EFFECTEN VAN ONDERWIJSVORMEN OP SCHOOLSE BETROKKENHEID

Dockx J, De Fraine B. & Van den Branden N.

EFFECTEN VAN ONDERWIJSVORMEN OP SCHOOLSE BETROKKENHEID

Dockx J., De Fraine B. & Van den Branden N.

Promotor: B. De Fraine

Research paper SONO/2019.OL1.1/6

Gent, januari 2019

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. & Van den Branden N. (2019). Effecten van onderwijsvormen op schoolse betrokkenheid. 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

Deze studie onderzoekt of bepaalde onderwijsvormen gekenmerkt worden door een hogere mate van inzet bij de leerlingen dan andere onderwijsvormen. In de wetenschappelijke literatuur wordt deze inzet 'schoolse betrokkenheid' genoemd. Schoolse betrokkenheid omvat een geheel van attitudes en gedragingen die een indicatie zijn van deze onderliggende betrokkenheid. Vaak wordt daarbij een onderscheid gemaakt tussen gedragsmatige betrokkenheid en emotionele betrokkenheid. De internationale literatuur vermeldt een aantal hypotheses over de relatie tussen de verwijzing naar specifieke onderwijsvormen en de betrokkenheid van de leerlingen. Er is echter weinig empirisch onderzoek dat deze hypotheses ook toetst.

In het Vlaams 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. In de praktijk spreken leerlingen, ouders en scholen vaak 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 het keuzegedeelte en de basisopties in de eerste graad 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. In internationaal wetenschappelijk onderzoek wordt het inrichten van verschillende onderwijsvormen *tracking* genoemd, waarbij de onderwijsvormen *tracks* heten. In het Vlaams secundair onderwijs veranderen een groot aantal leerlingen tijdens hun schoolcarriere van onderwijsvorm. Dat zijn leerlingen met zogenaamde 'watervalloopbanen'.

Over hoe onderwijsvormen de schoolse betrokkenheid van leerlingen beïnvloeden wordt vaak de *anti-school cultuur hypothese* vermeld. Deze hypothese, waarvoor een zekere empirische evidentie bestaat, is dat onderwijsvormen meer gekenmerkt worden door een *anti-school cultuur* naarmate ze meer arbeidsmarktgericht zijn (Van Houtte, 2009). De reden hiervoor is dat een doorverwijzing naar een meer arbeidsmarkgerichte onderwijsvorm als een falen wordt ervaren en leerlingen hiernaar handelen. Namelijk, omdat leerlingen het gevoel hebben gefaald te hebben in het onderwijs, zouden ze geen moeite meer willen doen voor een reeds verloren zaak. Op basis hiervan kan voorspeld worden dat meer arbeidsmarktgerichte onderwijsvormen een negatieve invloed zullen hebben op schoolse betrokkenheid. Daarnaast kunnen we ook een alternatieve hypothese beargumenteren. Het is bijvoorbeeld mogelijk dat de doorverwijzing van leerlingen naar studieomgevingen die beter aangepast zijn aan hun interesses en competenties net protectief is voor hun schoolse betrokkenheid. Verder leeft ook de idee dat het veranderen van onderwijsvorm een negatief effect heeft op de schoolse betrokkenheid van leerlingen. Om de *anti-school cultuur* hypothese en het effect van verandering van onderwijsvorm te toetsen stelden we de volgende onderzoeksvragen:

1. Wat is het effect van een onderwijsvorm op de gemiddelde schoolse betrokkenheid van vergelijkbare leerlingen?
2. Is er een effect van onderwijsvormverandering op de gemiddelde schoolse betrokkenheid van vergelijkbare leerlingen?

Voor dit onderzoek gebruiken we de gegevens van het onderzoek 'Loopbanen in het Secundair Onderwijs' (LiSO-project). De substeekproef bestaat uit 5417 leerlingen die in september 2013 startten in het secundair onderwijs. Er waren vier *tracks*: (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 *tracks* 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*' in het eerste jaar als volgt: 1419 leerlingen zaten in KT, 2229 leerlingen zaten in MW, 1033 leerlingen zaten in TO en 736 leerlingen zaten in BV. Veel van deze leerlingen veranderden echter van *track* doorheen het secundair onderwijs. 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), op het einde van eerste leerjaar van de tweede graad (mei 2016) en op het einde van tweede leerjaar van de tweede graad (mei 2017).

Gedragsmatige schoolse betrokkenheid en emotionele schoolse betrokkenheid werden gemeten op elk van deze momenten (zie ook Van den Branden, 2016). De gedragsmatige betrokkenheid en de emotionele betrokkenheid van de leerlingen werd telkens bevraagd met dezelfde items in de leerlingvragenlijst:

Gedragsmatige betrokkenheid	**Emotionele betrokkenheid**
Ik doe erg mijn best om het goed te doen op school.	Wanneer we aan iets werken in de klas, ben ik geïnteresseerd.
Ik werk zo hard als ik kan in de klas.	Ik vind het leuk om in de klas te zijn.
Ik neem actief deel aan het klasgebeuren.	Wanneer ik in de klas ben, voel ik me goed.
Ik luister aandachtig in de klas	Ik vind het fijn om nieuwe dingen te leren in de klas.
Ik let op in de klas.	

Om vergelijkbare leerlingen in verschillende *tracks* te vinden gebruikten we *marginal structural mean models*. Deze methode is gericht op het schatten van effecten van behandelingen (in dit onderzoek is dat de toewijzing aan een bepaalde *track*), waarbij personen van behandeling kunnen veranderen (in dit onderzoek zijn dat leerlingen die van *track* veranderen). Deze methode staat toe om onvertekende effecten te schatten wanneer er voldoende over de achtergrond van de leerlingen gekend is. De achtergrond van leerlingen werd beschreven op basis van schoolse prestaties, sociaaleconomische achtergrond en psychosociale variabelen die gemeten waren in september 2013. Ook werd er rekening gehouden met het verschil in de evolutie in schoolse prestaties en non-cognitieve uitkomsten voor leerlingen die van *track* veranderen en leerlingen die in dezelfde *track* blijven. Bij het gebruik van deze methode bleek dat er enkel (voldoende) vergelijkbare leerlingen waren tussen bepaalde *tracks*. We moesten ons dus beperken tot paarsgewijze vergelijkingen tussen de *tracks*: KT wordt daarom vergeleken met 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. Om dezelfde reden was het enkel mogelijk om de effecten van verandering van *track* te onderzoeken voor leerlingen die eenmaal van een zogenaamde hogere *track* naar een zogenaamde lagere *track* veranderen.

Voor de eerste onderzoeksvraag vinden we voor vergelijkbare leerlingen in verschillende *tracks* dat er geen effecten zijn van tracks op gedragsmatige en emotionele schoolse betrokkenheid. Voor de tweede onderzoeksvraag komen we tot een gelijkaardige bevinding. Er is namelijk doorgaans geen noemenswaardig effect van het veranderen van onderwijsvorm op schoolse betrokkenheid. Enkel bij de KT met MW vergelijking is er een klein negatief effect voor emotionele schoolse betrokkenheid wanneer leerlingen van KT naar MW veranderen.

Een sterk punt van dit onderzoek is dat met de gebruikte methode konden nagaan hoe vergelijkbare leerlingen zouden presteren als ze in een andere *track* zouden zitten. Met deze methode konden we de effecten van eenmalige *track*verandering ook onderzoeken. 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. Ook tussen leerlingen die eenmaal van track veranderen en leerlingen die in hun track blijven vonden we steeds voldoende vergelijkbare leerlingen.

