, & der Klaauw, 2001) can be used to investigate track effects by comparing students who are close to this threshold value. Students close to this threshold value are considered to be randomly allocated to their track. Using this approach in the Netherlands, Korthals and Dronkers (2016) found that higher track placement affects intelligence and reading skills positively, but no effect

was found for mathematics. Furthermore, Kuzmina and Carnoy (2016) found that gains in academic performance across tracks are more or less equal in four central European countries.

Concluding, most studies found significant track effects on academic performance and intelligence. However, this was not the case for every outcome in every study. This is somewhat surprising, considering that track effects are considered as the key explanation why tracking causes more inequality. Previous studies also did not assess how track effects may change over time or across student characteristics. Consequently, further research on track effects within education systems is required.

The current study

The purpose of this study was to investigate whether being in a higher track affects academic performance within the Flemish education system. This question is apt for Flanders, with PISA 2015 results showing that, out of 57 regions, Flanders has the first to third largest spread of academic performance in mathematics, reading and science for 15-year olds (OECD, 2016, p442-p444). Typically, this inequality is at least partially attributed to its highly selective tracking system (Van Houtte & Stevens, 2015). However, recent moves towards a more comprehensive system have been halted, fearing that this might undermine performance of the whole education system. Hence, any policy discussion on the tracking system closely resembles the academic discussion on the equality-efficiency trade-off due to tracking (for an overview see Brunello & Checchi, 2007, p3-p6). Although it is tempting to apply conclusions from cross-national comparative research to a specific education system, each system's unique characteristics make such inference questionable (Trautwein et al., 2006).

Flemish secondary education has its own characteristic tracking system. Tracking starts at the age of 12, when students leave primary school and have to choose a secondary school

(OECD, 2012, p57). They must choose between the following tracks: classical, modern, technical, and vocational. While the first three tracks share a common core of educational goals during the first two years, each has a unique curriculum. The classical and modern track are academically focused, with students expected to follow tertiary education after these tracks. The classical track will orient towards university while the modern track will more often orient towards college-like tertiary education. The educational programs in the classical track have more content than educational programs in the modern track. The technical track offers pathways towards both college-like tertiary education and the labor market. The vocational track primarily prepares for the labor market. Track is a class-level variable, with most schools offering two tracks, some schools only one track and a few schools three or four tracks. There is no standardized testing in Flanders, hence track choice is completely free if a student has attained a certificate of primary education. A secondary school can reject students who received no certificate from all but the vocational track. Typically, the track choice is a joint decision of the parents and students, often based on the primary school's advice. The tracks have a clear hierarchy in attracting students with different academic abilities and socioeconomic background (Van Houtte, 2004). There is flexibility in changing tracks during the early years of secondary education. Over time, students primarily remain in their track or go down in the hierarchy of tracks (Boone & Van Houtte, 2013). In short, the Flemish education system is typically characterized as a selective tracking system with free track choice.

This study concerned itself with the research question if being in a higher track in Flemish secondary education affects academic performance for students who are in the higher track. To put it otherwise, does the higher track student benefit from being in a higher track? We derive this research question from the literature on tracking where higher track allocation is

considered to benefit higher track students. Any study on these track effects needs to account for the differential intake of students across tracks, a form of selection bias. Therefore, we used a matching approach, matching students who are comparable across tracks, disentangling the effects of tracks and different student selection across tracks. To make our findings more robust we used propensity score matching (Schafer & Kang, 2008), Mahalanobis distance matching (Rosenbaum & Rubin, 1985) and coarsened exact matching (Iacus, King, & Porro, 2012). We also deemed it plausible that differences in academic performance between tracks may change over time, warranting the description of learning growth. Both multilevel latent growth curve models (Duncan, Duncan, & Strycker, 2013) and generalized estimating equations (Hardin & Hilbe, 2003) were used, as these methods allowed us to discern whether the effects of tracks change over time. Furthermore, we gauged if track effects differ across student characteristics.

Our main hypothesis was that higher track placement benefits the higher track students' academic performance. We had no specific hypothesis on how these track effects change over time or across student characteristics. In the following section, the sample and methods are described in more detail.

Making any causal inference rests on several assumptions that, although untestable, should be plausible. We first highlight the ignorable treatment assumption (sometimes called conditional exchangeability assumption, Rosenbaum & Rubin, 1983). In our study this assumption entails that all confounders predicting both track allocation and the outcome of interest should be accounted for. Thus, a track effect can only be attributed to that track if all relevant selection bias due to confounders is removed (Steiner & Cook, 2014). Hence, we require a rich dataset of measured confounders and apt methods to remove the selection bias.

The second assumption we highlight, the stable unit treatment value assumption (Rubin, 1990), entails that a track effect on a student's outcome should only be a function of that student's track allocation. A track effect on a student's outcome should not be influenced by other students' track allocation. However, in practice the track effects are also thought of as partially resulting from the differences in student composition between tracks. If tracks were to suddenly attract different students, changing the student composition in tracks, the track effects will likely change. Hence, track effects in this study should be interpreted under the idea of an individual student having potentially chosen another track, but that differences in student composition across tracks remain stable (Imbens & Rubin, 2015, p.11).

Method

Sample

The research questions were addressed through analyses of the large-scale longitudinal dataset of the LiSO (Dutch acronym for Educational Trajectories in Secondary Education, LiSO-project, 2018) data collection. This ongoing project follows a cohort of 6158 students in 48 schools who started in secondary education in the school year 2013-2014. A regional sampling strategy with complete enumeration was used, meaning that almost all the students belonging to the aforementioned cohort in all the classes in all the schools within a certain area are being studied. Three schools de-track their students in first grade, therefore the 675 students (10.96%) from these schools were excluded from the analyses. Furthermore, 2278 students of the remaining subsample of 5483 students (41.55%) changed track during the first three years of secondary education and were excluded from the analyses as well. This resulted in a subsample of 3205 students in 338 classes in 45 schools at the start of secondary education in September 2013 (the first month of the school year in Flanders). The 84 students (2.62%) who repeated a

grade during the time of the study were kept in the dataset. 691 students were in the classical track, 1285 were in the modern track, 663 students were in the technical track and 566 were in the vocational track. There were slightly more girls (53.73%) than boys in the total sample. 9.86% of students in this sample do not speak Dutch at home, while 21.40% of student parents in this sample are eligible for allowance fees. There were four measurement points throughout secondary education: the start of secondary education in the first grade September 2013 (T0), the end of the first grade May 2014 (T1), the end of the second grade May 2015 (T2) and the end of the third grade May 2016 (T3). Between T0 and T1 there was a time interval of eight months while the subsequent time intervals were twelve months.

At each measurement point students had to complete several tests and fill in a student questionnaire within their own classroom, supervised by their teachers. At the first measurement point the parents of students were also given a parent questionnaire, which they could fill in in their own time. Teachers were also given questionnaires at each measurement point. The tests and questionnaires used in the data collection were specifically developed for the data collection.

Treatment variables

Of main interest was the effect of going to a higher track on academic performance of higher track students, compared to going to a lower track. Our sample contained students from the classical track, modern track, technical track and vocational track. To discern the effect of each track, pairwise comparisons were made of tracks that are consecutive in the hierarchy of tracks. It was not possible to compare nonconsecutive tracks, due to the absence of comparable students. Therefore, the following three comparisons were made: the classical track with the modern track, the modern track with the technical track and the technical track with the vocational track. In each pairwise comparison the hierarchically lower track was the control track

while the hierarchically higher track was the treatment track. Hence, a positive effect would indicate that a hierarchically higher track predicts higher academic performance.

Measures

Outcomes

The outcome of interest was student academic performance. Two measures for academic performance were used: academic performance in mathematics and (Dutch) reading comprehension. Both mathematics and reading comprehension are core parts of the curriculum during the first three years of Flemish secondary education in all tracks.

Mathematics performance was measured at T0, T1, T2 and T3. The number of items ranged from 32 to 42 and encompassed following domains: algebra, geometry, geometric calculation, and data- and information processing. The tests were based on the educational goals set by the government and are considered valid measurements in the Flemish context. Each test had a mix of multiple-choice and open-ended questions. Item Response Theory was used during test development to vertically link, test for differential item functioning and select items in a broad range of difficulty parameters with high discrimination parameters (Embretson & Reise, 2000). Scoring was done using Warm's weighted likelihood estimation (Warm, 1989) based on a two-parameter model. The reliabilities of the tests ranged from 0.83 to 0.87 using Cronbach's Alpha. The retest stability between measurement occasions ranged from 0.65 to 0.75.

Dutch reading comprehension was developed and measured in a similar way as mathematics, however, it was measured only at T0 and T3. The number of items ranged from 32 to 42 and encompassed texts which for a correct answer either required a descriptive analysis, structuring of text elements or personal judgement on different text elements. The texts also varied in complexity of sentence formulation, complexity of text structuring, extensiveness of the

texts, extent of visual support, familiarity of content and if the text was more practical or abstract. This was based on the educational goals set by the government and the tests are considered valid in this context. The reliabilities of the tests were 0.80 and 0.82 using Cronbach's Alpha. The retest stability between both measurement occasions was 0.50.

Baseline covariates

Although including every variable predicting track assignment seems a safe route to reduce selection bias, every included variable decreases the efficiency of the estimators (Golinelli, Ridgeway, Rhoades, Tucker, & Wenzel, 2012; Myers et al., 2011; Pearl, 2010). Hence, matching literature suggests using only those variables that predict both the treatment and the outcome of interest, in this case learning gains (e.g. Brookhart et al., 2006; Myers et al., 2011). Accordingly, we only selected those variables from the total dataset which had a tangible relation with the treatment allocation (whether their track average values differ) and the outcome of interest ($r > 0.05$ across students in two tracks). In the next paragraphs we give an overview of the 25 covariates selected for matching by describing their theoretical backgrounds, the instruments used to measure them, and their reliabilities. Table 1 gives a brief overview of these variables, including their correlations with both mathematics and reading comprehension at T3 and the percentages of missingness. These variables were measured or obtained at T0.

Performance measures

Performance indicators for baseline measurement include the already described measures of mathematics and Dutch reading comprehension at T0. It also included a baseline measurement of French reading comprehension, for French is part of the curriculum in Flemish primary education. The test had 36 items, with each item having a short piece of text (i.e. a restaurant menu or a postcard), differing in complexity of sentence formulation, complexity of text

structuring, extensiveness of the texts, and extent of visual support. The French texts were of a practical nature, with content familiar to the lifeworld of students. The reliability was 0.79 using Cronbach's Alpha.

Gender and age

Two binary variables from governmental records describing the student personal backgrounds were used. The first variable indicates whether the student is a boy. The second variable indicates whether the student is older than normally progressing students (mostly due to grade retention).

Social background

Socioeconomic status was modeled as a latent variable using a Nominal Response Model (Bock, 1972) based on seven indicators: 1) diploma father (10 categories), 2) diploma mother (10 categories), 3) employment status father (4 categories), 4) employment status mother (4 categories), 5) occupational level father (8 categories), 6) occupational level mother (8 categories) and 7) income (7 categories). These indicators were acquired from a parent questionnaire conducted at the start of the data collection. The empirical reliability was 0.87, with Expected A Posteriori (EAP) estimation used for factor scores.

Three binary variables from governmental records describing the social background were used as well. The first variable indicates whether a student's family is eligible for a monthly allowance for their child due low family income. The second variable indicates whether the mother has not attained a certificate of the last year of secondary education. The third variable indicates whether the student does not speak Dutch with more than one family member.

Academic self-concept

Student academic self-concept was measured with measures for general academic self-concept and domain-specific self-concepts for mathematics, Dutch and French. These measures were part of a student questionnaire. General academic self-concept had 4 items, translated to Dutch from the shortened Self-Description Questionnaire II (Marsh, Ellis, Parada, Richards, & Heubeck, 2005).

The items of the domain specific measures are Dutch translations of the Self-Description Questionnaire III (Marsh & O'Neill, 1984), while being reduced to 4 items on perceived competence instead of the original 10 items. Students answered on the items on a five-point response scale, ranging from “not true” to “true”. Using Cronbach’s Alpha, reliabilities for general academic self-concept, self-concept in mathematics, self-concept in Dutch and self-concept in French were .77, .91, .86 and .92 respectively. Maximum A Posteriori (MAP) estimation was used for factor scores.

