SONO
STEUNPUNT
ONDERWIJS ONDERZOEK

ONDERWIJSVORMEN EN SCHOOLSE PRESTATIES

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

ONDERWIJSVORMEN EN SCHOOLSE PRESTATIES

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

Promotor: B. De Fraine

Research paper SONO/2017.OL1.1/12
Herwerking

Gent, januari 2018

Het Steunpunt Onderwijsonderzoek is een samenwerkingsverband van UGent, KU Leuven, VUB, UA en ArteveldeHogeschool.

Gelieve naar deze publicatie te verwijzen als volgt:
Dockx J., De Fraine B. & Vandecandelaere M. (2017). Onderwijsvormen en schoolse prestaties. Steunpunt Onderwijsonderzoek, Gent.

Voor meer informatie over deze publicatie jonas.dockx@kuleuven.be;
info@lisoproject.be

Deze publicatie kwam tot stand met de steun van de Vlaamse Gemeenschap, Ministerie voor Onderwijs en Vorming.

p.a. Coördinatie Steunpunt Onderwijsonderzoek
UGent – Vakgroep Onderwijskunde
Henri Dunantlaan 2, BE 9000 Gent

Deze publicatie is ook beschikbaar via www.steunpuntsono.be en www.lisoproject.be

Beleidssamenvatting

In het Vlaamse secundair onderwijs zijn er vier onderwijsvormen: het algemeen secundair onderwijs (aso), het technisch secundair onderwijs (tso), het beroepssecundair onderwijs (bso) en het kunstsecundair onderwijs (kso). Binnen het aso wordt daarbij vaak een onderscheid gemaakt tussen klassieke talen en moderne studierichtingen. Deze onderwijsvormen worden pas formeel ingericht vanaf de tweede graad van het secundair onderwijs, maar in de praktijk spreken leerlingen, ouders en scholen al in termen van onderwijsvormen in de eerste graad. In heel wat scholen zijn de onderwijsvormen reeds 'te herkennen' in het onderwijsaanbod van de eerste graad. In het tweede leerjaar van de eerste graad worden namelijk basisopties ingericht die aansluiten op deze onderwijsvormen. De meeste scholen gebruiken hun pedagogische vrijheid voor het invullen van lesuren in het eerste leerjaar ook als voorbereiding op de onderwijsvormen in de bovenbouw. In de eerste graad bereiden het eerste leerjaar B en het beroepsvoorbereidend leerjaar voor op het bso.

Bij beleidsmakers is er discussie over mogelijke effecten van deze onderwijsvormen op schoolse prestaties. Voorstanders argumenteren dat onderwijsvormen die aansluiten op de vaardigheden en interesses van leerlingen de schoolse prestaties van leerlingen verbeteren. Tegenstanders argumenteren echter dat sociale ongelijkheid in schoolse prestaties tussen leerlingen versterkt wordt doordat de onderwijsvormen verschillen in hun mogelijkheden tot leerwinst.

In wetenschappelijk onderzoek wordt het inrichten van verschillende onderwijsvormen *tracking* genoemd. Er zijn diverse studies die onderwijssystemen met *tracking* (categoriale onderwijssystemen) vergelijken met onderwijssystemen zonder *tracking* (eerder comprehensieve onderwijssystemen). Deze studies tonen over het algemeen geen positief of negatief effect van *tracking* op de gemiddelde schoolse prestaties van onderwijssystemen. Een meerderheid van deze studies toont wel dat *tracking* de sociale ongelijkheid in schoolse prestaties versterkt, maar het effect is doorgaans beperkt. Studies die onderwijssystemen vergelijken beschrijven echter alleen gemiddelde verschillen tussen groepen van onderwijssystemen. De internationaal vergelijkende studies gaan dus niet in op de precieze effecten van de (gepercipieerde) hiërarchie tussen *tracks* binnen een land. In dit rapport willen we daarom inzoomen op de effecten van *tracking* binnen Vlaanderen.

Onderzoek naar de effecten van *tracks* op schoolse prestaties van leerlingen in Vlaanderen is vereist om na te gaan of deze de ongelijkheid tussen leerlingen versterken. Als *tracks* namelijk deze ongelijkheid versterken, dan verwachten we dat *tracks* met een instroom van initieel sterker presterende leerlingen ook meer leerwinst maken. Hiervoor moet de gemiddelde leerwinst per *track* vergeleken worden. Eventuele verschillen in leerwinst tussen *tracks* zijn dan wel mogelijk toe te schrijven aan verschillen in instroom van leerlingen. Daarom moet ook onderzocht worden of er effecten zijn van *tracks* op vergelijkbare leerlingen die in verschillende *tracks* zitten. Er zijn dus twee onderzoeksvragen:

1. Verschillen *tracks* in gemiddelde leerwinst?
2. Verschillen *tracks* in gemiddelde leerwinst voor vergelijkbare leerlingen?