We concluderen dat leerlingen niet beïnvloed worden in hun schoolse betrokkenheid door de onderwijsvorm waarin ze zitten. De *anti-school cultuur* hypothese voor schoolse betrokkenheid wordt dus verworpen.

Inhoud

1 Introduction

In most education systems tracking is used during secondary education (OECD, 2012, pp. 57-58), placing students into different groups, called tracks, according to ability and interest (e.g., Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006; Van de Werfhorst & Mijs, 2010). The assignment of students into separate tracks has been linked to increased inequality in academic performance (e.g., Van de Werfhorst & Mijs, 2010). An often mentioned explanation is that in lower tracks an anti-school culture exists, which limits effort and involvement, and consequently limits learning progress in lower tracks (e.g., Carbonaro, 2005, p. 4; Van Houtte, 2016, pp. 874-875).

This study tests the anti-school culture hypothesis by assessing if being assigned to a lower track negatively affects school engagement. Inverse probability treatment weights with marginal structural mean models were used to control for differences in student intake between tracks and track changes across school years. A Flemish dataset was used for describing the development of school engagement during the first four years of secondary education. In the following sections, we describe literature on tracking, anti-school culture and school engagement in more detail.

1.1 Tracking

Tracking allows instructional practices to be tailored to the specific abilities and needs of more homogeneous student groups (e.g., Hanushek & Wößmann, 2006). Furthermore, it allows for the development of specialized skillsets that are required by the labor market (e.g., Kelly & Price, 2011; Shavit & Müller, 2000; Van de Werfhorst & Mijs, 2010). Hence, policymakers have several reasons to implement tracking at some time point during secondary education.

Two types of tracking systems are often distinguished, the early tracking system and the late tracking system (e.g., Van de Werfhorst & Mijs, 2010). The early tracking system places students in different tracks at a young age (i.e., 12 years or younger), whereas the late tracking system places students in different tracks at an older age (i.e., 14 years or older, Brunello & Rocco, 2017; Schütz, Ursprung, & Wößmann, 2008; Shavit & Müller, 2000). Usually, tracks are hierarchically organized, which leads to a distinction between higher and lower tracks, and is based on the average academic performance of students. Most studies comparing education systems have shown that education systems with early tracking increase inequality in academic performance between students, when compared to education systems with late tracking (e.g., Ammermüller, 2005; Bauer & Riphahn, 2006; Brunello & Checchi, 2007; Hanushek & Wößmann, 2006; Horn, 2009; Lavrijsen & Nicaise, 2015; Marks, 2005; Schütz et al., 2008; Wößmann, 2008), but not all (e.g., Waldinger, 2007). Studies that investigated educational reform of tracking practices also found that early tracking increases inequality in academic performance (Hall, 2012; Jakubowski, Patrinos, Porta, & Wiśniewski, 2016; Kerr, Pekkarinen, & Uusitalo, 2013; Malamud & Pop-Eleches, 2011; Piopiunik, 2014). Hence, early tracking increases inequality in academic performance.

The larger inequality in academic performance in early tracking systems is typically explained by students being allocated to tracks based on their academic performance and their socioeconomic background (e.g., Erikson, Goldthorpe, Jackson, Yaish, & Cox, 2005; Jackson et al., 2007; Kloosterman, Ruiter, De Graaf, & Kraaykamp, 2009). Accordingly, studies show that students with lower academic abilities and disadvantageous socioeconomic backgrounds are placed in lower tracks, which negatively affects academic performance (e.g., Becker, Lüdtke, Trautwein, Köller, & Baumert, 2012; Guill, Lüdtke, & Köller, 2017; Gustafsson, 2008; Korthals & Dronkers, 2016; Retelsdorf, Becker, Köller, & Möller, 2012). Therefore, increased inequality in academic performance is explained by the separation of student across lower and higher tracks, with lower track allocation negatively affecting academic performance. However, the exact mechanisms that lead to lower academic performance in lower tracks remain unclear.

1.2 Anti-school culture

One of the main hypotheses that explains why lower tracks limit learning progress is that lower tracks have an anti-school culture. While there is no generally agreed upon definition of an anti-school culture, it is often described as students being opposed to the school system (e.g., Carbonaro, 2005, p. 4; Van Houtte, 2016, pp. 874-875). This is reflected in lower track students showing less effort and setting less ambitious academic goals, which limits learning progress.

The existence of an anti-school culture in lower tracks has typically been explained by using the differentiation-polarization theory (Hargreaves, 1967; Lacey, 1966). This theory states that students experience being assigned to a lower track as a failure, for this indicates that they did not meet the academic demands of the higher tracks. Accordingly, research shows that students are aware of the status of their student group (e.g., Abraham, 2008; Susan Hallam & Ireson, 2006, 2007) and consider going to a lower track as a negative choice (Ainsworth & Roscigno, 2005). Because educational attainment can no longer give status to these students, they start looking for alternatives sources of status (Van Houtte, 2006a), which results in an anti-school culture.

Related to the differentiation-polarization theory is how students appraise the future rewards education can give (Van Houtte, 2006a). Lower tracks typically prepare students for less esteemed and lower wage jobs (e.g., Brunello & Rocco, 2017; Schütz et al., 2008; Shavit & Müller, 2000), whereas higher tracks offer more pathways to higher education (e.g., Wang & Eccles, 2012; Wolbers, 2007). Hence, lower tracks students expect fewer future rewards because of education. Van Houtte (2016) indeed showed that lower track students have a higher sense of futility, the feeling of no control over success in education. In sum, an anti-school culture results from lower track students expecting fewer rewards from education.

Alternative explanations for an anti-school culture can be given as well. For example, lower tracks are more likely to attract students from disenfranchised backgrounds who may already have developed anti-school attitudes prior to track allocation. This higher prevalence of anti-school attitudes may create a compositional effect that affects all lower track students (Müller & Hofmann, 2016). Another alternative explanation is that teachers in lower tracks have lower expectations of students and set less ambitious goals, and students will act accordingly (Abraham, 2007; Van Houtte, 2006b). Hence, this may lead to an anti-school culture as well.

In sum, an anti-school culture is often used to explain how lower tracks limit the learning progress of students. However, there are different theories on why an anti-school culture would exist in these tracks.

1.3 School engagement

The main challenge in assessing whether students have been affected by an anti-school culture is how that lack of effort and involvement should be operationalized. There is no consensus in the literature on what measure should be used. However, in educational research the concept of school engagement is typically used to describe a student's effort and involvement. In the following paragraphs we explore this concept.

School engagement has been described as a range of attitudes and behaviors that reflect the level of involvement and effort of students in school. Engaged students seek and participate in those activities that make them successful in their learning activities (e.g., Fredricks, 2011; Johnson et al., 2001; Skinner et al., 2008, 2009). Fredricks, Blumenfeld, & Paris (2004) define school engagement as a multidimensional concept that consists of three related components: a behavioral, an emotional and a cognitive component. Behavioral school engagement refers to the level of positive behaviors and active involvement in the learning activities. Emotional school engagement describes the student's positive emotions and attitudes towards teachers, classmates and school in general. Cognitive school engagement describes a student's level of conscious investment in the learning activities (e.g., Fredricks, 2011; Fredricks et al., 2004; Meece, Blumenfeld, & Hoyle, 1988; Skinner et al., 2009; Turner et al., 2002).