Motivation

Student motivation was measured with measures for controlled motivation and autonomous motivation, which were part of a student questionnaire. The items of these measures are Dutch translations of the Academic Self-Regulation Questionnaire (SQR-A, Ryan & Connell, 1989). Controlled motivation had 8 items, a reduction of the original 9 items, while autonomous motivation had 4 items, a reduction of the original 7 items. Students answered on the items on a five-point response scale, ranging from “not true” to “true”. Using Cronbach’s Alpha, reliabilities for controlled motivation and autonomous motivation were respectively .81 and .77. MAP estimation was used for factor scores.

Engagement

Engagement was measured with measures for behavioral engagement, emotional engagement, behavioral disengagement and emotional disengagement, which were part of a student questionnaire. They are Dutch translations of an English scale (Skinner, Kindermann, & Furrer, 2009). Behavioral engagement had 5 items, emotional engagement had 4 items, behavioral disaffection had 5 items and emotional disaffection had 6 items. For emotional engagement this meant a reduction of 1 item from the original scale, while for emotional disaffection this meant a reduction of 6 items from the original scale. Students answered on the items on a five-point response scale, ranging from “not true” to “true”. Using Cronbach’s Alpha, reliabilities for behavioral engagement, emotional engagement, behavioral disengagement and emotional disengagement were .78, .77, .68, and .63 respectively. MAP estimation was used for factor scores.

Mindset

A measure of mindset was part of a student questionnaire. It measures whether a student considers intelligence as a fixed trait or an ability that can grow. The scale is a Dutch translation of the Theory of Intelligence Scale (Dweck, Chiu, & Hong, 1995). It had 3 items, with students answering on the items on a five-point response scale, ranging from “not true” to “true”. Using Cronbach’s Alpha, the reliability was .55. Maximum A Posteriori estimation (MAP) was used for factor scores.

School wellbeing

A measure of school wellbeing was part of a student questionnaire. It measures how well a student values daily life at school. The scale is of Dutch origin and often used as a diagnostic tool (Vorst & Smits, 2010). It had 9 items, with students answering on a five-point response scale,

ranging from “not true” to “true”. Using Cronbach’s Alpha, the reliability was .82. MAP estimation was used for factor scores.

Interest

Interest was with measures for mathematics, Dutch, French and technology, as part of a student questionnaire. Each of the scales had 2 items, the first item was on interest and the second on enjoyment. They are Dutch translations of an original English scale (Denissen, Zarrett, & Eccles, 2007). Students answered on a five-point response scale, ranging from “not true” to “true”. Using the Spearman-Brown Coefficient, reliabilities for interest in for mathematics, Dutch, French and technology were respectively .87, .86, .90, and .86. Sum scores were used instead of factor scores.

[Table 1]

Area of common support

A prerequisite for matching students in different tracks is that there should be overlap in the distribution of the baseline covariates across tracks, called the area of common support. Otherwise, no matching procedure can lead to balanced covariates (Austin, 2008). Furthermore, this overlap indicates for which (sub)population conclusions can be drawn from the analyses (Bryson, Dorsett, & Purdon, 2002). The literature primarily focuses on the average treatment effect (ATE) and the average treatment effect of the treated (ATT). In our study the ATE would be the average effect of being in a higher track for both lower track students and higher track students. The ATT is the average effect of being in a higher track for higher track students (Stuart, 2010). In this study we focused on the ATT. The area of common support is determined

by the overlap in propensity scores (Steiner & Cook, 2014). We therefore assessed the overlap in propensity scores of both tracks for each comparison with density plots.

Matching

Using matching methods, we first explicitly modeled how students are allocated into different tracks, using the aforementioned confounders (Schafer & Kang, 2008). The goal was to find comparable students across two consecutive tracks for each combination of confounder values. When comparable students were found, a matched dataset of students across tracks with equal confounder distributions could be constructed. Hence, any effect of the track could be causally ascribed to that track (i.e. the ignorable treatment assumption; Rubin, 1978; Winship & Morgan, 2007). Exactly how track allocation was modeled and how the matched set was constructed depended on the specific matching procedure (Stuart, 2010). In this study, several variations on three main matching procedures were used to match students: propensity score matching (e.g. Caliendo & Kopeinig, 2008), Mahalanobis distance matching, and coarsened exact matching (e.g. Iacus et al., 2012). These all have the same goal: a dataset with equal confounder distributions across tracks. However, they differ in whether either bias reduction or efficiency has priority and to what population the inferences apply. We chose for applying different matching methods due to the lack of consensus on which matching method is optimal under which conditions (Stuart, 2010). An overview of the different matching methods can be found in Table 4. Matching was done with the MatchIt 3.0.1 package (Ho, Imai, King, & Stuart, 2011) in R 3.3.2. See Appendix 1 for an example analysis using MatchIt.

After the matching procedures, balance in the matched datasets was assessed through standardized mean differences of covariates (SMD's, SD of lower track as denominator) between tracks. We investigated the mean, minimum and maximum of all SMD's. Mean SMD's should be

no higher than 0.05 while SMD's of specific covariates as a rule of thumb should not exceed 0.25 (Caliendo & Kopeinig, 2008). To indicate how the matching methods differ in efficiencies, we report the effective sample size (ESS, Kish, 1965) of the lower track in a track comparison. We interpret the ESS of the lower track after matching as how many observations from a random sample would give a sampling variation estimate equal to our weighted lower track group (Ridgeway, McCaffrey, Morral, Griffin, & Burgette, 2017). We do note that due to the correlation between the propensity scores and the outcome(s), the ESS will always be an underestimation of efficiency.

Propensity score matching

Using propensity score matching (PSM), the probability of track allocation was modeled by estimating propensities of respondents of being allocated to the higher track. The logits of these propensities were then used to match respondents across tracks (Rosenbaum & Rubin, 1983). The theoretical foundation is that conditional on these propensities, the allocation of students was random (Imbens & Rubin, 2015). We used logit models to estimate propensities of higher track assignment, with higher track assignment as outcome and a selection of baseline covariates as predictors. Per pairwise comparison of tracks and per outcome, baseline covariates were included based on their predictive power being larger than a 0.05 correlation for performance at T3 in any of the track comparisons (Austin, 2011; Austin, Grootendorst, Normand, & Anderson, 2007; Myers et al., 2011). When matching for mathematics four baseline covariates were not used (Autonomous motivation, emotional engagement, interest in Dutch and interest in French) during matching. When matching for Dutch one baseline covariate was not used (Wellbeing) during matching. Next, matching procedures were applied in which students of both tracks were matched based on their logit propensity scores (e.g. Caliendo & Kopeinig,

2008). We applied two types of matching procedures: nearest neighbor caliper matching and full matching.

Using nearest neighbor caliper matching, a student in the higher track was matched to the student in the lower track who had the closest logit propensity value (Thoemmes & Kim, 2011). A higher tolerance of the maximum distance (i.e. the caliper) is more efficient, but also more biased. We used a 0.05SD propensity for matching. Using full matching, lower track students were matched to higher track students within the same logit propensity score interval (Stuart & Green, 2008). Weights were estimated per interval so that both tracks were equally represented per interval. We applied full matching to students with propensity scores between 0.05 and 0.95.

Mahalanobis distance matching

In Mahalanobis distance matching (MDM), the selection mechanism is controlled for by matching students who have the shortest Mahalanobis distance. This measure of distance is based on the covariance matrices estimated on the baseline covariates of both groups (Rosenbaum & Rubin, 1985). We used the baseline covariates from Table 1 as described for propensity score matching. Matching students across tracks on this distance metric approximates a stratified random sample. We used a specific variation of this method whereby only students within a 0.25 propensity score caliper were considered for MDM (see Rosenbaum & Rubin, 1985, p35). In this implementation, we used 1:1 nearest neighbor matching without replacement.

Coarsened exact matching

In coarsened exact matching (CEM), the selection mechanism is controlled for by matching students across tracks based on coarsened baseline covariates (Iacus et al., 2012). The idea is that exact matching on covariates is unnecessary as small differences typically are not meaningful. Accordingly, it suffices to match on coarsened covariates to reduce most bias. This

matching procedure approximates a stratified sample. Coarsening was based on two characteristics of the baseline covariates: their predictive power and reliability. For predictive power, the more a confounder predicts learning gains by the correlation coefficient, the more bins are created. The number of bins was calculated by how many $0.2SD$ differences in learning gains were predicted by the baseline covariate within a $6SD$ interval. For reliability, the standard error of the variables was used to determine which values are different, with 95% certainty. These standard errors were derived from a measurement model using either confirmatory factor analysis or item response theory. Hence, bins were constructed so that this threshold was not exceeded. Based on both the predictive power and reliability, the smallest number of bins resulting from both approaches was taken. The resulting bins with both students in the higher and lower track were then reweighted to have equal numbers in both tracks.

Outcome analyses: GEE and MLGC

Two methods from different research traditions were used to estimate the effects of being in a higher track (McNeish, Stapleton, & Silverman, 2017). Generalized estimating equations (GEE's) account for the design effect on the efficiency of parameter estimates through a correlation matrix for sampling units (Hardin, 2002). Multilevel latent growth curve models (MLGC's) account for cluster effects on the efficiencies of parameter estimates by partitioning the variance in between-cluster and within-cluster variance (Duncan, Duncan, & Strycker, 2013). Both methods were used in this study, for there is no consensus on which method is optimal under which conditions (e.g. McNeish, Stapleton, & Silverman, 2016). For both methods the difference in academic performance at each time point will be divided by the SD of the lower track in the unmatched sample at the start of secondary education. Differences in academic

performance between tracks at the end of each of the first three years of secondary education will be reported as d_{T1} , d_{T2} and d_{T3} . Interpretation of the effect sizes is by Cohen's d (Cohen, 1977).

GEE's account for clustering in linear regression models by specifying a working correlation matrix for the error terms of the sampling units (Robins, Hernan, & Brumback, 2000). We specified an autoregressive working correlation matrix to account for the repeated measurements of students (Liang & Zeger, 1986). Robust sandwich standard errors were reported, which are valid even if the working correlation matrix is misspecified (Joffe, Ten Have, Feldman, & Kimmel, 2004). Newton-Raphson was used for parameter estimation. For mathematics each respondent i the outcome y at time t is related to track allocation in the following model:

$$Y_{it} = \beta_0 + \beta_{1t}LowerTrack_i + \beta_{2t}HigherTrack_i + \beta_3X_i + e_{it}$$

Y_{it} represents the performance of person i at time t . Y_{it} is a function of: parameter β_0 , the average performance at $t = 0$ for the lower track students; parameter β_{1t} and indicator $LowerTrack_i$, describing the average change in performance compared to the intercept if person i is in the lower track at times $t = 1$ to $t = 3$; parameter β_{2t} and indicator $HigherTrack_i$ describing the average change in performance compared to the intercept if person i is in the higher track at times $t = 0$ to $t = 3$; β_3X_i , where β_3 is a vector of parameters describing the average change in performance for the standardized values of the covariates in vector X_i measured for person i ; and the error term e_{it} for person i at time t . Added to this baseline model are the inverse probability weights resulting from the different matching procedures (these weights equal one using 1:1 matching with caliper $0.05SD$ and MDM). The covariates X_i added to this model are those used in the propensity score estimation, - removing any remaining confounding (i.e. doubly robust; Schafer & Kang, 2008). For reading comprehension the same model was used, but parameters at

$t = 1$ and $t = 2$ were not estimated.

To examine the differences between the treatment conditions at each time point, contrasts between two conditions were tested using one degree of freedom Wald tests (Kuhn, Weston, Wing, & Forester, 2016). The GEE-models were estimated using the *geepack* 1.2-1 package (Højsgaard, Halekoh, & Yan, 2006) in R 3.3.2 while the contrasts were estimated using the *contrast* package 0.21 (Kuhn et al., 2016) in R 3.3.2. See Appendix 2 for an example analysis using *geepack*.