Voor dit onderzoek gebruiken we de gegevens van het onderzoek 'Loopbanen in het Secundair Onderwijs' (LiSO-project). De substeekproef bestaat uit 3025 leerlingen die in september 2013 startten in het secundair onderwijs in 45 Vlaamse scholen. We onderscheiden vier groepen van studiekeuzes in het eerste jaar secundair onderwijs: (1) klassieke talen (KT), (2) moderne wetenschappen (MW), (3) technisch onderwijs (TO) en (4) beroepsvoorbereidend onderwijs (BV). Hoewel er in het eerste jaar secundair onderwijs nog geen officiële onderwijsvormen onderscheiden worden, sluit de studiekeuze in het eerste jaar SO wel sterk aan bij de onderwijsvormen die in de bovenbouw zullen volgen. In dit Engelstalige rapport wordt daarom wél gesproken over '*tracking*' in het eerste jaar secundair onderwijs, omdat het gaat over het groeperen van leerlingen voor een volledig schooljaar voor (quasi) alle vakken.

De steekproef is verspreid over de vier 'tracks' als volgt: 691 leerlingen zaten in KT, 1285 leerlingen zaten in MW, 663 leerlingen zaten in TO en 566 leerlingen zaten in BV. Enkel leerlingen die de eerste drie jaar van het secundair onderwijs in dezelfde *track* zitten werden opgenomen in deze substeekproef. Drie LiSO-scholen die kiezen voor een heterogene klassamenstelling in het eerste jaar, werden geschrapt uit de steekproef van deze studie omdat er dus niet aan *tracking* wordt gedaan. Toetsen en vragenlijsten werden afgenomen aan de start van het secundair onderwijs (september 2013), op het einde van het eerste leerjaar van de eerste graad (mei 2014), op het einde van het tweede leerjaar van de eerste graad (mei 2015) en op het einde van eerste leerjaar van de tweede graad (mei 2016). Prestaties voor wiskunde werden gemeten op elk van deze momenten. Dit onderzoek beschrijft dus de effecten van *tracks* tijdens de eerste drie jaar van het secundair onderwijs op wiskunde.

Om vergelijkbare leerlingen in verschillende *tracks* te vinden gebruiken we *matching* methoden. Deze zijn gericht op het vinden van vergelijkbare personen in verschillende omgevingen. Leerlingen werden *gematched* op basis van schoolse prestaties, sociaaleconomische achtergrond en psychosociale variabelen die gemeten waren in september 2013. In totaal werd de vergelijkbaarheid van de leerlingen bepaald aan de hand van 25 variabelen. Om onze resultaten methode-onafhankelijk te maken gebruiken we verschillende *matching*-methoden. Bij elk van deze methoden bleek dat er enkel (voldoende) vergelijkbare leerlingen waren tussen bepaalde *tracks*. KT wordt daarom vergeleken met het MW, MW wordt vergeleken met TO en TO wordt vergeleken met BO. Er moet opgemerkt worden dat het aantal vergelijkbare leerlingen tussen TO en BV eerder beperkt is. Verschillen tussen *tracks* in gemiddelde leerwinst worden tweemaal berekend: (1) zonder het *matchen*, dus voor alle leerlingen, en (2) na het *matchen* van vergelijkbare leerlingen in verschillende *tracks*.

Voor de eerste onderzoeksvraag vinden we dat er bij het begin van het secundair onderwijs grote verschillen zijn in schoolse prestaties tussen de *tracks*. We stellen vast dat de leerlingen in KT de hoogste gemiddelde aanvangsscores hebben voor wiskunde. Daarna volgen de leerlingen in MW, de leerlingen in TO en de leerlingen in BO. Voor wiskunde maken de vier *tracks* een gelijke leerwinst over de eerste drie jaar secundair onderwijs. Anders gezegd: de kloof in wiskundeprestaties tussen de vier groepen blijft ongeveer dezelfde tijdens de drie eerste jaren in het secundair onderwijs. De wiskundeverschillen tussen de vier groepen worden niet groter, maar ook niet kleiner. Voor wiskunde wordt dus niet voldaan aan de hypothese dat tracks met een sterkere instroom van leerlingen meer leerwinst maken.

Voor de tweede onderzoeksvraag vinden we voor vergelijkbare leerlingen in verschillende *tracks* dat er in *tracks* met een gemiddeld sterkere leerlinginstroom significant meer leerwinst gemaakt wordt. We zien echter dat bij de vergelijking KT en MW, en bij de vergelijking MW en TO het effect klein is. Enkel voor de vergelijking TO en BV is het effect groot. Een leerling die kiest voor BV, maakt dus gemiddeld minder leerwinst voor wiskunde dan een vergelijkbare leerling die kiest voor TO. We hebben ook zicht op wanneer dit verschil ontstaat en zien dat dit effect zich vooral in het eerste jaar manifesteert. Het effect vergroot niet meer in de daaropvolgende jaren.

Een sterk punt van dit onderzoek is dat door de matching-methode nagegaan kan worden hoe vergelijkbare leerlingen zouden presteren als ze in een andere *track* zouden zitten. Dit is vooral mogelijk doordat *tracking* in Vlaanderen een eigenschap heeft die niet kenmerkend is voor de meeste andere onderwijssystemen. In Vlaanderen verloopt het verdelen van leerlingen in *tracks* immers niet op basis van objectieve criteria (bijvoorbeeld een instaptoets). Hierdoor verschillen de *tracks* wel gemiddeld op het vlak van instroomniveau, maar vinden we nog steeds veel vergelijkbare leerlingen terug in verschillende *tracks*. In andere onderwijssystemen zien we dat er minder of nauwelijks vergelijkbare leerlingen zijn in verschillende tracks.