Note that a form of negative school engagement can also be described, which is called disaffection (Skinner et al., 2009) or disengagement (Eccles, 2004). It encompasses a range of negative attitudes and behaviors, such as lack of initiation, lack of effort, mental withdrawal and giving up. It is also multidimensional with a behavioral, emotional and cognitive component. However, discussion remains whether school engagement and disaffection are on a single continuum or (partially) represent different concepts (e.g., Jimerson, Campos, & Greif, 2003).

Research has shown that school engagement outcomes generally decline throughout students' school careers, with the steepest declines reported during secondary education (e.g., Skinner & Pitzer, 2012; Wylie & Hodgen, 2012). This decline has been attributed to a mismatch between the developmental needs of adolescence and the educational environment (e.g., Christenson, Reschly, & Wylie, 2012; Eccles et al., 1993). Furthermore, studies have shown that during secondary education school engagement predicts both academic performance and dropout (e.g., Archambault, Janosz, Fallu, & Pagani, 2009; Fall & Roberts, 2012; Green et al., 2012; Martin, Anderson, Bobis, Way, & Vellar, 2012; Reschly, Huebner, Appleton, & Antaramian, 2008).

1.4 This study

The goal of this study was to investigate if being allocated to a lower track negatively affects effort and involvement within the Flemish education system. School engagement measures were used for assessing the level of effort and involvement. This research question derived from studies on tracking that often mention that the anti-school culture in lower tracks negatively affects effort and involvement, which in its turn negatively affects academic performance. Accordingly, our hypothesis was that lower track allocation negatively affects school engagement.

Flemish education is an early tracking system where students choose freely between four tracks at age 12 (OECD, 2012, p. 57). These tracks are hierarchically organized by average academic abilities of students (Van Houtte, 2004). From high to low these are the classical track, the modern track, the technical track and the vocational track. The classical and modern track mainly provide pathways to tertiary education, but they do not prepare for a transition to the labor market. The technical track offers pathways to both tertiary education and a direct transition labor market. The vocational track primarily prepares for a direct transition to the labor market. Tracks are organized at the class-level, and all classmates are in the same track.

While track choice is free, students who do not meet the needs of higher track at the end of school year will be forced to a lower track. This change of a higher track to a lower track in the hierarchy is often described as the "educational waterfall" (Boone & Van Houtte, 2013). Accordingly, many students finish their secondary education in a lower track than their initial track choice. Because the classical, modern and technical do not officially exist in the first stage (first and second year) of secondary education, it is difficult to assess how many students change track in the Flemish student population. However, the vocational track is already distinct from the other tracks at the start of secondary education. Accordingly, the statistical yearbook of Flemish education of the school year 2017-2018 reports that during the first stage of secondary education 14.69% of students are in the vocational track, whereas during the third stage of secondary education (fifth and sixth year) 28.90% of students are in the vocational track. In short, the percentage of students in the vocational track doubles from the first stage to the third stage.The phenomenon of students changing track is not unique to Flanders, for it also happens in Germany (Becker, Neumann, & Dumont, 2016, p. 12) and the Netherlands (OECD, 2016a, p. 16, p. 71).

Any estimated effect of being allocated to a lower track would be biased if we did not account for tracks attracting students with different background characteristics. On average, higher tracks attract students with higher academic performance and more advantageous socioeconomic backgrounds. Moreover, many students change from a higher to lower track over time in Flemish education. Consequently, we used marginal structural mean models (MSMMs), a method which uses inverse probability weighting (IPW) to reweight datasets so no differences prior to track allocation can bias the effect. Furthermore, this method accounted for the track changes that occur over time.

2 Methodology

2.1 Sample

We used data from a study that follows a cohort of 6328 students in 48 schools (LiSO-project, 2018). A subsample was taken where students from de-tracked schools, students in a sports or arts program and students who were redoing their first year in secondary education were removed. Hence, our final subsample consisted of 5417 students. At the start of secondary education, 1419 students were in the classical track. At the end of the fourth year, 620 of those students remained in this track. At the start of secondary education, 2229 students were in the modern track. At the end of the fourth year, 1079 of those students remained in that track. At the start of secondary education, 1033 students were in the technical track. At the end of the fourth, year 590 of those students remained in that track. At the start of secondary education, 736 students were in vocational track. At the end of the fourth year, 625 of those students remained in that track.

There were five measurement occasions, as shown in Figure 1: the start of secondary education September 2013 (T0), the end of the first year of secondary education May 2014 (T1), the end of the second year of secondary education May 2015 (T2), the end of the third year of secondary education May 2016 (T3), and the end of the fourth year of secondary education May 2017 (T4).

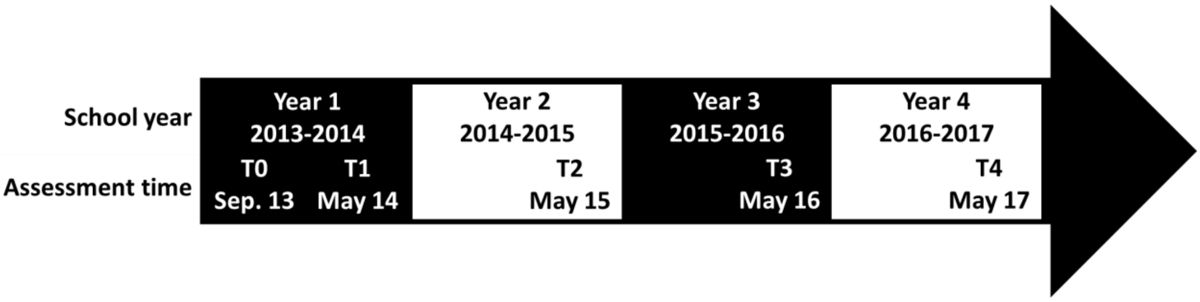

Figure 1. Timeline assessment occasions and school years.

2.2 Variables

2.2.1 Treatment variable

The treatment variable was track allocation (Z_t). Pairwise comparisons were made between tracks that are consecutive in the hierarchy of tracks. It was not possible to compare nonconsecutive tracks, due to the absence of comparable students (this will be further explained in the section 'Weighting procedure'). Lower track allocation was the active treatment condition (Z_t = 1), whereas higher track allocation was the control condition (Z_t = 0). 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. For each comparison of two tracks five track allocation histories were distinguished: staying in the higher track continuously (0,0,0,0), starting in the higher track but changing to the lower track after T3 (0,0,0,1), starting in the higher track but changing to the lower track after T2 (0,0,1,1), starting in the higher track but changing to the lower track after T1 (0,1,1,1), and staying in the lower track continuously (1,1,1,1).

2.2.2 Outcomes

School engagement was operationalized with a measure for behavioral school engagement and a measure for emotional school engagement. Multiple group factor analyses were used to investigate measurement invariance across measurement occasions (Baumgartner & Steenkamp, 2006; Cheung & Rensvold, 2002). This implied assessing whether factor loadings and means of the indicators could be held equal over time and across tracks with sufficient model fit. The cutoff criteria from Hu and Bentler's (1999) were used for fit indices CFI, TLI and RMSEA.