MLGC's account for clustering by a shared residual for each unit in the same school cluster, modeling the heterogeneity in the error terms (e.g. Goldstein, 2011; Hox, 2010; Snijders & Bosker, 2012). The measured performance at each time point was modeled as deriving from a latent growth curve (Curran & Hussong, 2002; Duncan et al., 2013). We specified a latent growth curve where the functional form was freely estimated (i.e. a latent basis model; Grimm, Ram, & Hamagami, 2011). Based on testing several functional forms in the total sample we found that freely estimating the functional form resulted in the best fit indices. We also found that the functional form changed across matched datasets, leading us to freely estimate the functional form in each matched dataset. Furthermore, we had no specific theory on a most plausible functional form. For parameter estimation we used maximum likelihood with bootstrapped standard errors (using the resampling method for clustered data, Asparouhov & Muthén, 2010). The latent growth curve model for mathematics can be summarized with the following equations:

$$Y_{it} = IC_i + SL_{it} + e_{it}$$

$$IC_i = \beta_1 LowerTrack_i + \beta_2 HigherTrack_i + \beta_3 X_i + u_{1i} LowerTrack_i + u_{2i} HigherTrack_i$$

$$SL_{it} = \lambda_{1t} LowerTrack_i + \lambda_{2t} HigherTrack_i + \beta_4 X_i$$

Y_{it} is the performance of person i at time t . Y_{it} is a function of: IC_i , the average performance at $t = 0$ for student i ; SL_{it} , describing the average change in performance relative to the intercept for student i at times $t=1$ to $t=3$; and e_{it} is the random effect for student i at time t . IC_i is a function of: parameter β_1 and indicator $LowerTrack_i$, describing the average performance at $t = 0$ for the lower track students; parameter β_2 and indicator $HigherTrack_i$, describing the average performance at $t = 0$ for the higher track students; $\beta_3 X_i$, where β_3 is a vector of parameters describing the average change in Y_{it} for the standardized values of the covariates in vector X_i measured for person i ; $u_{1i} LowerTrack_i$, describing the deviation for person i in the lower track group from the average IC_i in the lower track; $u_{2i} HigherTrack_i$, describing the deviation for person i in the higher track group from the average IC_i in the higher track. SL_{it} is a function of: parameter λ_{1t} and indicator $LowerTrack_i$, describing the average change in performance at $t = 1$ to $t = 3$ relative to the performance at $t = 0$ for students in the lower track; parameter λ_{2t} and indicator $HigherTrack_i$, describing the average change in performance at $t = 1$ to $t = 3$ relative to the performance at $t = 0$ for students in the higher track; $\beta_4 X_i$, where β_4 is a vector of parameters describing the average change in SL_{it} for the standardized values of the covariates in vector X_i measured for person i . See Appendix 3 for an example analysis using Mplus. For reading comprehension the same model was used, but parameters at $t = 1$ and $t = 2$ were not estimated, e_{it} was constrained to zero, and u_{3i} and u_{4i} were added as the deviations for person i from the average IC_i of a track group. The latent growth curve models for mathematics and reading comprehension are also shown graphically in Figure 1.

[Figure 1]

For mathematics the mean intercept and mean slope were freely estimated, while the intercept variance was also freely estimated. The slope variance was constrained to zero, for often during the analyses it would be either close to zero or negative after incorporating predictors into the model. The error terms were constrained to equality across time points, as is common in the specification of latent growth curve models (Singer & Willett, 2003). For reading comprehension both the mean intercept and mean slope were freely estimated, while their variances were freely estimated (for a latent growth curve model with two measurements points, see Duncan & Duncan, 2009, p.3). The error terms were constrained to zero.

To compare pairs of tracks, two different latent growth curves were estimated using a multigroup model with the two tracks. The differences in learning gains at T1 and T2 were assessed by the estimated factor loadings at T1 and T2 multiplied by the estimated slopes. The differences in learning gains between tracks at T3 were assessed comparing the estimated slopes. These differences were reported as d_{T1} , d_{T2} and d_{T3} , using Cohen's d (Cohen, 1977) for interpretation. The significance was tested by assessing whether the estimated difference between both tracks differed significantly from zero. In these models, covariates used during matching were incorporated as predictors of the intercept and slope for double robustness (Schafer & Kang, 2008). These models were specified in Mplus 7.4 (Muthén & Muthén, 2015).

Assessing differential track effects

Assessing the differences in track effects over time for mathematics was done by estimating differences in track effects in the MLGC-model and whether they differed significantly from zero. Comparisons were between d_{T1} and d_{T2} , between d_{T2} and d_{T3} and between d_{T1} and d_{T3} . These tests were conducted on the datasets resulting from 1:1 matching with caliper 0.05SD.

Assessing differential track effects across student characteristics was done by estimating track effects for different student groups according to a dichotomization of baseline performance, propensity score and SES. The means for each of these variables in the datasets resulting from 1:1 matching with caliper $0.05SD$ were used to create a dichotomous variable. Hence partitioning the two tracks of each comparison in four groups. Subsequently, track effects were estimated separately for low and high performance students, low and high SES students, and low and high propensity score students. Differences in track effects were estimated by estimating differences in track effects in the MLGC-model and whether they differed significantly from zero.

Missing data

In our sample, 3.43% of the data was missing on average at T0 (see Table 1). We used multiple imputation by chained equations to attain unbiased and efficient estimates for missing values (Schafer & Graham, 2002). Due to schools as clusters in our data, the multilevel pan-approach was used during imputation (Lüdtke, Robitzsch, & Grund, 2017). Missing values of the 25 baseline covariates, of which 20 had missing values, were imputed with the 25 baseline covariates and outcomes values at T1, T2 and T3 used as predictors (Leyrat et al., 2017; White, Royston, & Wood, 2011). Convergence was reached after 15 iterations and was determined by the autocorrelation functions and trace plots. Recent literature suggests as many imputed datasets as the average missing rate multiplied by ten (Bodner, 2008; White et al., 2011). However we played safe by estimating ten imputed datasets, while combining their results as described by Rubin's (1987) rules. The relative efficiencies attained (against a perfect efficiency of 100%) for the outcomes of interest ranged from 91.04% to 99.79% with an average of 97.36%. Hence, the results were unlikely to notably differ in precision from the perfect efficiency case. The

imputations were estimated using the packages mice 2.30 (van Buuren & Groothuis-Oudshoorn, 2011) and pan 1.4 (Zhao & Schafer, 2016) in R 3.3.2.

Regarding the outcomes of interest, some students were censored due to missingness. For mathematics, 10.55% was censored at T1, 4.49% at T2 and 5.93% at T3. For reading comprehension 9.83% is censored at T3. To obtain unbiased and efficient estimates, two approaches were used to handle censoring, tailored to the two outcome analysis methods. For GEE's, time-varying inverse probability censoring weights were estimated per time point, with each uncensored student receiving a weight which accounts for comparable censored students (Robins et al., 2000). The censoring weights are estimated as a function of the baseline covariates. Due to time-varying weights not being useable in MLGC's, full information maximum likelihood (FIML; Enders & Bandalos, 2001) was incorporated into the estimation of the parameters, yielding efficient and unbiased estimates. FIML is also generally considered a superior approach. The censoring weights were estimated using the ipw 1.0-11 package (van der Wal & Geskus, 2011) in R 3.3.2. FIML is a part of the estimation procedure of latent growth curves in Mplus 7.4 by using maximum likelihood (Raykov, 2005).

Results

Track differences before matching

Table 2 shows the means for each baseline covariate per track, with the covariates, while the last three columns show the SMD's between each of the track comparisons. Overall, the hierarchy in tracks from classical, modern, technical to vocational is reflected in the differences in academic performance, socio-economic status and academic self-concept. Moreover, there is a general trend of the lower tracks attracting more students who speak no Dutch at home. Interestingly, there is a substantially larger difference in mathematics performance between the

technical and vocational track, relative to the difference between the other two track comparisons.

[Table 2]

Table 3 describes the LGCM's across tracks for mathematics and reading comprehension, describing per track the mean baseline performance at T0, mean performance at T1, mean performance at T2, mean performance at T3 and amount of growth between T0 and T3. The fit indices for this model were satisfactory for mathematics (CFI = 0.984, TLI = 0.984, RMSEA = 0.063). The fit indices for reading comprehension indicate no misfit, due to being a completely saturated model. Constraining the slope to equality across all tracks yielded no significantly worse fit for mathematics ($\chi^2(3, N = 3205) = 1.202, p = 0.753$), indicating comparable mean learning gains across tracks between T0 and T3. Constraining the slope to equality across all tracks for did yield significantly worse fit for reading comprehension ($\chi^2(3, N = 3205) = 198.526, p < 0.01$), indicating significant differences in mean learning gains across tracks between T0 and T3. The vocational track even showed negative growth for reading comprehension.

[Table 3]

In Figure 2 three pairs of density plots are shown, one pair for each track comparison. The x-axis shows the logit propensities of going to the higher track predicted by the mathematics propensity score model. The area of common support can be described as the percentage of higher track students for who there is at least one lower track student with an equal or lower

propensity score. For 94.07% of classical track students there was at least one comparable modern track student, while for 94.09% of modern track students there was at least one comparable technical track student and for 91.05% of technical track students there was at least one comparable vocational track student. The logit propensities predicted by the reading comprehension propensity score model were highly comparable. Table 5, 6 and 7 show how many students were matched per matching procedure per track comparison, with almost all having enough matches to conduct outcome analyses. Only CEM for mathematics for the technical and vocational track comparison did not yield enough matches, thus no analyses on this matched dataset were conducted.

[Figure 2]

In short, the baseline covariates and assessment of overlap between tracks showed that differences in student selection exist across tracks. However, they also showed a substantial area of common support between tracks, a required condition for any matching procedure. Although the overlap between the technical and vocational track was smaller, matched comparisons were almost always possible.

Produced samples after matching

Critical to the samples produced by the matching procedure is balance, which we assessed with SMD's and differences in propensity scores. Table 4 shows the mean, minimum and maximum of all SMD's, as well as the mean difference in propensity scores for each matching procedure and track comparison.

[Table 4]

Across all matching procedures and comparisons, the mean SMD's were under the 0.05 threshold. However, it was slightly exceeded for CEM and MDM in the modern and technical track comparison. Using caliper matching and full matching, the mean propensity score difference between tracks is always close to zero for each criterion used. Thus, in general satisfactory balance is achieved between tracks per comparison when assessing mean SMD's and propensity scores.

When assessing SMD's for specific covariates across the three track comparisons and matching procedures, the classical and modern track comparison have SMD's under the 0.25 threshold, except when using CEM. For the modern and technical track comparison, all SMD's are under 0.25. For the technical and vocational track comparison, the SMD's point to difficulties with the matching procedures. For only matching with caliper $0.05SD$ has all SMD's under the 0.25 threshold. This is likely due to strong differences between these tracks impeding the success of the matching procedure (Steiner & Cook, 2013). Hence, caution is needed when making inferences for the technical and vocational track comparison.

Across the different matching procedures, the SMD's of the baseline measurement of mathematics for the classical and modern track comparison range from -0.05 to 0.07, with an average of 0.03. For the modern and technical track comparison it ranges from -0.01 to 0.16, with an average SMD of 0.06. For the technical and vocational track comparison it ranges from 0.11 to 0.37, with an average SMD of 0.22. For the baseline measurement of reading comprehension for the classical and modern track comparison it ranges from -0.01 to 0.13, with an average SMD of 0.05. For the modern and technical track comparison it ranges from 0.04 to

0.20, with an average SMD of 0.10. For the technical and vocational track comparison it ranges from -0.04 to 0.11, with an average SMD of 0.12.

Assessing the differences in ESS's indicates that 1:1 matching with caliper 0.05SD and MDM attain comparable datasets when it comes to efficiency. Full matching and CEM show smaller ESS's than 1:1 matching with caliper 0.05SD and MDM. Full matching in the technical and vocational track gives an ESS which is especially lower. Hence, indicating lower efficiencies for full matching and CEM compared to 1:1 matching with caliper 0.05SD and MDM in this study. Across the different matching procedures, the SMD's of the baseline measurement of mathematics for the classical and modern track comparison range from -0.05 to 0.07, with an average of 0.03. For the modern and technical track comparison it ranges from -0.01 to 0.16, with an average SMD of 0.06. For the technical and vocational track comparison it ranges from 0.11 to 0.37, with an average SMD of 0.22. For the baseline measurement of reading comprehension for the classical and modern track comparison it ranges from -0.01 to 0.13, with an average SMD of 0.05. For the modern and technical track comparison it ranges from 0.04 to 0.20, with an average SMD of 0.10. For the technical and vocational track comparison it ranges from -0.04 to 0.11, with an average SMD of 0.12.