We concluderen dat zowel in KT, MW als TO de gemiddelde leerlingen in elk van deze *tracks* een eerder gelijkaardige leerwinst maken voor wiskunde. Wanneer we naar vergelijkbare leerlingen kijken in aso KT, MW en TO vinden we wel positieve effecten van naar een *track* gaan met een sterkere leerlinginstroom. Deze effecten zijn echter vrij klein en lijken eerder beperkt bij te dragen aan ongelijkheid in schoolse prestaties. Voor vergelijkbare leerlingen in het TO en BV vinden we wel grote negatieve effecten van naar het BV gaan. Hierdoor wordt de ongelijkheid tussen leerlingen in schoolse prestaties wel merkbaar versterkt.

De resultaten geven hoofdzakelijk weer hoe leerlingen beïnvloed worden door de huidige structuur van het secundair onderwijs. We tonen dat 'hoog mikken', een strategie die vaak gebruikt wordt bij studiekeuze, slechts een beperkt positief effect heeft op schoolse prestaties voor de vergelijking KT en MW, en de vergelijking in MW en TO. Dit klein positief effect ontstaat daarbij enkel in het eerste jaar, wat naar onze mening vooral aantoont dat sterkere leerlingen in MW en TO tijdens het eerste jaar enigszins meer uitdaging nodig hebben. Hoog mikken in de vergelijking TO en BV toont een groter effect, echter heeft deze uitspraak enkel betrekking op sterkste leerlingen in het BV. De sterkste leerlingen van het BV hebben dus meer uitdaging nodig dan hun nu geboden wordt. Wanneer we de resultaten van dit onderzoek samenleggen met de resultaten van het onderzoek naar de effecten van tracks op academisch zelfconcept (SONO/2017.OL1.1/13), dan blijkt er een interessante afweging te zijn. Bij de vergelijking MW en TO, en de vergelijking TO en BV is 'hoog mikken' immers net negatief voor academisch zelfconcept. Verder onderzoek naar de effecten van *tracks* op een breder scala van uitkomsten is echter lopende, alsook de effecten van *track* verandering. Zo kunnen de effecten van *tracks* binnen het Vlaamse onderwijs beter geduid worden en welke afwegingen bij studiekeuze gemaakt moeten worden.

Inhoud

1. Introduction

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.

Ascertaining the effects of tracks requires data and methods that can control for selection bias resulting from differential student intake across tracks. Moreover, longitudinal data and methods that can describe academic development are preferable (e.g. Raudenbush, 2001; Robins, 1997). To control for selection bias, we matched comparable students across different tracks and compared academic outcomes between higher and lower tracks. In the following sections, we describe tracking literature and the research strategy in more detail.

1.1. 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.

1.2. 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 starting 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.

1.3. 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 low tracks memorization and disciplining students are emphasized (Kunter & Baumert, 2006; Retelsdorf, Butler, Streblow, & Schiefele, 2010; Van Houtte, 2004). Though, 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.

1.4. 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 is 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. Further, previous studies did not asses how track effects may change over time or across student characteristics. Consequently, further research on track effects within education systems is required.

1.5. 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, p3p6). 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, or vocational. Each track has its own educational program, though the first three tracks share a common core of educational goals. Tracks are 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 student academic performance. 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 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.

2. Method

2.1. 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) 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 remaining subsample of 5483 students (41.55%) change track during the first three years of secondary education; these students were excluded from the analyses as well. This resulted in a subsample of 3205 students in 338 classes in 45 schools in grade 7 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 students 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 through 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.

2.2. Treatment variables

Of main interest was the effect of going to a higher track on student academic performance, 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.

2.3. Measures

2.3.1. Outcome

The outcome of interest was student academic performance in mathematics. 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 educational goals set by the government and are considered a valid measurement 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 Cronbachs' Alpha. The retest stability between measurement occasions ranged from 0.65 to 0.75.

2.3.2. 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). Table 1 gives a brief overview of the 25 covariates used during the different matching procedures by giving an indication of their theoretical background, their reliability, with what instrument they were measured, and their correlation with both mathematics at T3. These variables were measured or obtained at T0.

Table 1 Baseline covariates at T0

Variable	Description	Rel.	Info	r_{math}	Mis
Math. T0	IRT-score achievement in mathematics T0	.85	AT	.84	.03
Dutch T0	IRT-score achievement in Dutch T0	.82	AT	.54	.02
French T0	IRT-score achievement in French T0	.79	AT	.58	.05
Boy	Binary indicator for boy		OR	.07	.00
Age	Categorical variable years behind grade		OR		.00
SES	Factor score socioeconomic status: based on seven indicators: (1) Highest diploma father, (2) Highest diploma mother, (3) Employment status father, (4) Employment status mother, (5) Occupational level father, (6) Occupational level mother and (7) Income.	.87	PQ	.50	.11
Allowance	Binary variable whether family is eligible for an allowance due to low income		OR	-.25	.00
Ed. mother	Binary variable whether mother is lowly educated		OR	-.33	.00
Other lang.	Binary variable whether the home language is not Dutch		OR	-.16	.00
ASC General	Factor score general academic self-concept based on 4 items	.77	SQ	.30	.04
ASC Math.	Factor score academic self-concept mathematics based on 6 items	.91	SQ	.47	.04
ASC Dutch	Factor score academic self-concept Dutch based on 6 items	.86	SQ	.13	.04
ASC French	Factor score academic self-concept French based on 6 items	.92	SQ	.18	.04
Wellbeing	Factor score wellbeing based on 9 items	.82	SQ	.05	.04
Mindset	Factor score mindset (i.e. if intelligence in considered as static or flexible) based on 3 items	.55	SQ	-.12	.04
Aut. Mot.	Factor score autonomous motivation based on 4 items	.77	SQ	.01	.04
Contr. Mot.	Factor score controlled motivation based on 8 items	.81	SQ	-.03	.04
Beh. Eng.	Factor score behavioral engagement based on 5 items	.78	SQ	.06	.04
Em. Eng.	Factor score emotional engagement based on 4 items	.77	SQ	.01	.04
Beh. Dis.	Factor score behavioral disengagement based on 5 items	.68	SQ	-.17	.04
Em. Dis.	Factor score emotional disengagement based on 6 items	.63	SQ	-.15	.04
Int. Math.	Sum score interest in mathematics based on 2 items		SQ	.30	.05
Int. Dutch	Sum score interest in Dutch based on 2 items		SQ	-.04	.05
Int. French	Sum score interest in French based on 2 items		SQ	.02	.05
Int. Tech.	Sum score interest in technology based on 2 items		SQ	-.08	.05