Behavioral school engagement was measured with five items, and emotional school engagement was measured with four items. Both measures were part of a student questionnaire which was conducted at T0, T1, T2, T3 and T4. The items were Dutch translations of the items of an English measure (Skinner et al., 2009). Factor analyses showed that a one-factor structure with assumed measurement invariance fitted well for behavioral school engagement. No item loading or intercept had to be unconstrained (CFI= .97, TLI = .97, RMSEA = .06). For emotional school engagement one item intercept had to be unconstrained at T0 to achieve satisfactory model fit (CFI = .98, TLI = .98, RMSEA = .05). Composite reliabilities ranged from 0.78 to 0.85 for behavioral school engagement, and 0.78 to 0.80 for emotional school engagement. Maximum a posteriori (MAP) estimation was used for estimating student factor scores with a zero mean and unit variance.

2.2.3 Independent variables

To procure unbiased estimates of track effects we needed to account for all differences between students that predict both track change and the outcome at each time point. In literature on MSMMs this is referred to as the sequential conditional exchangeability assumption (Robins & Hernán, 2009). Most authors agree that all variables that predict both the treatment (i.e., track allocation and track change in this study) and the outcome (i.e., school engagement in this study) should be included. If sample size allows it, all variables related to the outcome should also be included (e.g., Brookhart et al., 2006; Myers et al., 2011; Stuart, 2010). Hence, we included those variables that predict school engagement, the initial track allocation and track changes. We used 25 variables, which are discussed in the following paragraphs and shown in Table 1.

Table 1

Descriptions, information sources and properties of time-varying and time-fixed confounder measures

Variable	Description	Info	T0	T1	T2	T3	T4
Mathematics	IRT-score achievement in mathematics	AT	X	X	X	X	
Dutch	IRT-score achievement in Dutch reading comprehension	AT	X				
French	IRT-score achievement in French	AT	X				
Gender	Indicator for boy	OR	X				
Age	Indicator whether student is older than normally progressing	OR	X				
SES	Factor score socioeconomic status	PQ	X				
Allowance	Indicator whether family has an allowance due to low income	OR	X				
Education mother	Indicator whether mother is lowly educated	OR	X				
Other language	Indicator whether home language is not Dutch	OR	X				
Global self-concept	Factor score global academic self-concept	SQ	X	X	X	X	
Self-concept mathematics	Factor score self-concept mathematics	SQ	X	X	X	X	
Self-concept Dutch	Factor score self-concept Dutch	SQ	X	X	X	X	
Self-concept French	Factor score self-concept French	SQ	X	X	X		
Wellbeing	Factor score wellbeing	SQ	X	X	X	X	
Mindset	Factor score mindset	SQ	X	X	X	X	
Autonomous motivation	Factor score autonomous motivation	SQ	X	X	X	X	
Controlled motivation	Factor score controlled motivation	SQ	X	X	X	X	
Behavioral engagement	Factor score behavioral engagement	SQ	X	X	X	X	X
Emotional engagement	Factor score emotional engagement	SQ	X	X	X	X	X
Behavioral disengagement	Factor score behavioral disengagement	SQ	X	X	X	X	
Emotional disengagement	Factor score emotional disengagement	SQ	X	X	X	X	
Interest mathematics	Sum score interest in mathematics	SQ	X	X	X	X	
Interest Dutch	Sum score interest in Dutch	SQ	X	X	X	X	
Interest French	Sum score interest in French	SQ	X	X	X		

Variable	Description	Info	T0	T1	T2	T3	T4
Interest technology	Sum score interest in technology	SQ	X	X	X	X	

Note: T0, T1, T2, T3, T4 = measured at T0, T1, T2, T3, T4; AT = achievement test; OR = official records; PQ = parent questionnaire; SQ = student questionnaire

2.2.3.1 Academic performance

Three measures of academic performance were used: a mathematics test, a Dutch reading comprehension test and a French test. Each test consisted of multiple-choice items and open questions. Item Response Theory (IRT) with 2-parameter and 3-parameter models was used. Warm's weighted likelihood estimation (Warm, 1989) was used for estimating student ability scores. Mathematics tests were conducted at T0, T1, T2 and T3, with the Cronbach's Alphas ranging from 0.83 to 0.87. A Dutch reading comprehension test was conducted at T0 with a Cronbach's Alpha of 0.82. A French test was conducted at T0, with a Cronbach's Alpha of 0.79.

2.2.3.2 Biographical background

Two variables from governmental records were used. The first variable indicates whether the student is a boy. The second variable indicates whether the student is older than normally progressing students.

2.2.3.3 Socioeconomic background

Socioeconomic status was modeled with a Nominal Response Model (Bock, 1972) based on seven indicators: (1 & 2) diploma father and diploma mother, (3 & 4) employment status father and employment status mother, (5 & 6) occupational level father and occupational level mother and (7) family income. These indicators were acquired from a parent questionnaire conducted at T0. Expected A Posteriori (EAP) estimation was used for estimating factor scores. Socioeconomic status had an empirical reliability of 0.87.

Three variables from governmental records were also used. The first variable indicates whether a student's family is eligible for a governmental financial grant due low family income. The second variable indicates whether the mother has not attained a diploma of secondary education. The third variable indicates whether the student does not speak Dutch with more than one family member.

2.2.3.4 Psychosocial development

For psychosocial development repeated measurements of engagement, disengagement, academic self-concept, motivation, mindset, school wellbeing and interest were conducted at T0, T1, T2 and T3 with a student questionnaire. Note that the aforementioned school engagement measures, which were also the outcomes, were also part of the student questionnaire All these measures were on a five-point response scale, which ranged from "not true" to "true". Most were Dutch translations of English measures. MAP estimation was used for estimating student factor scores.

Behavioral and emotional school engagement were already discussed in the outcomes section. Behavioral disengagement and emotional disengagement were also assessed. These originated from the same English measure (Skinner et al., 2009). Academic self-concept was assessed with measures for global academic self-concept and domain-specific self-concepts for mathematics, Dutch and French. These came from the Self-Description Questionnaire III (Marsh & O'Neill, 1984) and shortened Self-Description Questionnaire II (Marsh, Ellis, Parada, Richards, & Heubeck, 2005). Student motivation was assessed with measures for controlled motivation and autonomous motivation. These came from the Academic Self-Regulation Questionnaire (SQR-A, Ryan & Connell, 1989). Mindset was assessed with the Theory of Intelligence Scale (Dweck, Chiu, & Hong, 1995). School wellbeing was assessed with a measure of Dutch origin (Vorst & Smits, 2010). Interest was assessed for mathematics, Dutch, French and technology (Denissen, Zarrett, & Eccles, 2007).

2.3 Marginal structural mean model

2.3.1 Background

The goal of a marginal structural mean model (e.g., Vandecandelaere, Vansteelandt, De Fraine, & Van Damme, 2016) is to construct comparable groups of respondents (i.e., students) across treatment conditions (i.e., attending the higher track or the lower track) at each time point that a treatment exposure can occur (i.e., track changing between school years). To achieve this, inverse probability treatment weighting (IPTW; Austin, 2011) is used. IPTW's main rationale is that when estimating the treatment probabilities on a set of covariates, the estimated probabilities summarize the pretreatment differences in covariates between treatment conditions (Imbens, 2000, pp. 708; Rosenbaum & Rubin, 1983). By weighting with the inverse of these probabilities each treatment condition is made to resemble the population prior to the treatment exposure. Accordingly, the MSMM estimates the average outcome if the entire population would have been allocated to a certain treatment history. This estimated average outcome is called the marginal mean (Robins & Hernán, 2009).