Analysis of track effects

The treatment effects of the three pairwise comparisons of a higher track versus a lower track are presented in the following sections. For each comparison, the difference in mean value between both tracks at T1 (only mathematics), T2 (only mathematics) and T3 are estimated using both GEE's and MLGC's. The results of the pairwise comparisons of the classical and modern track, the modern and technical track, and the technical and vocational track are shown in Table 5, 6 and 7 respectively. Figure 3 shows the growth curves of these comparisons for matching

with caliper $0.05SD$. In the following paragraphs, we discuss the general trends in each pairwise comparison.

For the classical and modern track comparison, all positive effects of being in a higher track on mathematics range from $d = 0.19$ to $d = 0.29$ at T1, from $d = 0.12$ to $d = 0.20$ at T2 and from $d = 0.12$ to $d = 0.22$ at T3 using GEE's, with MLGC's yielding similar results. However, significance is not always reached, even for comparable effect sizes. For reading comprehension, the effects range from $d = 0.21$ to $d = 0.23$ using GEE's, with MLGC's yielding similar results. Again, significance is not always reached, even for comparable effect sizes.

[Table 5]

For the modern and technical track comparison, all effects on mathematics range from $d = 0.19$ to $d = 0.25$ at T1, from $d = 0.44$ to $d = 0.52$ at T2 and from $d = 0.17$ to $d = 0.27$ at T3 using GEE's, with MLGC's yielding similar results, though in some cases slightly lower. However, significance is not always reached, even for comparable effect sizes. For reading comprehension, the effects range from $d = 0.43$ to $d = 0.55$ using GEE's, with MLGC's yielding similar results. Significance is always reached.

[Table 6]

For the technical and vocational track comparison, all effects on mathematics range from $d = 0.65$ to $d = 0.69$ at T1, from $d = 0.38$ to $d = 0.44$ at T2 and from $d = 0.74$ to $d = 0.85$ at T3 using GEE's. MLGC's yield similar results in general, with some unsystematic differences.

Significance is mostly reached, but not always. For reading comprehension, the effects range from $d = 0.40$ to $d = 0.53$ using GEE's, with MLGC's yielding similar results. Significance is always reached.

[Table 7]

Sensitivity analyses of track effects

Sensitivity analyses of possible departures from the ignorable treatment assumption were conducted on the estimated track effects. We used Vanderweele & Arah's (2011) procedure. This meant assessing how strongly an unobserved confounder needs to differ between tracks to completely explain the observed track effect. This was investigated for a hypothetical unobserved confounder which has a relationship of small effect size ($r = 0.2$), a moderate effect size ($r = 0.4$) or a large effect size ($r = 0.6$). These sensitivity analyses were performed for each track effect. For brevity only those for matching with caliper $0.05SD$ and MLGC's as outcome analyses are reported. Concerning the classical and modern track comparison for mathematics, an unobserved confounder with a moderate (small/large) relation to mathematics at T3 needs to differ between both tracks with a SD of 0.5 ($1.0/0.3$). For reading comprehension, an unobserved confounder with a moderate (small/large) relation to reading comprehension at T3 needs to differ between both tracks with a SD of 0.6 ($1.2/0.4$). Concerning the modern and technical track comparison for mathematics, an unobserved confounder with a moderate (small/large) relation to mathematics at T3 needs to differ between both tracks with a SD of 0.6 ($1.2/0.4$). For reading comprehension, an unobserved confounder with a moderate (small/large) relation to reading comprehension at T3 needs to differ between both tracks with a SD of 1.2 ($2.3/0.8$). Concerning the technical and vocational track comparison for mathematics, an unobserved confounder with a

moderate (small/large) relation to mathematics at T3 needs to differ between both tracks with a *SD* of 1.9 (3.8/1.3). For reading comprehension, an unobserved confounder with a moderate (small/large) relation to reading comprehension at T3 needs to differ between both tracks with a *SD* of 1.3 (2.5/0.8).

Furthermore, Steiner et al. (2010) suggested that variation in the geographical location of the treatment and control groups may lead to confounded treatment effects. Hence, sensitivity analyses were conducted whether the estimated track effects change when either only matching students in schools in urban areas or only matching students across tracks in the same school. These sensitivity tests were conducted using caliper matching with $0.05SD$ and MLGC's as outcome analyses. These sensitivity tests could not be conducted for the technical and vocational track comparison, due to too few matched students. For brevity, we only report the results at T3. Concerning the classical and modern track comparison the effect is $d_{T3} = 0.28$ ($SE = 0.23$) for mathematics and $d_{T3} = 0.11$ ($SE = 0.10$) for reading comprehension when only matching students in urban schools. When only matching students across tracks in the same schools $d_{T3} = 0.23$ ($SE = 0.10$) for mathematics and $d_{T3} = 0.31$ ($SE = 0.12$) for reading comprehension. Concerning the modern and technical track comparison the effect is $d_{T3} = 0.19$ ($SE = 0.21$) for mathematics and $d_{T3} = 0.43$ ($SE = 0.11$) for reading comprehension when only matching students in urban schools. When only matching students across tracks in the same schools $d_{T3} = 0.42$ ($SE = 0.16$) for mathematics and $d_{T3} = 0.71$ ($SE = 0.12$) for reading comprehension. These results show that only matching students in urban areas yields effect sizes rather comparable to the effect sizes in the whole sample. The results for students in different tracks in the same school does show effect sizes larger than in the whole sample for the modern and technical track comparison, significantly so for reading comprehension. However, this runs opposite to the hypothesis that

the effects of our original analysis may still have been confounded when not controlling for environment. Hence, our conclusion is that there is no indication our main analyses were confounded by differences in environments.

Differences in track effects over time

For the classical and modern track comparison no significant differences in track effects over time were found. However, the modern and technical track comparison does show that d_{T2} is larger than d_{T1} ($F(1,430) = 4.12, p < .05$) and that d_{T2} is larger than d_{T3} ($F(1,430) = 2.05, p < .05$). However, there is no significant difference between d_{T1} and d_{T3} ($F(1,430) = 0.93, p < .05$). For the technical and vocational track comparison, d_{T1} is significantly larger than d_{T2} ($F(1,430) = 2.08, p < .05$) and d_{T3} is significantly larger than d_{T2} ($F(1,430) = 2.31, p < .05$), but there is no significant difference between d_{T1} and d_{T3} . However, when applying different matching procedures, some of these effects are insignificant, even when the effect size is comparable.

[Figure 3]

Differences in track effects over student groups

Table 8 shows the track effects for each track comparison per student group, with the differences in track effects between groups shown as well. No difference in track effects between groups is found to be significant.

[Table 8]

However, assessing the standard errors in Table 8 shows that statistical power is only adequate to distinguish medium to large effects per time point. To increase statistical power, we tested the null hypothesis whether the difference in track effects for mathematics across all three time points are significantly different from zero. Comparing low versus high performing students only for the modern and technical track comparison a significant effect was found ($F(1,830) = 2.04, p < .05$) with mean ES = -0.25. This shows that the effect of the modern track versus the technical track is stronger for low academic performance students than high academic performance students. However, for all other comparisons no significant effect was found.

Discussion

This study investigated whether tracks affect academic performance for mathematics and reading comprehension during the first three years of Flemish secondary education. Academic performance was compared between four tracks: the classical, the modern, the technical and the vocational track. First, students were matched across every pair of tracks which are hierarchically consecutive. Second, the academic performance for comparable students across pairs of tracks was assessed. Supporting our hypothesis, positive effects of going to a higher track for higher track students were found without exception.

However, a more nuanced picture is revealed when assessing effects sizes, how they differ between the different track comparisons and over time. At the end of the third year we found two small track effects ($d = 0.20$ and $d = 0.23$) and one large effect ($d = 0.76$) for mathematics. For reading comprehension there was one small track effect ($d = 0.23$) and two moderate effects ($d = 0.46$ and $d = 0.50$). Our results showed that these track effects do not differ according to student characteristics. However, our study did reveal that for two out of three track comparisons there were significant differences in the track effects over time. Between the end of

the first year and the end of the third year though, there was no significant difference in track effect. This indicates that the benefit of being in a higher track for mathematics does not expand over time (Figure 3 illustrates this). Relatedly, when assessing the difference in relative learning gain (the learning gain in the lower track divided by the learning gain in the higher track), the advantage of being in the higher track diminishes over time. For example, in the classical and modern track comparison, the relative learning gain of the lower track is only 16%, while at the end of the third year this rises to 81%. A comparable trend was found for the other track comparisons (57% to 82% and -3% to 43%). In sum, the effect size of being allocated to a higher track differs over time, performance indicators and track comparisons, indicating that there is no single track effect within an education system. Moreover, a comparison of the relative learning gains reveals that the track effect on academic performance may be limited over time in a relative sense.

The heterogeneity in our effect sizes raises the question how these compare to the effect sizes of other quasi-experimental studies on track effects. Becker et al. (2012) and Guill et al. (2016) respectively found average effect sizes of 0.40 and 0.31 when comparing an academic to a non-academic track across four years for intelligence development. Our own average effect size at the end of the third year of 0.39 across all track comparisons, matching methods and outcome analyses is very alike. Somewhat small compared to our own average effect size, Retelsdorf et al. (2012) found an effect size of 0.21 for decoding speed and a non-significant effect for reading comprehension which they deemed too small to even mention. However, given the heterogeneity in our own results, such heterogeneity across studies on different tracking systems does not seem strange. In sum, the effect sizes in our study generally resemble the findings of prior quasi-experimental studies on track effects.

Given the results of our and former quasi-experimental studies, the argument that track effects cause selective systems to have higher inequality than comprehensive systems is given more weight. For the explanation is that tracks exacerbate already existing differences in academic performance due to the higher tracks benefitting student academic performance (e.g. Brunello & Checchi, 2007). Based on our study, this argument is certainly plausible for Flemish education.

Limitations and strengths

The main goal of matching is to reduce selection bias for causal inferences, its success the ignorable treatment assumption (Rosenbaum & Rubin, 1983). If after a matching procedure a confounder remains that predicts both track assignment and the outcome, the track effect is biased through that confounders' effect. Thus, a track effect can only be attributed to a track if all relevant selection bias due to confounders is removed (Steiner & Cook, 2014). We applied different matching methods to see whether this mattered for bias reduction. All methods yielded comparable effect sizes, indicating that the choice for a specific matching method to reduce selection bias is somewhat trivial in this study. Furthermore, we investigated the tenability of the ignorable treatment assumption through sensitivity tests of the treatment effect (Caliendo & Kopeinig, 2008; Vanderweele & Arah, 2011) across track comparisons. Generally, we found that small effects could still be somewhat plausibly explained by an unobserved confounder. However, track effects of moderate size or larger could not be plausibly explained by an unobserved confounder in our view. If considering the track effect at the end of the third year, this means three out of six track effects are robust for unobserved confounding. Furthermore, potential confounding due to variation in the geographical location of the treatment and control

groups (Steiner et al., 2010) was also assessed. No indication was found that our main analyses were confounded.

Potential unmeasured confounders remain however. For example, we have no information on the parents' educational aspirations, a well-known predictor of track choice (e.g. Boone & Van Houtte, 2013) and academic performance (e.g. Spera, Wentzel, & Matto, 2009). Another example would be that we had no information on student's intelligence, also a potential confounder. Whether these unmeasured confounders bias our result to a tangible degree depends on how well the measured confounders in our matching procedure can also account for the biasing effect of the unmeasured confounders.