Note: Rel. = Reliability; Info. = Information source; Mis = % of students with missing data; Math. = Mathematics; AT = Achievement Test; OR = Official Records; SQ = Student Questionnaire; PQ = Parent Questionnaire; ASC = Academic Self-Concept

2.4. Area of common support

A prerequisite for matching students in different tracks is that there is 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 of the treated (ATT) and the local average treatment effect (LATE). In our study, the ATT is the average effect of being in a higher track for higher track students. However, if there is no complete overlap in the baseline covariates, the ATT cannot be estimated (Stuart, 2010). In our study, this meant that the average effect of the higher track could only be estimated for a specific subpopulation, this is the LATE (Imbens, 2010). How close a LATE approximates an ATT depends on the area of common support available and the matching procedure used. The area of common support is determined by the overlap in propensity scores (Steiner & Cook, 2012). We therefore assessed the overlap in the density plots of propensity scores of both tracks for each comparison.

2.5. 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, for track allocation in the matched dataset would be uncorrelated with said outcome (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.

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).

2.5.1. 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. These propensities were than 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 (Austin, 2011; Austin, Grootendorst, Normand, & Anderson, 2007; Myers et al., 2011) (see Table 1). Next, matching procedures were applied, in which students of both tracks were matched based on their 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 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. Further, the number of students that are matched within one matched set can vary. Particularly, one lower track student can be matched to a single higher track student, or one lower track student can be matched to multiple higher track students (i.e., replacement). The latter is less biased, but also less efficient. Therefore, matching with caliper 0.05SD was conducted both with and without replacement. Lastly, multiple lower track students within the caliper of one higher track student can be matched. Allowing for this multiple matching should increase efficiency, but also the bias. Therefore, the matching with caliper 0.05SD was conducted both as one-to-one (1:1) matching, one-to-three (1:3) matching and one-to-three (1:3) matching with replacement.

Using full matching, lower track students were matched to higher track students within the same propensity score interval (Stuart & Green, 2008). Weights were estimated per interval so that both tracks are equally represented per interval. More extreme weights occur in intervals with extreme propensities. Therefore, a trade-off is made between limiting the propensity score intervals for which weights are estimated and maximizing the number of students in the matched dataset. Accordingly, we varied the matching procedures by minimum and maximum propensity score included during full matching. We applied both full matching to students with propensity scores between 0.05 and 0.95 and full matching to students with propensity scores between 0.10 and 0.90.

2.5.2. 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. 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.

2.5.3. 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 is 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 than reweighted to have equal numbers in both tracks.

2.6. 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 betweencluster 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 dT1, dT2 and dT3. 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 independence 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). NewtonRaphson was used for parameter estimation. The baseline model is shown below:

$$E(Y_T) = \beta_0 + \sum_{t=1}^{3} \beta_t LowerTrack + \sum_{t=0}^{3} \beta_{t+4} HigherTrack$$

Y_T represents the average performance if the population would have followed track T at a specific time point. β_0 indicates average performance at T0 for the lower track. Parameters β_1 to β_3 represent the average change in performance for the lower track from time point T1 to T3. β_4 to β_7 represent the average difference in performance between the higher and lower track from time point T0 to T3. 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 0.05*SD* and MDM). Furthermore, covariates used in the propensity score estimation are added to this model as well, removing any remaining confounding (i.e. double robustness; Schafer & Kang, 2008). 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.

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). 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 is shown in Figure 1.

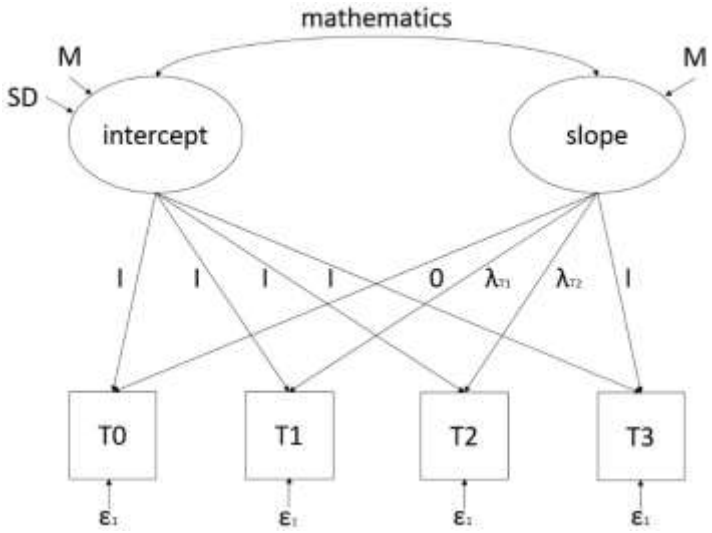

Figure 1. *Latent growth curve model mathematics.*

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).