The application of a MSMM to estimate the marginal means of each treatment history consists of three main steps (e.g., Vandecandelaere, et al., 2016). First, a structural model for the marginal means is defined. Second, weights are estimated. Third, a Horvitz-Thompson type estimator with the weighs is used for estimating the structural model. The differences between the marginal means of the structural model are unbiased if the sequential conditional exchangeability assumption holds (Robins & Hernán, 2009). We describe each of these steps in the following sections.

2.3.2 Structural model

As a first step in applying the MSMM, we linked the marginal mean of each treatment history to a structural model. The marginal mean here is either average behavioral school engagement or average emotional school engagement. For the marginal mean at measurement occasion T4 we have the following equation:

$$\mathrm{E}[Y_4(z_1, z_2, z_3, z_4)] = \beta_0 + \beta_1 z_1 z_2 z_3 z_4 + \beta_2 (1 - z_1) z_2 z_3 z_4 + \beta_3 (1 - z_1)(1 - z_2) z_3 z_4 + \beta_4 (1 - z_1)(1 - z_2)(1 - z_3) z_4 \quad (1)$$

In this equation $\mathrm{E}[Y_4(z_1, z_2, z_3, z_4)]$ is the marginal mean, and parameters β_1, β_2, β_3 and β_4 respectively describe the average treatment effects of track allocation histories (1,1,1,1), (0,1,1,1), (0,0,1,1) and (0,0,0,1). β_0 is therefore equal to the track allocation history of always being in the higher track (0,0,0,0). Note that for the marginal means at measurement occasions T1, T2 and T3 equivalent structural models were specified.

2.3.3 Weighting procedure

The second step in applying the MSMM was the estimation of weights in order to remove pretreatment differences across treatment conditions. There were four time points of treatment exposure: T1, T2, T3 and T4. Accordingly, for each time point a time-specific weight was estimated.

The total weight at T4, $\overline{SW}_4$, was the product of time-specific weights, which are described in the following equation:

$$\overline{SW}_4 = SW_1 * SW_2 * SW_3 * SW_4 =$$

$$\frac{\mathrm{P}[Z_1 = 1]}{\mathrm{P}[Z_1 = 1|\boldsymbol{x}_0]} * \frac{\mathrm{P}[Z_2 = 1|Z_1 = 0]}{\mathrm{P}[Z_2 = 1|\boldsymbol{x}_0, \boldsymbol{l}_1, Z_1 = 0]} * \frac{\mathrm{P}[Z_3 = 1|Z_2 = 0]}{\mathrm{P}[Z_3 = 1|\boldsymbol{x}_0, \boldsymbol{l}_2, Z_2 = 0]}$$

$$* \frac{\mathrm{P}[Z_4 = 1|Z_3 = 0]}{\mathrm{P}[Z_4 = 1|x_0, l_3, Z_3 = 0]} \quad (5)$$

In this equation $\overline{SW}_4$ was the total weight at T4, and SW_1, SW_2, SW_3 and SW_4 were the time-specific stabilized weights. Each time-specific weight was estimated based on the baseline measure of the covariates measured at T0 ($\boldsymbol{x}_0$) and the repeated measures of the time-varying covariates (l_{t-1}) measured prior to the track change. When a time-specific weight was not estimated for a student (i.e., student was already in the lower track), it was replaced by value one (i.e., the total weight is unchanged). Note that for the total weights $\overline{SW}_1$, $\overline{SW}_2$ and $\overline{SW}_3$ at T1, T2 and T3 respectively, it was only necessary to multiply time-specific weights until that measurement occasion.

Before estimating weights, we assessed the area of common support between the different track allocation histories of a track comparison. This was achieved by estimating the propensity score of being continuously in the higher track. The overlap in the resulting propensity scores were then used to assess the area of common support (Steiner & Cook, 2013). To prevent extreme weights, students with probabilities close to either one or zero were removed from the sample prior to estimating weights. We used a minimum of 0.05 and a maximum of 0.95 as cutoff values (e.g., Crump, Hotz, Imbens, & Mitnik, 2009). This step will be shown first in the results section.

Furthermore, to attain stable weights estimates, we chose not to use the entire history of time-varying covariates but only their values at T0 (i.e., the baseline measures) and their values directly prior to the track allocation. This appeared to be enough to balance the entire covariate history and stabilize the weight estimation. To estimate the propensity score we used generalized boosted regression models (GBMs, McCaffrey, Ridgeway, & Morral, 2004), which is considered best practice for propensity score estimation (e.g., Stuart, 2010). However, for the technical and vocational track comparison we used the covariates as linear predictors, for this lead to a better balance. We also used 99th percentile truncation (e.g., Lee, Lessler, & Stuart, 2011). The twang 1.5 package (Ridgeway, McCaffrey, Morral, Griffin, & Burgette, 2017) in R 3.4.3 was used for weight estimation.

We assessed balance in the covariates after applying the weights with standardized mean differences of covariates (SMDs). The SMD is the difference between two covariate means of track allocation histories, divided by the pooled *SD* of both track allocation histories (Rubin, 2001). The SMDs were assessed before and after applying weights. Mean SMDs should be no higher than 0.05, whereas SMDs of specific covariates as a rule of thumb should not exceed 0.25 (Caliendo & Kopeinig, 2008).

2.3.4 Estimation

The third step in applying the MSMM was fitting the structural models as a generalized linear model. The model is estimated with the time-varying outcome on the left hand-side, and the combination of treatment-history indicators with time point indicators on the right-hand side. We used GEEs with the Newton-Raphson algorithm and specified an independent correlation matrix (Liang & Zeger, 1986) with sandwich standard errors (Joffe, Ten Have, Feldman, & Kimmel, 2004). Using a non-independent correlation structure or multilevel models would bias results by making outcome measurements dependent on future treatment exposures (Robins, Hernan, & Brumback, 2000, p. 554). To examine differences between track allocation histories at each time point, contrasts were tested using one degree of freedom Wald tests (Kuhn, Weston, Wing, & Forester, 2016). The geepack 1.2-1 package in R 3.4.3 was used for estimation (Højsgaard, Halekoh, & Yan, 2006). Cohen's *d* was used for effect size interpretation (Cohen, 1977).

2.4 Missing values

The rates of missingness for the track allocation histories of interest were 2.98%, 14.17%, 8.18%, 12.86%, and 15.87% (only outcomes) at T0, T1, T2, T3 and T4 respectively. We used multiple imputation by chained equations for missing values in the outcomes and covariates (Schafer & Graham, 2002) with the package mice 2.30 (van Buuren & Groothuis-Oudshoorn, 2011) in R 3.4.3. This should result in unbiased and efficient estimates under the missing at random assumption (MAR, e.g., Moodie, Delaney, Lefebvre, & Platt, 2008). We estimated ten imputed datasets, and combined the results as described by Rubin's (1987) rules. The average relative efficiencies for contrasts at T4 was 99.12% for the classical and modern track comparison, 98.23% for the modern and technical track comparison, and 97.59% for the technical and vocational track comparison.

There were also students who did not have a track allocation history as described in the 'Treatment variable' section, but at some time point went to a sports program, arts program, special method program, special education, changed track multiple times or changed to a track not part of the comparison. Simply removing these students from the analysis could bias results. Therefore, these students were included in the analysis up until the time point they went to an alternative track allocation history. From that time point onwards, they were considered censored. Censoring weights for these students were estimated just as IPTW for the MSMM, but now the probability of being censored was estimated. The final weights used in the analysis were a product of the IPTW weights and censoring weights.