Any estimate deriving from a matched dataset is also limited in inference to the area of common support for which enough statistical power exists in both groups (Stuart, 2010). Initially, for 91% to 94% of students in the higher track a comparable student was found in the lower track across the three pairwise comparisons. Although not allowing for estimating the average treatment effect for the treated, having more than 90% of the higher track represented in a matched dataset seems intuitively close. However, while the different matching methods attain comparable effect sizes, the standard errors of these effects differ. Generally, those matching methods that apply greater weights, in order to resemble the higher track population, do this at a cost of efficiency (i.e. full matching). Hence, we are cautious in making inferences on the effect of being in a higher track for the entire higher track population (the ATT). We think our inferences can only apply to a population of students who are adequately represented in each track. This limitation in the area of common support is not unique to his study, for comparable studies on track effects show even smaller overlap in propensity scores (e.g. Becker et al., 2012; Guill et al., 2016; Retelsdorf et al., 2012).

Analyzing the differences in outcomes due to track effects also required a model to describe learning gains. Discussion remains under which conditions multilevel models or GEE's are preferable (e.g. McNeish, Stapleton, & Silverman, 2016). Furthermore, discussion within the matching literature on the estimation of standard errors is also ongoing (Abadie & Imbens, 2008). Due to a lack of consensus in the literature, we choose to apply both MLGC's with bootstrapped standard errors and autoregressive GEE models. Interestingly, both methods yielded close results for effect sizes, with MLGC's having substantially larger standard errors. Even though we cannot answer which method gives the correct assessment of standard errors, we consider their accordance in effect sizes a strength of this study. Hence, applying either solely MLGC's or solely GEE's would have led to the same conclusions in our study.

Furthermore, it should be noted that though we found significant track effects on academic performance, what causes tracks to differ remains unknown. As mentioned, track effects are likely because tracks differ in pedagogical design and peer-group composition (e.g.), but how these effects interact and contribute to track effects largely remains unknown. Moreover, track effects should perhaps not always be sought within the school context. For example, being in a higher track might pressure a student to read more in his/her spare time.

As in any study we were limited by our data, with only two academic performance indicators available across a three-year span. With some irony this limitation seems more important given our own results, which show that describing the functional form of longitudinal learning gains can lead to a more nuanced interpretation. However, our own data do not allow for assessing how the track effect changes after the first three years of secondary education. Neither could we assess the functional form for development in reading comprehension. Perhaps more limiting is that our performance indicators only offer a narrow view on student's academic

performance. Indeed, the main rationale behind tracking is that different tracks offer skill development in different areas (e.g. Bol & van de Werfhorst, 2013). Correspondingly, it seems necessary to assess this specific claim. If different tracks are meant to prepare for different demands of the labor market, it should be evaluated whether different tracks do adequately prepare students for these different demands.

Conclusion

Assessing whether tracks affect academic performance for mathematics and reading comprehension during the first three years of Flemish secondary education revealed that being allocated to a higher track is beneficial for academic performance. A comparison of these effects reveals that the effect size differs over time, performance indicators and track comparisons, but not according to student characteristics. Interestingly though, the gap caused by the track effect does not enlarge over time. These results are in line with former research on track effects, giving further credence to the argument that tracks are responsible for exacerbating differences in academic performance.

References

- Abadie, A., & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6), 1537–1557.
- Ammermüller, A. (2005). Educational Opportunities and the Role of Institutions. *ZEW Discussion Paper*, 5–44.
- Asparouhov, T., & Muthén, B. (2010). Resampling methods in Mplus for complex survey data. *Structural Equation Modeling*, 14(4), 535–569.
- Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, 27(12), 2037–2049.
- Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, 10(2), 150–161.
- Austin, P. C., Grootendorst, P., Normand, S.-L. T., & Anderson, G. M. (2007). Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Statistics in Medicine*, 26(4), 754–768.
- 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.
- Barth, J. M., Dunlap, S. T., Dane, H., Lochman, J. E., & Wells, K. C. (2004). Classroom environment influences on aggression, peer relations, and academic focus. *Journal of School Psychology*, 42(2), 115–133.
- Bauer, P., & Riphahn, R. T. (2006). Timing of school tracking as a determinant of intergenerational transmission of education. *Economics Letters*, 91(1), 90–97.

- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., ... Tsai, Y.-M. (2010). Teachers' Mathematical Knowledge, Cognitive Activation in the Classroom, and Student Progress. *American Educational Research Journal*, 47(1), 133–180.
- Becker, M. (2009). *Kognitive Leistungsentwicklung in differenziellen Lernumwelten: Effekte des gegliederten Sekundarschulsystems in Deutschland [Cognitive development in differential learning environments: Effects of the tracked secondary school system in Germany]*. Berlin, Germany: Max-Planck-Institut für Bildungsforschung.
- Becker, M., Lüdtke, O., Trautwein, U., Köller, O., & Baumert, J. (2012). The differential effects of school tracking on psychometric intelligence: Do academic-track schools make students smarter? *Journal of Educational Psychology*, 104(3), 682–699.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29–51.
- Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling*, 15(4), 651–675.
- Bol, T., & van de Werfhorst, H. G. (2013). Educational Systems and the Trade-Off between Labor Market Allocation and Equality of Educational Opportunity. *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), 1545–1572.
- Boone, S., & Van Houtte, M. (2013). Why are teacher recommendations at the transition from primary to secondary education socially biased? A mixed-methods research. *British Journal of Sociology of Education*, 34(1), 20–38.

- Boudon, R. (1974). *Education, opportunity, and social inequality: Changing prospects in western society*. New York City, New York: Wiley.
- Breen, R., & Goldthorpe, J. H. (1997). Explaining Educational Differentials Towards a Formal Rational Action Theory. *Rationality and Society*, 9(3), 275–305.
- Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12), 1149–1156.
- Brunello, G., & Checchi, D. (2007). Does school tracking affect equality of opportunity? New international evidence. *Economic Policy*, 22(52), 781–861.
- Bryson, A., Dorsett, R., & Purdon, S. (2002). *The use of propensity score matching in the evaluation of active labour market policies* (Working Paper No. 4). London, United Kingdom: Department for Work and Pensions.
- Buser, T., Niederle, M., & Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *Quarterly Journal of Economics*, 129(3), 1409–1447.
- Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1), 31–72.
- Cohen, J. (1977). *Statistical power analysis for the behavioural sciences*. New York City, New York: Academic press.
- Curran, P. J., & Hussong, A. M. (2002). Structural equation modeling of repeated measures data: Latent curve analysis. In D. S. Moskowitz & S. L. Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data* (pp. 59–85). Mahwah, New Jersey: Lawrence Erlbaum.
- Denissen, J. J. A., Zarrett, N. R., & Eccles, J. S. (2007). I like to do it, I'm able, and I know I am:

- Longitudinal couplings between domain-specific achievement, self-concept, and interest. *Child Development*, 78(2), 430–447.
- Ditton, H., Krüsken, J., & Schauenberg, M. (2005). Bildungsungleichheit — der Beitrag von Familie und Schule. *Zeitschrift Für Erziehungswissenschaft*, 8(2), 285–304.
- Duncan, T. E., & Duncan, S. C. (2009). The ABC's of LGM: an introductory guide to latent variable growth curve modeling. *Social and Personality Psychology Compass*, 3(6), 979–991.
- Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2013). *An introduction to latent variable growth curve modeling: Concepts, issues, and application*. Mahwah, New Jersey: Erlbaum.
- Dupriez, V., Dumay, X., & Vause, A. (2008). How do school systems manage pupils' heterogeneity? *Comparative Education Review*, 52(2), 245–273.
- Duru-Bellat, M., & Suchaut, B. (2005). Organisation and Context, Efficiency and Equity of Educational Systems: what PISA tells us. *European Educational Research Journal*, 4(3), 181–194.
- Dweck, C. S., Chiu, C., & Hong, Y. (1995). Implicit theories and their role in judgments and reactions: A word from two perspectives. *Psychological Inquiry*, 6(4), 267–285.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, New Jersey: Lawrence Erlbaum.
- Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, 8(3), 430–457.
- Erikson, R., Goldthorpe, J. H., Jackson, M., Yaish, M., & Cox, D. R. (2005). On class differentials in educational attainment. *Proceedings of the National Academy of Sciences of*

the United States of America, 102(27), 9730–9733.

- Gamoran, A. (1992). Synthesis of research: Is ability grouping equitable? *Educational Leadership*, 50(2), 1–18.
- Goldstein, H. (2011). *Multilevel statistical models*. Hoboken, New Jersey: Wiley.
- Golinelli, D., Ridgeway, G., Rhoades, H., Tucker, J., & Wenzel, S. (2012). Bias and variance trade-offs when combining propensity score weighting and regression: with an application to HIV status and homeless men. *Health Services and Outcomes Research Methodology*, 12(2), 104–118.
- Grimm, K. J., Ram, N., & Hamagami, F. (2011). Nonlinear growth curves in developmental research. *Child Development*, 82(5), 1357–1371.
- Guill, K., Lüdtke, O., & Köller, O. (2017). Academic tracking is related to gains in students' intelligence over four years: Evidence from a propensity score matching study. *Learning and Instruction*, 47, 43–52.
- Gustafsson, J. E. (2008). Schooling and intelligence: Effects of track of study on level and profile of cognitive abilities. In P. C. Kyllonen, R. D. Roberts, & L. Stankov (Eds.), *Extending intelligence: Enhancement and new constructs* (pp. 31–49). London, United Kingdom: Routledge.
- Hahn, J., Todd, P., & der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1), 201–209.
- Hall, C. (2012). The effects of reducing tracking in upper secondary school evidence from a large-scale pilot scheme. *Journal of Human Resources*, 47(1), 237–269.
- Hallam, S., & Ireson, J. (2003). Secondary school teachers' attitudes towards and beliefs about ability grouping. *British Journal of Educational Psychology*, 73(3), 343–356.

- Hanushek, E. A., & Wößmann, L. (2006). Does educational tracking affect performance and inequality? differences-in-differences evidence across countries. *The Economic Journal*, 116(510), C63–C76.
- Hanushek, E. A., & Wößmann, L. (2010). Sample Selectivity and the Validity of International Student Achievement Tests in Economic Research. *National Bureau of Economic Research Working Paper*.
- Hardin, J. W., & Hilbe, J. (2003). *Generalized estimating equations*. Boca Raton, Florida: Chapman and Hall/CRC.
- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt : Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, 42(8), 1–28.
- Højsgaard, S., Halekoh, U., & Yan, J. (2006). The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software*, 15(2), 1–11.
- Horn, D. (2009). Age of selection counts: a cross-country analysis of educational institutions. *Educational Research and Evaluation*, 15(4), 343–366.
- Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. New York City, New York: Routledge.
- Huang, J., van den Brink, H. M., & Groot, W. (2009). A meta-analysis of the effect of education on social capital. *Economics of Education Review*, 28(4), 454–464.
- Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1), 1–24.
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge, United Kingdom: University Press.
- Ireson, J., Hallam, S., & Plewis, I. (2001). Ability grouping in secondary schools: Effects on

- pupils' self-concepts. *British Journal of Educational Psychology*, 71(2), 315–326.
- Jackson, M., Erikson, R., Goldthorpe, H., & Yaish, M. (2007). Primary and secondary effects in class differentials in educational attainment: the transition to A-level courses in England and Wales. *Acta Sociologica*, 50(3), 211–229.
- Jakubowski, M., Patrinos, H. A., Porta, E. E., & Wiśniewski, J. (2016). The effects of delaying tracking in secondary school: evidence from the 1999 education reform in Poland. *Education Economics*, 24(6), 557–572.
- Jenkins, S. P., Micklewright, J., & Schnepf, S. V. (2008). Social segregation in secondary schools: how does England compare with other countries? *Oxford Review of Education*, 34(1), 21–37.
- Joffe, M. M., Ten Have, T. R., Feldman, H. I., & Kimmell, S. E. (2004). Model selection, confounder control, and marginal structural models: review and new applications. *The American Statistician*, 58(4), 272–279.
- Kerckhoff, A. C. (2001). Education and Social Stratification Processes in Comparative Perspective. *Sociology of Education*, 74(Extra Issue), 3–18.
- Kerr, S. P., Pekkarinen, T., & Uusitalo, R. (2013). School Tracking and Development of Cognitive Skills. *Journal of Labor Economics*, 31(3), 577–602.
- Kish, L. (1965). *Survey sampling*. New York City, New York: Wiley.
- Kloosterman, R., Ruiter, S., De Graaf, P. M., & Kraaykamp, G. (2009). Parental education, children's performance and the transition to higher secondary education: trends in primary and secondary effects over five Dutch school cohorts (1965–99). *The British Journal of Sociology*, 60(2), 377–398.
- Korthals, R. A., & Dronkers, J. (2016). Selection on performance and tracking. *Applied*