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. The significance was tested by assessing whether the estimate 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).

2.7. Assessing differential track effects

Assessing the differences in track effects over time for Mathematics was done by placing equality constraints between the track effects in the MLGC-model. Equality constraints were placed between dT1 and dT2, between dT2 and dT3 and between dT1 and dT3 for each track comparison. Subsequently, the loglikelihood was compared to the unconstrained model. These tested 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 jointly estimated separately for low and high performance students, low and high SES students, and low and high propensity score students. Equality constraints for the track effects were added, and model fit was compared to the unconstrained model.

2.8. 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). All 25 baseline covariates were included in the imputation model (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 & GroothuisOudshoorn, 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. 10.55% was censored at T1, 4.49% at T2 and 5.93% at T3. To obtain unbiased and efficient estimates, two approaches were used to handle censoring, tailored to the two outcomes 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).

3. Results

3.1. Track differences before matching

Table 2 shows the distribution for each baseline covariate per track, 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 Differences between tracks in baseline covariates and standardized differences between tracks

Baseline covariate	Classical		Modern		Technical		Vocational		Comp1	Comp2	Comp3
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*SMD*	*SMD*	*SMD*
Math. T0	0.86	0.61	0.30	0.61	-0.20	0.63	-1.51	0.67	0.92	0.79	1.96
Dutch T0	0.88	0.83	0.12	0.84	-0.50	0.81	-0.77	0.78	0.90	0.77	0.35
French T0	0.80	0.73	0.26	0.74	-0.36	0.76	-1.15	0.83	0.73	0.82	0.95
Gender (boy)	0.40	0.49	0.42	0.49	0.56	0.50	0.53	0.50	-0.04	-0.28	0.06
Age	-0.04	0.27	0.08	0.30	0.17	0.38	0.46	0.50	-0.40	-0.24	-0.58
SES	0.76	0.88	0.13	0.88	-0.29	0.75	-0.89	0.83	0.72	0.56	0.72
Allowance	0.09	0.29	0.18	0.38	0.25	0.43	0.40	0.49	-0.24	-0.16	-0.31
Ed. mother	0.05	0.22	0.14	0.35	0.19	0.39	0.47	0.50	-0.26	-0.13	-0.56
Other lang.	0.05	0.23	0.10	0.30	0.07	0.25	0.18	0.38	-0.17	0.12	-0.29
ASC General	0.55	0.81	0.03	0.93	-0.34	0.96	-0.34	1.10	0.56	0.39	0.00
ASC Math.	0.47	0.82	0.06	0.93	-0.25	1.01	-0.42	1.08	0.44	0.31	0.16
ASC Dutch	0.39	0.88	0.04	0.94	-0.24	0.99	-0.30	1.10	0.37	0.28	0.05
ASC French	0.50	0.78	0.08	0.94	-0.31	0.96	-0.43	1.10	0.45	0.41	0.11
Wellbeing	0.18	0.94	-0.04	0.95	-0.05	1.02	-0.07	1.12	0.23	0.01	0.02
Mindset	-0.17	1.06	-0.05	0.98	0.05	0.99	0.27	0.93	-0.12	-0.10	-0.24
Aut. Mot.	0.18	0.93	-0.13	0.96	-0.02	1.04	0.08	1.08	0.32	-0.11	-0.09
Contr. Mot.	-0.04	1.00	0.01	0.94	-0.10	1.04	0.14	1.07	-0.05	0.11	-0.22
Beh. Eng.	0.16	0.91	-0.03	0.98	-0.06	1.02	-0.05	1.10	0.19	0.03	-0.01
Em. Eng.	0.15	0.96	-0.07	0.96	-0.07	1.02	0.06	1.08	0.23	0.00	-0.12
Beh. Dis.	-0.27	0.95	-0.02	0.94	0.07	0.98	0.29	1.13	-0.27	-0.09	-0.19
Em. Dis.	-0.20	0.91	0.00	0.93	0.02	1.04	0.23	1.16	-0.22	-0.02	-0.18

Interest Math.	0.23	0.90	0.04	0.97	-0.13	1.01	-0.22	1.11	0.20	0.17	0.08
Interest Dutch	0.14	0.96	-0.02	0.96	-0.11	1.01	0.01	1.10	0.17	0.09	-0.11
Interest French	0.29	0.88	0.00	0.97	-0.22	0.99	-0.10	1.12	0.30	0.22	-0.11
Interest Tech.	-0.28	1.04	-0.13	0.97	0.27	0.93	0.33	0.93	-0.15	-0.43	-0.06

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

Table 3 describes the LGCM's across tracks, 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). 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.