3 Results

3.1 Balance after weighting

Figure 2 shows the overlap in logit propensity scores of higher track allocation between track allocation histories of each comparison. Substantial overlap exists, but for each comparison very low propensity scores and very higher propensity scores occurred. Accordingly, by applying the cut-off values of .05 and .95, students were removed from the dataset. The resulting sample size for each comparison across time points is shown in Table 2.

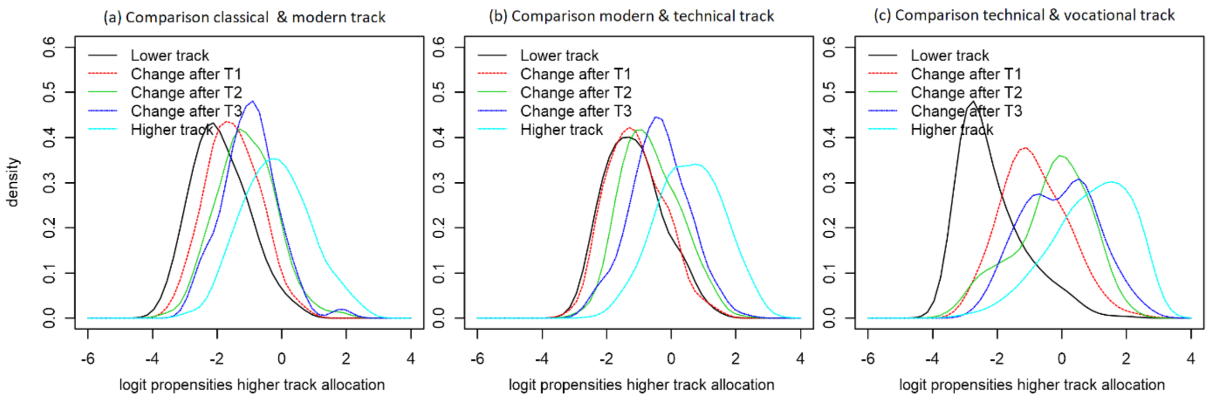

Figure 2. Overlap in logit propensity scores for higher track allocation across treatment histories of each track comparison.

Table 2
Overview treatment histories of classical and modern track comparison, modern and technical track comparison, and technical and vocational track comparison

Treatment history	T1	T2	T3	T4
Classical and modern track comparison				
Classical track continuous (0,0,0,0)	1240	978	673	608
Classical to modern after T3 (0,0,0,1)				61
Classical to modern after T2 (0,0,1,1)			291	281
Classical to modern after T1 (0,1,1,1)		242	223	196
Modern track continuous (1,1,1,1)	955	881	728	652
To other program		94	280	397
Modern and technical track comparison				
Modern track continuous (0,0,0,0)	2182	1780	1304	1078
Modern to technical track T3 (0,0,0,1)				187
Modern to technical track T2 (0,0,1,1)			338	315
Modern to technical track T1 (0,1,1,1)		284	219	192
Technical track continuous (1,1,1,1)	887	691	586	541
To other program		314	622	756
Technical and vocational track comparison				
Technical track continuous (0,0,0,0)	1026	772	645	588
Technical to vocational track T3 (0,0,0,1)				44
Technical to vocational track T2 (0,0,1,1)			89	86
Technical to vocational track T1 (0,1,1,1)		162	143	139
Vocational track continuous (1,1,1,1)	544	497	473	454
To other program		139	220	259

Note: T1 = Number of students first year after removal extreme propensity scores; T2 = Number of students second year after removal extreme propensity scores; T3 = Number of students third year after removal extreme propensity scores; T4 = Number of students fourth year after removal extreme propensity scores

Afterwards, weights were estimated for each track comparison. Figures 3, 4 and 5 show the absolute SMDs for each track comparison. Table 3 shows the mean, minimum, and maximum of all SMDs after applying the weights. The results show a substantial reduction in the pretreatment differences between tracks. For the classical and modern track comparison, general academic self-concept measured directly prior to a track allocation slightly exceeds the .25 cutoff value at T1, T2 and T3. The same is observed for the modern and technical track comparison. However, at T4 there is remaining imbalance in mathematics self-concept and also mathematics performance. For the technical and vocational track comparison, imbalance remains for interest in technology and gender at T2 and T3. At T4 imbalance remains for self-concept in mathematics and mathematics performance.

Table 3
SMDs after weighting

	Classical and modern track comparison				Modern and technical track comparison				Technical and vocational track comparison			
	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4
Average	-.03	-.03	-.03	-.03	-.02	-.06	-.05	-.04	.03	-.06	-.06	-.04
Minimum	-.17	-.30	-.30	-.27	-.15	-.32	-.30	-.32	-.25	-.41	-.35	-.32
Maximum	.09	.18	.18	.19	.11	.18	.19	.20	.18	.12	.10	.26