- Economics*, 48(30), 2836–2851.
- Kuhn, M., Weston, S., Wing, J., & Forester, J. (2016). The contrast Package. *CRAN Package Repository*, 1–14.
- Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, 9(3), 231–251.
- Kuzmina, J., & Carnoy, M. (2016). The effectiveness of vocational versus general secondary education. *International Journal of Manpower*, 37(1), 2–24.
- Lavrijzen, J., & Nicaise, I. (2015). New empirical evidence on the effect of educational tracking on social inequalities in reading achievement. *European Educational Research Journal*, 14(3), 206–221.
- LeTendre, G. K., Hofer, B. K., & Shimizu, H. (2003). What Is Tracking? Cultural Expectations in the United States, Germany, and Japan. *American Educational Research Journal*, 40(1), 43–89.
- Leyrat, C., Seaman, S. R., White, I. R., Douglas, I., Smeeth, L., Kim, J., ... Williamson, E. J. (2017). Propensity score analysis with partially observed covariates: How should multiple imputation be used? *Statistical Methods in Medical Research*, Advance online publication.
- Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22.
- LiSO-project. (2018). Het LiSO-project kort samengevat [The LiSO-project summarized]. Retrieved March 30, 2018, from <https://lisoproject.be/>
- Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple Imputation of Missing Data in Multilevel Designs: A Comparison of Different Strategies. *Psychological Methods*, 22(1), 141–165.

- Malamud, O., & Pop-Eleches, C. (2011). School tracking and access to higher education among disadvantaged groups. *Journal of Public Economics*, 95(11–12), 1538–1549.
- Marks, G. N. (2005). Cross-national differences and accounting for social class inequalities in education. *International Sociology*, 20(4), 483–505.
- Marsh, H. W., Ellis, L. A., Parada, R. H., Richards, G., & Heubeck, B. G. (2005). A Short Version of the Self Description Questionnaire II: Operationalizing Criteria for Short-Form Evaluation With New Applications of Confirmatory Factor Analyses. *Psychological Assessment*, 17(1), 81–102.
- Marsh, H. W., & O'Neill, R. (1984). Self Description Questionnaire III: The Construct Validity of Multidimensional Self-Concept Ratings by Late Adolescents. *Journal of Educational Measurement*, 21(2), 153–174.
- McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, 22(1), 114–140.
- Mons, N. (2007). *Les nouvelles politiques éducatives: La France fait-elle les bons choix? [New educational policies: Is France making the right choices?]*. Paris, France: Presses Universitaires de France.
- Muthén, B., & Muthén, L. (2015). *Mplus Statistical Analysis With Latent Variables User's Guide*. Los Angeles, California: Muthén & Muthén.
- Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., ... Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology*, 174(11), 1213–1222.
- OECD. (2012). *Equity and Quality in Education*. Paris, France: OECD.
- OECD. (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris,

France: OECD.

Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics*, 6(2), 1–59.

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.

Raykov, T. (2005). Analysis of longitudinal studies with missing data using covariance structure modeling with full-information maximum likelihood. *Structural Equation Modeling*, 12(3), 493–505.

Retelsdorf, J., Becker, M., Köller, O., & Möller, J. (2012). Reading development in a tracked school system: A longitudinal study over 3 years using propensity score matching. *British Journal of Educational Psychology*, 82(4), 647–671.

Retelsdorf, J., Butler, R., Streblow, L., & Schiefele, U. (2010). Teachers' goal orientations for teaching: Associations with instructional practices, interest in teaching, and burnout. *Learning and Instruction*, 20(1), 30–46.

Retelsdorf, J., & Möller, J. (2008). Entwicklungen von Lesekompetenz und Lesemotivation Schereneffekte in der Sekundarstufe? *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie*, 40(4), 179–188.

Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., & Burgette, L. (2017). twang: Toolkit for Weighting and Analysis of Nonequivalent Groups. *CRAN Package Repository*, 1–32.

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5), 550–560.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.

- Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1), 33–38.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1), 34–58.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York City, New York: Wiley.
- Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: examining reasons for acting in two domains. *Journal of Personality and Social Psychology*, 57(5), 749–761.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2), 147–177.
- Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods*, 13(4), 279–313.
- Schütz, G., Ursprung, H. W., & Wößmann, L. (2008). Education Policy and Equality of Opportunity. *Kyklos*, 61(2), 279–308.
- Shavit, Y., & Müller, W. (2000). Vocational secondary education, tracking, and social stratification. In M. Hallinan (Ed.), *Handbook of the sociology of education* (pp. 437–452). New York City, New York: Plenum.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford, United Kingdom: University press.
- Skinner, E. A., Kindermann, T. A., & Furrer, C. J. (2009). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral

- and emotional participation in academic activities in the classroom. *Educational and Psychological Measurement*, 69(3), 493–525.
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. London, United Kingdom: Sage.
- Spera, C., Wentzel, K. R., & Matto, H. C. (2009). Parental aspirations for their children's educational attainment: Relations to ethnicity, parental education, children's academic performance, and parental perceptions of school climate. *Journal of Youth and Adolescence*, 38(8), 1140–1152.
- Steiner, P. M., & Cook, T. D. (2014). Matching and propensity scores. In T. D. Litle (Ed.), *The oxford handbook of quantitative methods* (pp. 237–259). Oxford, United Kingdom: Oxford University Press.
- Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, 15(3), 250.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 25(1), 1–21.
- Stuart, E. A., & Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*, 44(2), 395–406.
- Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, 46(1), 90–118.
- Tieben, N., de Graaf, P. M., & de Graaf, N. D. (2010). Changing effects of family background on

- transitions to secondary education in the Netherlands: Consequences of educational expansion and reform. *Research in Social Stratification and Mobility*, 28(1), 77–90.
- Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology*, 98(4), 788–806.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 1–67.
- Van de gaer, E., Pustjens, H., Van Damme, J., & De Munter, A. (2006). Tracking and the effects of school-related attitudes on the language achievement of boys and girls. *British Journal of Sociology of Education*, 27(3), 293–309.
- Van de Werfhorst, H. G., & Mijs, J. J. (2010). Achievement inequality and the institutional structure of educational systems: A comparative perspective. *Annual Review of Sociology*, 36, 407–428.
- van der Wal, W. M., & Geskus, R. B. (2011). ipw: An R Package for Inverse Probability Weighting. *Journal of Statistical Software*, 43(13), 1–23.
- Van Houtte, M. (2004). Tracking effects on school achievement: A quantitative explanation in terms of the academic culture of school staff. *American Journal of Education*, 110(4), 354–388.
- Van Houtte, M., & Stevens, P. A. (2015). Tracking and sense of futility: the impact of between-school tracking versus within-school tracking in secondary education in Flanders (Belgium). *British Educational Research Journal*, 41(5), 782–800.
- Vandenberge, V. (2006). Achievement effectiveness and equity: the role of tracking, grade repetition and inter-school segregation. *Applied Economics Letters*, 13(11), 685–693.

- Vanderweele, T. J., & Arah, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology*, 22(1), 42–52.
- Vorst, H. C. M., & Smits, J. A. E. (2010). *Schoolvragenlijst voor basisonderwijs en voortgezet onderwijs*. Nijmegen, the Netherlands: Berkhout.
- Waldinger, F. (2007). Does ability tracking exacerbate the role of family background for students' test scores. *Unpublished Manuscript*.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3), 427–450.
- White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30(4), 377–399.
- Winship, C., & Morgan, S. (2007). *Counterfactuals and causal inference*. Cambridge, United Kingdom: Cambridge University Press.
- Wößmann, L. (2008). Efficiency and equity of European education and training policies. *International Tax and Public Finance*, 15(2), 199–230.
- Zhao, J. H., & Schafer, J. L. (2016). pan: Multiple imputation for multivariate panel or clustered data. *CRAN Package Repository*, 1–15.

Footnotes

1

Tables

Table 1

Baseline covariates at T0

Variable	Description	Rel.	r_{math}	r_{dutch}	Mis
Math. T0	IRT-score achievement in mathematics T0	.85	.84	.66	.03
Dutch T0	IRT-score achievement in Dutch T0	.82	.54	.61	.02
French T0	IRT-score achievement in French T0	.79	.58	.58	.05
Gender	Binary variable for boy		.07	-.15	.00
Age	Binary variable whether student is older than normally progressing students				.00
SES	Factor score socioeconomic status	.87	.50	.44	.11
Allowance	Binary variable whether family has an allowance due to low income		-.25	-.23	.00
Ed. mother	Binary variable whether mother is lowly educated		-.33	-.29	.00
Other lang.	Binary variable whether the home language is not Dutch		-.16	-.18	.00
SC Gen.	Factor score general academic self-concept	.77	.30	.26	.04
SC Math.	Factor score self-concept mathematics	.91	.47	.27	.04
SC Dutch	Factor score self-concept Dutch	.86	.13	.23	.04
SC French	Factor score self-concept French	.92	.18	.22	.04
Wellbeing	Factor score wellbeing	.82	.05	.05	.04
Mindset	Factor score mindset	.55	-.12	.13	.04
Aut. Mot.	Factor score autonomous motivation	.77	.01	.00	.04
Cont. Mot.	Factor score controlled motivation	.81	-.03	-.05	.04
Beh. Eng.	Factor score behavioral engagement	.78	.06	.06	.04
Em. Eng.	Factor score emotional engagement	.77	.01	.02	.04
Beh. Dis.	Factor score behavioral disengagement	.68	-.17	-.15	.04
Em. Dis.	Factor score emotional disengagement	.63	-.15	-.14	.04
Int. Math.	Sum score interest in mathematics	.87	.30	.13	.05
Int. Dutch	Sum score interest in Dutch	.86	-.04	.06	.05
Int. French	Sum score interest in French	.90	.02	.06	.05
Int. Tech.	Sum score interest in technology	.86	-.08	-.17	.05

Note: Rel. = Reliability; Mis = % of students with missing data; Math. = Mathematics; Ed. = education; Lang. = Language; ASC = Academic Self-Concept; Aut. = Autonomous; Contro. = Controlled; Beh = Behavioral; Eng = Engagement; Em. = Emotional; Dis = Disengagement; Tech = Technology

Table 2

Differences between tracks in baseline covariates and standardized differences between tracks

Baseline covariate	Classical	Modern	Technical	Vocational	Comp1	Comp2	Comp3
	<i>M</i>	<i>M</i>	<i>M</i>	<i>M</i>	<i>SMD</i>	<i>SMD</i>	<i>SMD</i>
Math. T0	0.86	0.30	-0.20	-1.51	0.92	0.79	1.96
Dutch T0	0.88	0.12	-0.50	-0.77	0.90	0.77	0.35
French T0	0.80	0.26	-0.36	-1.15	0.73	0.82	0.95
Gender	0.40	0.42	0.56	0.53	-0.04	-0.28	0.06
Age	-0.04	0.08	0.17	0.46	-0.40	-0.24	-0.58
SES	0.76	0.13	-0.29	-0.89	0.72	0.56	0.72
Allowance	0.09	0.18	0.25	0.40	-0.24	-0.16	-0.31
Ed. mother	0.05	0.14	0.19	0.47	-0.26	-0.13	-0.56
Other lang.	0.05	0.10	0.07	0.18	-0.17	0.12	-0.29
ASC General	0.55	0.03	-0.34	-0.34	0.56	0.39	0.00
ASC Math.	0.47	0.06	-0.25	-0.42	0.44	0.31	0.16
ASC Dutch	0.39	0.04	-0.24	-0.30	0.37	0.28	0.05
ASC French	0.50	0.08	-0.31	-0.43	0.45	0.41	0.11
Wellbeing	0.18	-0.04	-0.05	-0.07	0.23	0.01	0.02
Mindset	-0.17	-0.05	0.05	0.27	-0.12	-0.10	-0.24
Aut. Mot.	0.18	-0.13	-0.02	0.08	0.32	-0.11	-0.09
Contr. Mot.	-0.04	0.01	-0.10	0.14	-0.05	0.11	-0.22
Beh. Eng.	0.16	-0.03	-0.06	-0.05	0.19	0.03	-0.01
Em. Eng.	0.15	-0.07	-0.07	0.06	0.23	0.00	-0.12
Beh. Dis.	-0.27	-0.02	0.07	0.29	-0.27	-0.09	-0.19
Em. Dis.	-0.20	0.00	0.02	0.23	-0.22	-0.02	-0.18
Interest Math.	0.23	0.04	-0.13	-0.22	0.20	0.17	0.08
Interest Dutch	0.14	-0.02	-0.11	0.01	0.17	0.09	-0.11
Interest French	0.29	0.00	-0.22	-0.10	0.30	0.22	-0.11
Interest Tech.	-0.28	-0.13	0.27	0.33	-0.15	-0.43	-0.06