Table 3 Multilevel latent growth curves Mathematics four tracks

Track	Mathematics				
	M_{T0} (*SE*)	M_{T1} (*SE*)	M_{T2} (*SE*)	M_{T3} (*SE*)	*SL* (*SE*)
classical	0.86 (0.06)	0.98 (0.06)	1.14 (0.06)	1.51 (0.07)	0.65 (0.05)
modern	0.30 (0.05)	0.41 (0.05)	0.63 (0.06)	0.96 (0.05)	0.65 (0.03)
technical	-0.19 (0.06)	-0.08 (0.07)	-0.01 (0.09)	0.46 (0.09)	0.64 (0.05)
vocational	-1.50 (0.06)	-1.36 (0.06)	-1.01 (0.05)	-0.82 (0.05)	0.68 (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*

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. 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 the technical and vocational track comparison did not yield enough matches, thus no analyses on this matched dataset were conducted.

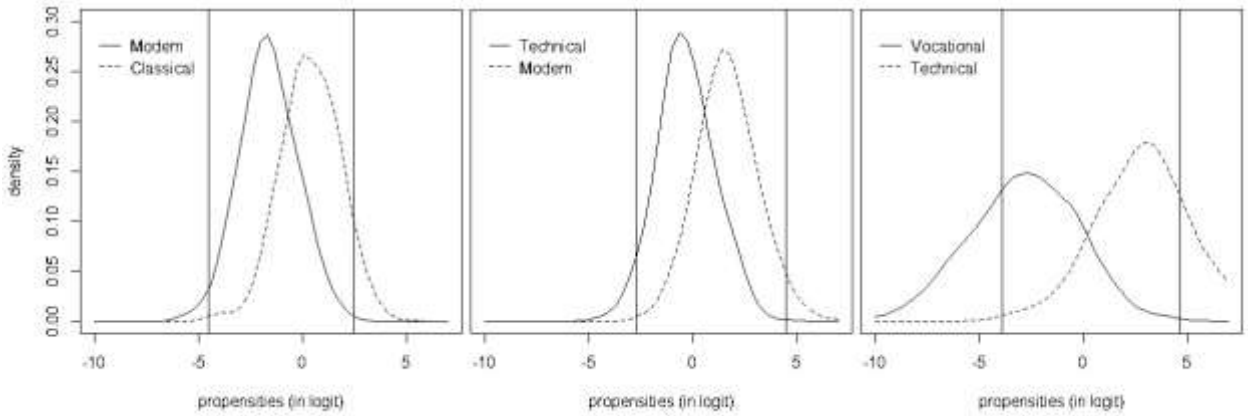

Figure 2. Overlap propensity scores in pairwise comparisons of tracks

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.

3.2. 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 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	M_d	M_{ps}	min_d	max_d	M_d	M_{ps}	min_d	max_d
Cal. 0.05 1:1	0.01	0.01	-0.07	0.05	0.02	0.01	-0.08	0.09	0.04	0.01	-0.16	0.25
Cal. 0.05 rep.	0.00	0.00	-0.10	0.16	0.03	0.00	-0.21	0.28	-0.03	0.00	-0.31	0.29
Cal. 0.05 1:3	0.00	0.00	-0.09	0.12	0.03	0.00	-0.18	0.29	-0.01	0.00	-0.31	0.30
Full .05-.95	0.01	0.01	-0.07	0.05	0.00	0.00	-0.22	0.11	0.04	0.00	-0.28	0.33
Full .10-.90	-0.01	0.00	-0.13	0.14	0.01	0.00	-0.19	0.17	0.00	0.00	-0.29	0.18
Maha.	0.03	0.05	-0.09	0.13	0.07	0.06	-0.04	0.23	0.05	0.07	-0.13	0.30
CEM	0.04	NA	0.00	0.26	0.07	NA	0.00	0.18	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; min_{d} = maximum SMD between tracks

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, satisfactory balance is achieved between tracks per comparison.

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, except for 1:1 matching with caliper 0.05SD with replacement and 1:3 matching with caliper 0.05SD with replacement. 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.

Although all matching procedures reach balance in mean SMD, except CEM, the resulting matched sets have different mean propensities of being in a higher track. Tables 5, 6 and 7 show the mean propensities per track per matched sample. Across the three comparisons matching with caliper 0.05SD has the lowest propensities, adding replacement heightens the propensities, while allowing for multiple matches causes no change. Full matching has higher mean propensities, trending higher when allowing more extreme propensities into the weighting scheme. MDM usually yields propensities somewhat comparable to 1:1 nearest neighbor 0.05 caliper matching

Another difference is the number of students in the matched samples, also shown in Tables 5, 6 and 7. matching with caliper 0.05SD produces the smallest matched sets. Allowing for replacement increases the number of students in the higher track, but lowers the number of students in the lower track. Allowing multiple matches increases the number of students in the lower tracks. Full matching with students between 0.05 and 0.95 propensity score yields the largest sample sizes. Full matching with students between 0.10 and 0.90 propensity score reduces the number of matches. MDM attains datasets comparable to matching with caliper 0.05SD. The number of matched students differs strongly for CEM, though it is generally small.

3.3. 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 Tables 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.