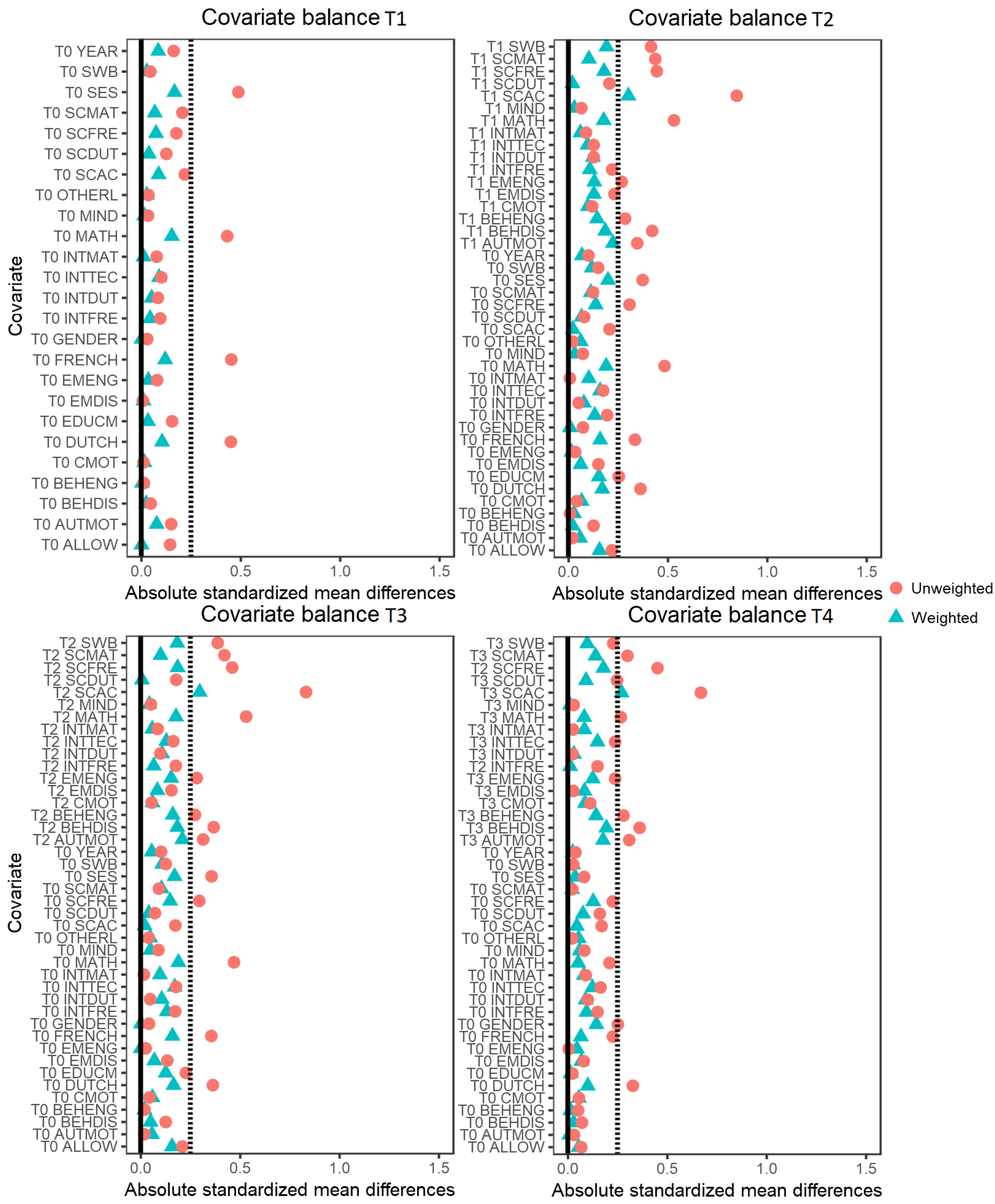

Figure 3. Absolute SMDs classical and modern track comparison at T1, T2, T3 and T4.

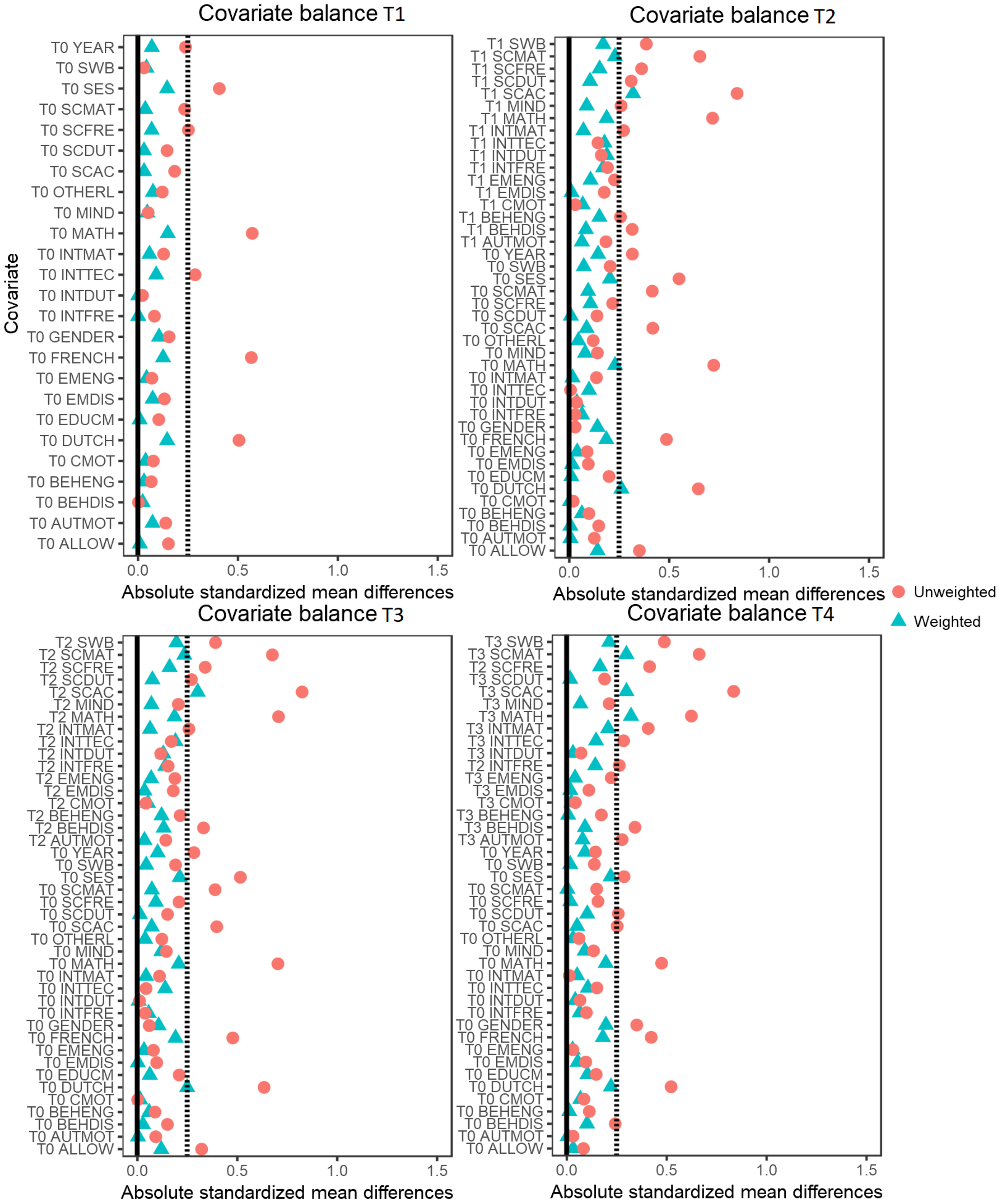

*Figure 4***.** Absolute SMDs modern and technical track comparison at T1, T2, T3 and T4.