Note: SMD = Standardized mean difference; Comp1 = difference classical and modern track; Comp2 = difference modern and technical track; Comp3 = difference technical and vocational track; Math. = Mathematics; Ed. = education; Lang. = Language; ASC = Academic Self-Concept; Aut. = Autonomous; Contro. = Controlled; Beh = Behavioral; Eng = Engagement; Em. = Emotional; Dis = Disengagement, Tech = Technology

Table 3

Multilevel latent growth curves Mathematics and Dutch four tracks

Track	Mathematics					Dutch		
	M_{T0} (<i>SE</i>)	M_{T1} (<i>SE</i>)	M_{T2} (<i>SE</i>)	M_{T3} (<i>SE</i>)	SL (<i>SE</i>)	M_{T0} (<i>SE</i>)	M_{T3} (<i>SE</i>)	SL (<i>SE</i>)
classical	0.86 (0.06)	0.98 (0.06)	1.14 (0.06)	1.51 (0.07)	0.65 (0.05)	0.87 (0.06)	1.18 (0.05)	0.30 (0.06)
modern	0.30 (0.05)	0.41 (0.05)	0.63 (0.06)	0.96 (0.05)	0.65 (0.03)	0.12 (0.05)	0.59 (0.06)	0.47 (0.05)
technical	-0.19 (0.06)	-0.08 (0.07)	-0.01 (0.09)	0.46 (0.09)	0.64 (0.05)	-0.50 (0.06)	-0.16 (0.07)	0.34 (0.06)
vocational	-1.50 (0.06)	-1.36 (0.06)	-1.01 (0.05)	-0.82 (0.05)	0.68 (0.05)	-0.78 (0.05)	-0.95 (0.04)	-0.17 (0.05)

Note: $M_{T0} - M_{T3}$ = Estimated mean achievement at T0 – T3 according to latent growth curve; SL = Slope as mean achievement growth between T0 and T3

Table 4

Indicators of remaining selection bias after application of matchings procedures

Matching procedure	Classical & modern					Modern & technical					Technical & vocational				
	M_d	M_{ps}	min_d	max_d	ESS	M_d	M_{ps}	min_d	max_d	ESS	M_d	M_{ps}	min_d	max_d	ESS
Cal. 0.05	.01	.01	-.07	.05	440.00	.02	.01	-.08	.09	419.00	.04	.01	-.16	.25	134.00
Full .05-.95	.01	.01	-.07	.05	125.16	.00	.00	-.22	.11	128.11	.04	.00	-.28	.33	12.76
Maha.	.03	.05	-.09	.13	217.87	.07	.06	-.04	.23	158.43	.05	.07	-.13	.30	153.00
CEM	.04	NA	.00	.26	486.00	.07	NA	.00	.18	454.00	NA	NA	NA	NA	NA

Note: M_d = mean SMD between tracks; M_{ps} = mean propensity score difference between tracks; min_d = minimum SMD between tracks; max_d = maximum SMD between tracks; ESS = effective sample size; Cal. = Caliper; Maha. = Mahalanobis; CEM = Coarsened exact matching; NA = Not applicable

Table 5

Differences classical and modern track in matched sample at T1, T2 and T3

Matching procedure	Track	N	Mathematics						N	Dutch	
			GEE			MLGC				GEE	MLGC
			d_{T1} (SE)	d_{T2} (SE)	d_{T3} (SE)	d_{T1} (SE)	d_{T2} (SE)	d_{T3} (SE)		d_{T3} (SE)	d_{T3} (SE)
Cal. .05	clas.	440	.29*	.20*	.22*	.27*	.20	.20*	429	.23*	.23*
	mod.	440	(.05)	(.08)	(.06)	(.08)	(.11)	(.09)	429	(.07)	(.09)
Full .05-.95	clas.	663	.21*	.12	.12	.19*	.15	.12	659	.21*	.21*
	mod.	1079	(.06)	(.08)	(.08)	(.09)	(.12)	(.11)	1071	(.07)	(.10)
Maha.	clas.	486	.26*	.15*	.21*	.23*	.13	.19*	475	.22*	.16
	mod.	486	(.05)	(.07)	(.06)	(.07)	(.10)	(.08)	475	(.05)	(.09)
CEM	clas.	481	.19*	.18*	.20*	.18*	.17	.19*	164	.21*	.22*
	mod.	315	(.07)	(.08)	(.07)	(.06)	(.10)	(.07)	145	(.10)	(.10)

Note. N = number of students in matched set per track; M_{PS} = Mean propensity score in matched group; GEE = Generalized estimating equations estimates; MLGC = multilevel latent growth curve model estimates; clas. = classical track; mod. = modern track; M_{ps} = Mean propensity score; $d_{T1} - d_{T3}$ = Difference between high track and low track divided by standard deviation low track at T0 – T3; Cal. = Caliper; Maha. = Mahalanobis; CEM = Coarsened exact matching; NA = Not applicable

* Significant at $\alpha = 0.05$

Table 6

Differences modern and technical track in matched samples at T1, T2 and T3

Matching procedure	Track	N	Mathematics						N	Dutch	
			GEE			MLGC				GEE	MLGC
			d_{T1} (SE)	d_{T2} (SE)	d_{T3} (SE)	d_{T1} (SE)	d_{T2} (SE)	d_{T3} (SE)			
Cal. .05	mod.	417	.20*	.52*	.25*	.17	.47*	.23*	413	0.46*	0.46*
	tech.	417	(.05)	(.07)	(.06)	(.09)	(.15)	(.10)	413	(0.07)	(0.10)
Full .05-.95	mod.	1045	.21*	.44*	.17	.18	.37*	.13	1047	0.46*	0.50*
	tech.	642	(.07)	(.09)	(.09)	(.10)	(.17)	(.11)	640	(0.07)	(0.10)
Maha.	mod.	454	.19*	.50*	.27*	.12	.40*	.22*	456	0.43*	0.39*
	tech.	454	(.05)	(.06)	(.06)	(.09)	(.14)	(.09)	456	(0.06)	(0.09)
CEM	mod.	824	.25*	.45*	.24*	.20	.41*	.22*	917	0.55*	0.51*
	tech.	514	(.07)	(.10)	(.07)	(.10)	(.19)	(.09)	495	(0.08)	(0.10)

Note. N = number of students in matched set per track; M_{PS} = Mean propensity score in matched group; GEE = Generalized estimating equations estimates; MLGC = multilevel latent growth curve model estimates; mod. = modern track; tech. = technical track; M_{ps} = Mean propensity score; $d_{T1} - d_{T3}$ = Difference between high track and low track divided by standard deviation low track at T0 – T3; Cal. = Caliper; Maha. = Mahalanobis; CEM = Coarsened exact matching; NA = Not applicable

* Significant at $\alpha = 0.05$

Table 7

Differences technical and vocational track in matched samples at T1, T2 and T3

Matching procedure	Track	<i>N</i>	Mathematics						<i>N</i>	Dutch	
			GEE			MLGC				GEE	MLGC
			<i>d</i> _{T1} (<i>SE</i>)	<i>d</i> _{T2} (<i>SE</i>)	<i>d</i> _{T3} (<i>SE</i>)	<i>d</i> _{T1} (<i>SE</i>)	<i>d</i> _{T2} (<i>SE</i>)	<i>d</i> _{T3} (<i>SE</i>)			
Cal. .05	tech.	134	.69*	.43*	.84*	.60*	.39*	.76*	132	.44*	.50*
	voc.	134	(.12)	(.10)	(.11)	(.18)	(.13)	(.14)	132	(.11)	(.10)
Full .05-.95	tech.	334	.65*	.44*	.74*	.76*	.58*	.83*	332	.53*	.61*
	voc.	290	(.15)	(.16)	(.15)	(.23)	(.22)	(.24)	286	(.13)	(.15)
Maha.	tech.	153	.68*	.38*	.85*	.48*	.23*	.67*	149	.40*	.50*
	voc.	153	(.09)	(.10)	(.09)	(.08)	(.08)	(.07)	149	(.11)	(.11)
CEM	tech.	NA	NA	NA	NA	NA	NA	NA	83	.47*	.58*
	voc.	NA	NA	NA	NA	NA	NA	NA	112	(.22)	(.14)

Note. N = number of students in matched set per track; M_{PS} = Mean propensity score in matched group; GEE = Generalized estimating equations estimates; MLGC = multilevel latent growth curve model estimates; tech. = technical track; voc. = vocational track; M_{ps} = Mean propensity score; $d_{T1} - d_{T3}$ = Difference between high track and low track divided by standard deviation low track at T0 – T3; Cal. = Caliper; Maha. = Mahalanobis; CEM = Coarsened exact matching; NA = Not applicable

* Significant at $\alpha = 0.05$

Table 8

Effect modification of track effects in matched samples at T1, T2 and T3

Group		Classical and modern				Modern and technical				Technical and vocational			
		Mathematics		Dutch		Mathematics		Dutch		Mathematics		Dutch	
		d_{T1} (<i>SE</i>)	d_{T2} (<i>SE</i>)	d_{T3} (<i>SE</i>)	d_{T3} (<i>SE</i>)	d_{T1} (<i>SE</i>)	d_{T2} (<i>SE</i>)	d_{T3} (<i>SE</i>)	d_{T3} (<i>SE</i>)	d_{T1} (<i>SE</i>)	d_{T2} (<i>SE</i>)	d_{T3} (<i>SE</i>)	d_{T3} (<i>SE</i>)
Aca. Perf.	high	.28*	.20	.24*	.10	.00	.27	.09	.44*	.25	.23	.63*	.45
		(.10)	(.14)	(.11)	(.11)	(.16)	(.26)	(.15)	(.14)	(.24)	(.19)	(.23)	(.25)
	low	.21*	.15	.11	.36*	.34*	.59*	.35*	.48*	.67*	.54*	1.07*	.48*
		(.09)	(.12)	(.10)	(.12)	(.09)	(.12)	(.10)	(.13)	(.23)	(.24)	(.29)	(.22)
PS	high - low	.08	.05	.13	-.26	-.34*	-.31	-.26	-.04	-.41	-.31	-.44	-.03
		(.12)	(.18)	(.13)	(.14)	(.16)	(.23)	(.17)	(.18)	(.34)	(.30)	(.43)	(.36)
	high	.26*	.24	.20	.14	.11	.37	.16	.48*	.39	.36	.72*	.72*
		(.10)	(.14)	(.13)	(.13)	(.11)	(.20)	(.13)	(.14)	(.27)	(.24)	(.33)	(.25)
SES	low	.26*	.15	.19*	.31*	.22	.51*	.33*	.44*	.63*	.46	.89*	.28
		(.08)	(.12)	(.10)	(.12)	(.11)	(.15)	(.11)	(.12)	(.3)	(.24)	(.25)	(.23)
	high - low	-.01	.08	.01	-.16	-.11	-.14	-.18	.04	-.23	-.10	-.17	.43
		(.11)	(.16)	(.15)	(.18)	(.14)	(.19)	(.15)	(.16)	(.45)	(.30)	(.46)	(.32)
SES	high	.29*	.27	.22	.16	.21	.37	.26	.46*	.65*	.43	.72*	.66*
		(.11)	(.17)	(.14)	(.14)	(.12)	(.19)	(.15)	(.14)	(.24)	(.23)	(.27)	(.27)
	low	.28*	.19	.24*	.35*	.11	.57*	.19	.40*	.57	.32	.87*	.30
		(.11)	(.13)	(.12)	(.11)	(.10)	(.13)	(.10)	(.12)	(.34)	(.26)	(.28)	(.24)
SES	high - low	.01	.08	-.01	-.20	.1	-.20	.06	.06	.08	.11	-.15	.36
		(.14)	(.18)	(.16)	(.16)	(.14)	(.16)	(.15)	(.16)	(.43)	(.36)	(.38)	(.36)

Note. $d_{T1} - d_{T3}$ = Difference between high track and low track divided by standard deviation low track at T0 – T3; Aca. Perf. = Academic performance; PS = Propensity score; SES = Socioeconomic status; low = lower half of both tracks; high = higher half of both tracks; low - high = difference track effects between low and high groups

* Significant at $\alpha = 0.05$

Latent growth curve model mathematics and reading comprehension.