Table 5 Differences classical and modern track in matched sample at T1, T2 and T3

Matching procedure	Track	*N*	M_{PS}	Mathematics					
				GEE			MLGC		
				d_{T1} (*SE*)	d_{T2} (*SE*)	d_{T3} (*SE*)	d_{T1} (*SE*)	d_{T2} (*SE*)	d_{T3} (*SE*)
Cal. .05	clas.	440	.44	.29*	.20*	.22*	.27*	.20	.20*
	mod.	440	.43	(.05)	(.08)	(.06)	(.08)	(.11)	(.09)
Cal. .05 with rep.	clas.	656	.55	.22*	.14	.14	.23*	.17	.16
	mod.	344	.55	(.06)	(.09)	(.08)	(.09)	(.12)	(.10)
Cal. .05 1 to 3	clas.	656	.55	.22*	.13	.14*	.23*	.17	.17
	mod.	623	.55	(.06)	(.08)	(.07)	(.08)	(.12)	(.10)
Full .05 .95	clas.	663	.57	.21*	.12	.12	.19*	.15	.12
	mod.	1079	.57	(.06)	(.08)	(.08)	(.09)	(.12)	(.11)
Full .10 .90	clas.	600	.55	.28*	.16	.22*	.31*	.22	.25*
	mod.	847	.55	(.06)	(.10)	(.08)	(.08)	(.13)	(.11)
Maha.	clas.	486	.46	.26*	.15*	.21*	.23*	.13	.19*
	mod.	486	.42	(.05)	(.07)	(.06)	(.07)	(.10)	(.08)
CEM	clas.	481	NA	.19*	.18*	.20*	.18*	.17	.19*
	mod.	315	NA	(.07)	(.08)	(.07)	(.06)	(.10)	(0.07)

Note. N = number of students in matched set per track; 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 ; NA = Not applicable. * Significant at α = 0.05

For the modern and technical track comparison, all effects on mathematics range from *d* = 0.19 to *d* = 0.25 at T1, from *d* = 0.43 to *d* = 0.52 at T2 and from *d* = 0.14 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.

Table 6 Differences modern and technical track in matched samples at T1, T2 and T3

Matching procedure	Track	*N*	M_{PS}	Mathematics					
				GEE			MLGC		
				d_{T1} (*SE*)	d_{T2} (*SE*)	d_{T3} (*SE*)	d_{T1} (*SE*)	d_{T2} (*SE*)	d_{T3} (*SE*)
Cal. .05	mod.	417	.58	.20*	.52*	.25*	.17	.47*	.23*
	tech.	417	.57	(.05)	(.07)	(.06)	(.09)	(.15)	(.10)
Cal. .05 with rep.	mod.	1285	.77	.20*	.43*	.14	.16	.40	.12
	tech.	349	.77	(.07)	(.09)	(.10)	(.15)	(.24)	(.17)
Cal. .05 1 to 3	mod.	1285	.77	.20*	.44*	.14	.17	.41	.12
	tech.	497	.77	(.07)	(.09)	(.10)	(.16)	(.24)	(.18)
Full .05 .95	mod.	1045	.73	.21*	.44*	.17	.18	.37*	.13
	tech.	642	.73	(.07)	(.09)	(.09)	(.10)	(.17)	(.11)

Full .10 .90	mod.	850	.68	.20*	.47*	.27*	.15	.38*	.22*
	tech.	582	.68	(.07)	(.10)	(.08)	(.08)	(.16)	(.09)
Maha.	mod.	454	.61	.19*	.50*	.27*	.12	.40*	.22*
	tech.	454	.55	(.05)	(.06)	(.06)	(.09)	(.14)	(.09)
CEM	mod.	824	NA	.25*	.45*	.24*	.20	.41*	.22*
	tech.	514	NA	(.07)	(.10)	(.07)	(.10)	(.19)	(.09)

Note. N = number of students in matched set per track; 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 ; NA = Not applicable. * Significant at α = 0.05

For the technical and vocational track comparison, all effects on mathematics range from d = 0.52 to d = 0.69 at T1, from d = 0.31 to d = 0.49 at T2 and from d = 0.68 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.

Table 7 Differences technical and vocational track in matched samples at T1, T2 and T3

Note. N = number of students in matched set per track; 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; NA = Not applicable. * Significant at α = 0.05

Matching procedure	Track	*N*	M_{PS}	Mathematics					
				GEE			MLGC		
				d_{T1} (*SE*)	d_{T2} (*SE*)	d_{T3} (*SE*)	d_{T1} (*SE*)	d_{T2} (*SE*)	d_{T3} (*SE*)
Cal. .05 with rep.	tech.	294	.70	.67*	.49*	.78*	.66*	.56*	.80*
	voc.	106	.70	(.13)	(.17)	(.15)	(.33)	(.25)	(.26)
Cal. .05 1 to 3	tech.	294	.70	.66*	.46*	.78*	.74*	.60*	.83*
	voc.	165	.70	(.13)	(.15)	(.14)	(.30)	(.26)	(.28)
Full .05 .95	tech.	334	.73	.65*	.44*	.74*	.76*	.58*	.83*
	voc.	290	.72	(.15)	(.16)	(.15)	(.23)	(.22)	(.24)
Full .10 .90	tech.	227	.65	.52*	.31*	.68*	.41*	.39*	.67*
	voc.	226	.65	(.12)	(.12)	(.10)	(.11)	(.12)	(.09)
Maha.	tech.	153	.55	.68*	.38*	.85*	.48*	.23*	.67*
	voc.	153	.48	(.09)	(.10)	(.09)	(.08)	(.08)	(.07)
CEM	tech.	NA	NA	NA	NA	NA	NA	NA	NA
	voc.	NA	NA						
Cal. .05	tech.	134	.52	.69*	.43*	.84*	.60*	.39*	.76*

3.4. 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, 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). Concerning the modern and technical track comparison, 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). Concerning the technical and vocational track comparison, 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).