Figure 1.

Overlap propensity scores in pairwise comparisons of tracks.

Figure 2.

Mathematics development in matched datasets.

Figure 3.

Appendix 1: matching with Matchit package in R

```

library(MatchIt) #activation package MatchIt

DataImputation1=read.csv("DataImputation1.csv",header=TRUE,sep=";",dec=".") #uploading
data from the first imputation to the R-environment

DataImputation1<- DataImputation1 [DataImputation1$Track %in% c("Classical","Modern"),]
#select students who were in the classical and modern track

DataImputation1$Treatment[DataImputation1$Track=="Classical"]<-1 #Students who were in
the classical track are coded as having treatment status 1

DataImputation1$Treatment[DataImputation1$Track=="Modern"]<-0 #Students who were in the
modern track are coded as having treatment status 0

DataImputation1.out<-
matchit(Treatment~MATH0+DUTC0+FREN0+ACSC0+SCMA0+SCFR0+SCDU0+BEHD0+B
EHE0+CMOT0+EMDI0+MIND0+INTT0+INTM0+SES+YEAR+EDUC+ALLO+LANG+GEN
D,
data= DataImputation,distance="logit",method="nearest",caliper=0.05)

DataImputation1.matched<- match.data(DataImputation1.out)

```

Appendix 2: outcome analysis with geepack and contrast packages in R

This code follows on “Appendix 1: matching with Matchit package in R”, after having constructed censoring weights and multiplying them with the matching weights and restructuring the data from wide form (one entry pefor each student) to long form (one entry fo each time point for each student)

```
library(geepack) #activation package geepack
```

```
library(contrast) #activation package contrast
```

```
DataImputation1.GEE=geese(formula=Math~Time*Treatment+
DUTC0+FREN0+ACSC0+SCMA0+SCFR0+SCDU0+BEHD0+BEHE0+CMOT0+EMDI0+MI
ND0+INTT0+INTM0+SES+YEAR+EDUC+ALLO+LANG+GEND,data=
DataImputation1.matched.long,id=STUDID,corstr="ar1",weights=TotWeight) #Fitting the GEE
outcome model to the data, with covariates as predictors for doubly robustness. The setting
“id=STUDID” specifies that observations are clustered within students, whereas corstr="ar1"
specifies using an autoregressive working correlation matrix.
```

```
DataImputation1.Contrast1<-contrast(DataImputation1.GEE,list(Treatment=1,Time="1",
DUTC0="0",FREN0="0",ACSC0="0",SCMA0="0",SCFR0="0",SCDU0="0",BEHD0="0",BE
HE0="0",CMOT0="0",EMDI0="0",MIND0="0",INTT0="0",INTM0="0",SES="0",YEAR="0
",EDUC="0",ALLO="0",LANG="0",GEND="0"), list(Treatment=0,Time="1",
DUTC0="0",FREN0="0",ACSC0="0",SCMA0="0",SCFR0="0",SCDU0="0",BEHD0="0",BE
HE0="0",CMOT0="0",EMDI0="0",MIND0="0",INTT0="0",INTM0="0",SES="0",YEAR="0
```

”,EDUC=“0”,ALLO=“0”,LANG=“0”,GEND=“0”)) #contrasting the prediction for students having treatment 1 with students having treatment 0 at time 1, while having the same predictor values.

```
DataImputation1.Contrast2<-contrast(DataImputation1.GEE,list(Treatment=1,Time="2",
DUTC0=“0”,FREN0=“0”,ACSC0=“0”,SCMA0=“0”,SCFR0=“0”,SCDU0=“0”,BEHD0=“0”,BE
HE0=“0”,CMOT0=“0”,EMDI0=“0”,MIND0=“0”,INTT0=“0”,INTM0=“0”,SES=“0”,YEAR=“0
”,EDUC=“0”,ALLO=“0”,LANG=“0”,GEND=“0”), list(Treatment=0,Time="2",
DUTC0=“0”,FREN0=“0”,ACSC0=“0”,SCMA0=“0”,SCFR0=“0”,SCDU0=“0”,BEHD0=“0”,BE
HE0=“0”,CMOT0=“0”,EMDI0=“0”,MIND0=“0”,INTT0=“0”,INTM0=“0”,SES=“0”,YEAR=“0
”,EDUC=“0”,ALLO=“0”,LANG=“0”,GEND=“0”)) #contrasting the prediction for students
having treatment 1 with students having treatment 0 at time 2, while having the same predictor
values.
```

```
DataImputation1.Contrast3<-contrast(DataImputation1.GEE,list(Treatment=1,Time="3",
DUTC0=“0”,FREN0=“0”,ACSC0=“0”,SCMA0=“0”,SCFR0=“0”,SCDU0=“0”,BEHD0=“0”,BE
HE0=“0”,CMOT0=“0”,EMDI0=“0”,MIND0=“0”,INTT0=“0”,INTM0=“0”,SES=“0”,YEAR=“0
”,EDUC=“0”,ALLO=“0”,LANG=“0”,GEND=“0”), list(Treatment=0,Time="3",
DUTC0=“0”,FREN0=“0”,ACSC0=“0”,SCMA0=“0”,SCFR0=“0”,SCDU0=“0”,BEHD0=“0”,BE
HE0=“0”,CMOT0=“0”,EMDI0=“0”,MIND0=“0”,INTT0=“0”,INTM0=“0”,SES=“0”,YEAR=“0
”,EDUC=“0”,ALLO=“0”,LANG=“0”,GEND=“0”)) #contrasting the prediction for students
having treatment 1 with students having treatment 0 at time 3, while having the same predictor
values.
```

Appendix 3: outcome analysis with Mplus

! This code continues from “Appendix 1: matching with Matchit package in R” with

DataImputation1.matched exported as a .csv file

TITLE: Outcome analyses matched dataset

DATA: FILE IS DataImputation1.matched.csv;

Format is free;

VARIABLE:

Names=

STUDID SCHLID

MATH0 MATH1 MATH2 MATH3

DUTC0 DUTC3

FREN0

GEN0 AGE SES ALLO EDUC LANG

ACSC0 SCMA0 SCDU0 SCFR0

WELL0 MIND0 AUTM0 CMOT0

BEHE0 EMEN0 BEHD0 EMDI0

INTM0 INTDU0 INTFR0 INTT0

TREAT WEIGHTS;

IDVARIABLE=STUDID;

MISSING ARE ALL (9999999);

USEVARIABLES=

MATH0 MATH1 MATH2 MATH3

DUTC0

FREN0

ACSC0 SCMA0 SCFR0 SCDU0

BEHD0 BEHE0 CMOT0 EMDI0

MIND0 INTT0 INTM0 SES YEAR

EDUC ALLO LANG GEND;

GROUPING IS TREAT (0=LOW 1=HIGH); !analysis will be conducted on two groups, the lower track and high track respectively coded as TREAT=0 and TREAT=1;

CLUSTER= SCHLID; !the clusters are schools, accounting for students being clustered in schools

WEIGHT=WEIGHTS; !the weights from the matching procedure are incorporated in the estimation

DEFINE:

center

DUTC0

FREN0

ACSC0 SCMA0 SCFR0 SCDU0

BEHD0 BEHE0 CMOT0 EMDI0

MIND0 INTT0 INTM0 SES YEAR

EDUC ALLO LANG GEND

(GRANDMEAN);

Analysis:

TYPE=COMPLEX; !accounting for the school clusters when estimating the marginal track effects

ESTIMATOR=ML;

REPSE=BOOTSTRAP;

BOOTSTRAP=500;

Model:

ic BY MATH0@1;

ic BY MATH1@1;

ic BY MATH2@1;

ic BY MATH3@1;

sl BY MATH0@0;

sl BY MATH1;

sl BY MATH2;

sl BY MATH3@1;

[MATH0@0];

[MATH1@0];

[MATH2@0];

[MATH3@0];

ic on

DUTC0 FREN0 (b_ic1-b_ic2)

ACSC0 SCMA0 SCFR0 SCDU0 (b_ic3-b_ic6)

BEHD0 BEHE0 CMOT0 EMDI0 (b_ic7-b_ic10)

MIND0 INTT0 INTM0 SES YEAR (b_ic11-b_ic15)

EDUC ALLO LANG GEND (b_ic16-b_ic19); ! a doubly robust outcome model by regressing the intercept on the baseline covariates. Parameters of the baseline covariates are assumed to be equal within the analysis.

sl on

DUTC0 FREN0 (b_sl1-b_sl2)

ACSC0 SCMA0 SCFR0 SCDU0 (b_sl3-b_sl6)

BEHD0 BEHE0 CMOT0 EMDI0 (b_sl7-b_sl10)

MIND0 INTT0 INTM0 SES YEAR (b_sl11-b_sl15)

EDUC ALLO LANG GEND (b_sl16-b_sl19); ! a doubly robust outcome model by regressing the intercept on the baseline covariates. Parameters of the baseline covariates are assumed to be equal within the analysis.

MODEL LOW:

[ic](ic1); ! a separate intercept is estimated for the low track

[sl](sl1); ! a separate slope is estimated for the low track

sl BY MATH1 (B1_14); ! how much MATH1 loads on the slope of the low track

sl BY MATH2 (B1_15); ! how much MATH2 loads on the slope of the low track

ic; ! the residual variance in the intercept is freely estimated for the low track

sl@0; ! the residual variance in the slope is constrained to zero in the low track

MATH0(V1);

MATH1(V1);

MATH2(V1);

MATH3(V1); ! the residual error terms in mathematics performance are assumed to be equal over time within the low track

MODEL HIGH:

[ic](ic2); ! a separate intercept is estimated for the high track

[sl](sl2); ! a separate slope is estimated for the high track

sl BY MATH1 (B2_14); ! how much MATH1 loads on the slope of the high track

sl BY MATH2 (B2_15); ! how much MATH2 loads on the slope of the high track

ic; !the residual variance in the intercept is freely estimated for the high track

sl@0; ! the residual variance in the slope is constrained to zero in the high track

MATH0(V2);

MATH1(V2);

MATH2(V2);

MATH3(V2); ! the residual error terms in mathematics performance are assumed to be equal over time within the high track

Model constraint:

new(M1_1 M1_2 M1_3

M2_1 M2_2 M2_3

D1 D2 D3);

M1_1=sl1*B1_14; !estimating the growth in mathematics performance at time 1 for the low track

M1_2=sl1*B1_15; !estimating the growth in mathematics performance at time 2 for the low track

M1_3=sl1*1; !estimating the growth in mathematics performance at time 3 for the low track

M2_1=sl2*B2_14; !estimating the growth in mathematics performance at time 1 for the high track

M2_2=sl2*B2_15; !estimating the growth in mathematics performance at time 2 for the high track

M2_3=sl2*1; !estimating the growth in mathematics performance at time 3 for the high track

$D1 = M2_1 - M1_1$; !estimating the difference in growth at time 1 between tracks

$D2 = M2_2 - M1_2$; !estimating the difference in growth at time 2 between tracks

$D3 = M2_3 - M1_3$; !estimating the difference in growth at time 3 between tracks