3.5. 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 dT2 is larger than dT1 ($F(1,430) = 4.12$, $p < .05$) and that dT2 is larger than dT3 ($F(1,430) = 2.05$, $p < .05$). However, there is no significant difference between dT1 and dT3 ($F(1,430) = 0.93$, $p < .05$). For the technical and vocational track comparison, dT1 is significantly larger than dT2 ($F(1,430) = 2.08$, $p < .05$) and dT3 is significantly larger than dT2 ($F(1,430) = 2.31$, $p < .05$), but there is no significant difference between dT1 and dT3. However, when applying different matching procedures, some of these effects are insignificant, even when the effect size is comparable.

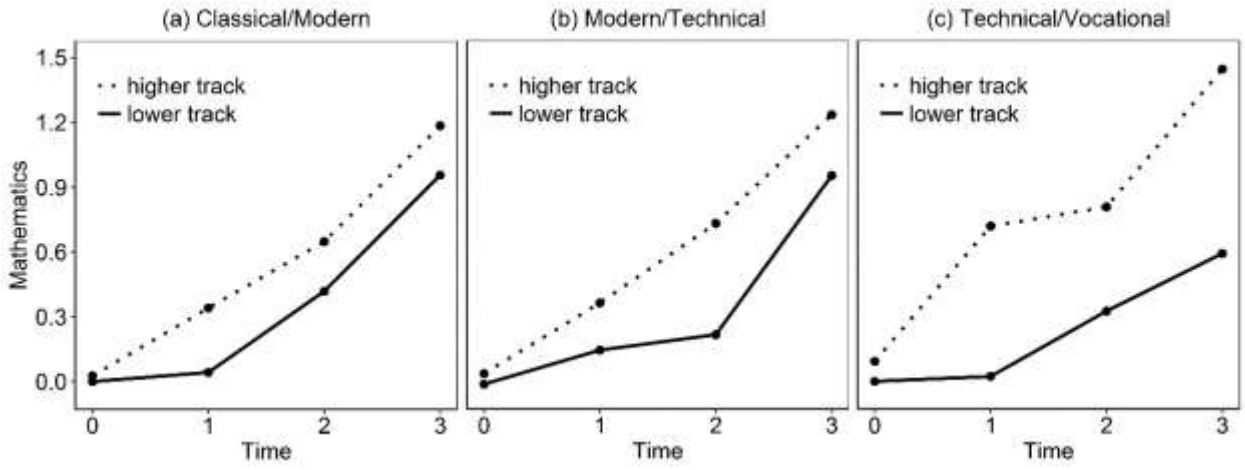

Figure 3. *Mathematics development in matched datasets.*

3.6. 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

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

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

Note. d_{T1} – d_{T3} = Difference between high track and low track divided by standard deviation low track at T0 – T3; low = lower performing half of both tracks; high = higher performing half of both tracks; low - high = difference track effects between low and high groups. * Significant at α = 0.05

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.

4. Discussion

This study investigated whether tracks affect academic performance for mathematics 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 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). 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 and track comparisons, indicating that there is no single track effect within an education system. Moreover, 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 nonacademic track across four years for intelligence development. Our own average effect size at the end of the third year of 0.40 across all track comparisons and matching methods 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.

5. 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, 2012). 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.

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 and caliper matching with replacement). 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 (the LATE). 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.

As in any study we were limited by our data, with only one academic performance indicator 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. Perhaps more limiting is that our performance indicator, mathematics performance, only offers 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.

6. Conclusion

Assessing whether tracks affect academic performance for mathematics auring 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 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.

Bibliografie

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: Max-PlanckInstitut 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.

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: 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). School Tracking and Equality of Opportunity. *Economic Policy*, (10), 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: 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.

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., & 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.

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. (2016). 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: 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.

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: 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. (2010). Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature*, *48*(2), 399–423.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Ireson, J., Hallam, S., & Plewis, I. (2001). Ability grouping in secondary schools: Effects on pupils' selfconcepts. *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 attaintment: 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., & Kimmel, 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.

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.

Lavrijsen, J., & Nicaise, I. (2015). New empirical evidence on the effect of educational tracking on social inequalities in reading achievement. *European Educational Research Journal*, *14*(3), 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.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

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.

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?* Paris: Presses Universitaires de France.

Muthén, B., & Muthén, L. (2015). *Mplus Statistical Analysis With Latent Variables User's Guide*. Los Angeles, CA: 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: OECD.

OECD. (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: 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.

Raudenbush, S. W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology*, *52*, 501–525.

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.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In M. Berkane (Ed.), *Latent Variable Modeling and Applications to Causality* (pp. 69–117). New York: Springer.

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*, 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: Wiley.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods*, *13*(4), 279.

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: Plenum.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford: University press.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. London: Sage.

Steiner, P. M., & Cook, T. D. (2012). Matching and propensity scores. In T. D. Litle (Ed.), *Oxford handbook of quantitative methods*. New York: Oxford University Press.

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.

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.

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 ninthgrade mathematics. *Journal of Educational Psychology*, *98*(4), 788.

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.

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, UK: 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.