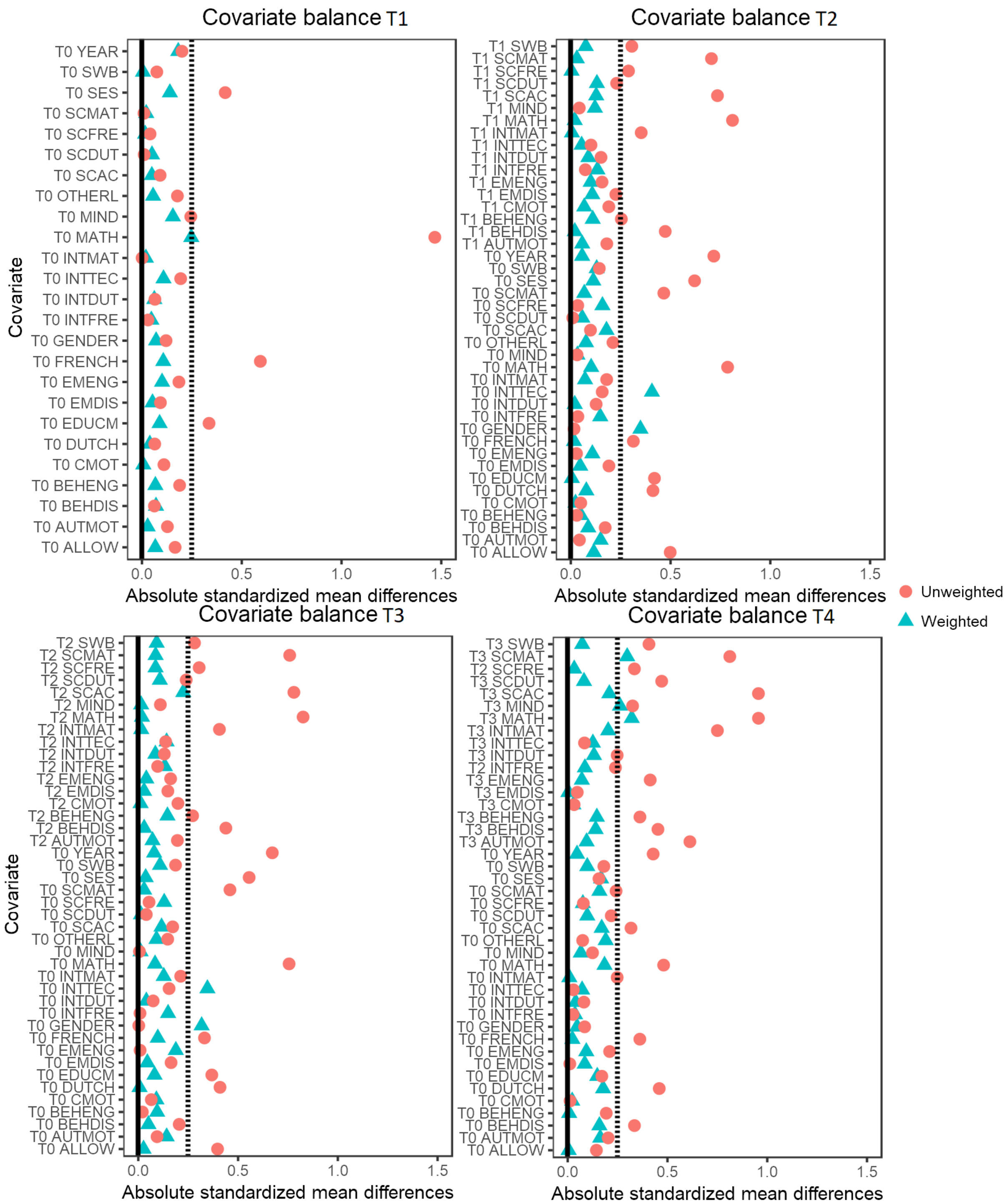

Figure 5. Absolute SMDs technical and vocational track comparison at T1, T2, T3 and T4.

3.2 Analysis of track effects

In the following paragraphs we describe the estimated contrasts across the three track comparisons. For brevity, we limit the description to general trends and the final contrast estimates at T4. Figures 6, 7 and 8 also show the development in behavioral school engagement and emotional school engagement for each track comparison. Note that across all outcomes and across all groups we only observe declines. Accordingly, we will describe whether the different groups of students decline more than the reference group, which is the group of students continuously in the higher track. The estimated contrasts are shown in Table 4, 5 and 6.

3.2.1 Classical and modern track comparison

For behavioral school engagement we see no substantial difference in the development between the group of students continuously in the higher track (classical track) and the group of students continuously in the lower track (modern track). Hence, at T4 the contrast is not significant. The groups of students who changed from the higher to lower track do decline more. However, only the group of students who changed from the higher to lower track after T1 has a significant contrast at T4, but it has a trivial effect size.

For emotional school engagement we see no substantial difference in the development between the group of students continuously in the higher track and the group of students continuously in the lower track. Hence, at T4 the contrast is not significant. The groups of students who changed from the higher to lower track do decline more. Accordingly, the group of students who changed from the higher to lower track after T1 and the group of students who changed from the higher to lower track after T2 have significant contrasts at T4, with small effect sizes.

Table 4
Contrast estimates behavioral engagement and emotional engagement in classical and modern track comparison

	Behavioral engagement		Emotional engagement	
	d	SE	*d*	SE
T1				
High – low	-0.06	0.04	0.07	0.04
T2				
High – low	0.02	0.05	0.05	0.04
High – T1 Change	0.15	0.09	0.21*	0.07
T3				
High – low	-0.01	0.07	0.09	0.05
High – T1 Change	0.27*	0.09	0.28*	0.07
High – T2 Change	0.18*	0.09	0.15*	0.07
T4				
High – low	0.00	0.07	0.06	0.06
High – T1 Change	0.19*	0.09	0.28*	0.08
High – T2 Change	0.17	0.09	0.23*	0.08
High – T3 Change	0.14	0.15	0.19	0.14

Note: *d* = contrast estimate; high = continuously in the higher track; low = continuously in the lower track; T1 change = changed from higher to lower track after T1; T2 change = changed from higher to lower track after T2; T3 change = changed from higher to lower track after T3
* Significant at $\alpha = 0.05$

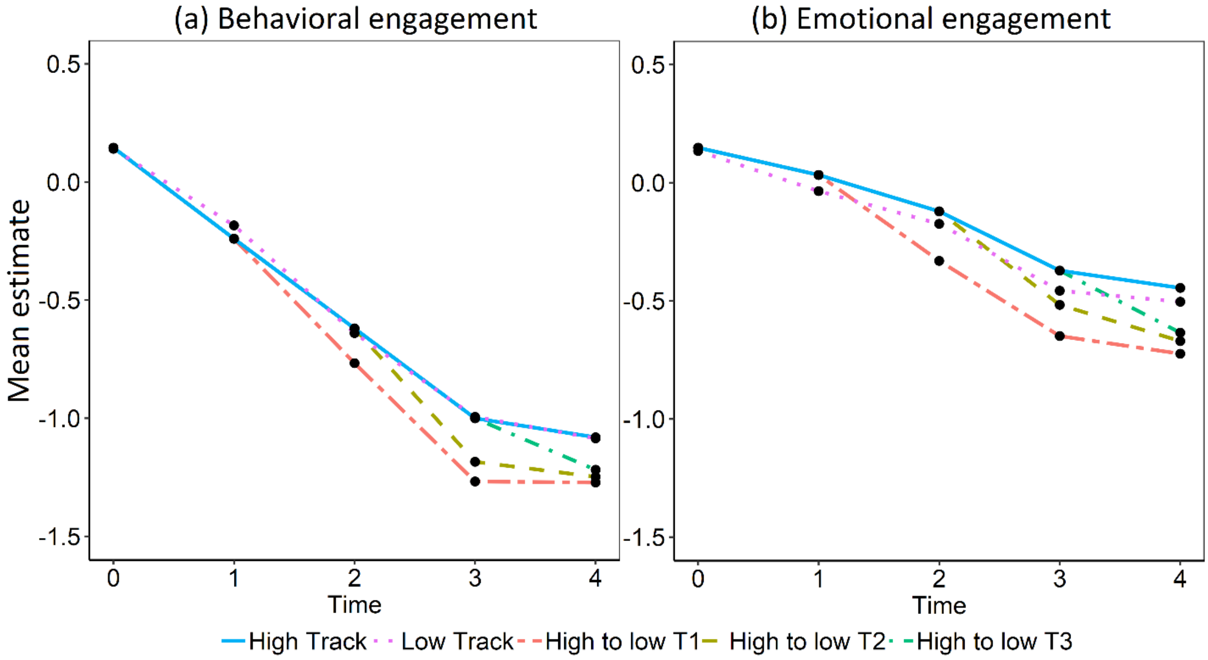

Figure 6. Development behavioral engagement and emotional engagement in classical and modern track comparison.

3.2.2 Modern and technical track comparison

For behavioral school engagement there is no substantial difference in the development between the group of students continuously in the higher track (modern track) and the group of students continuously in the lower track (technical track). Hence, at T4 the contrast is not significant. The groups of students who changed from the higher to lower track do decline more. However, only the group of students who changed from the higher to lower track after T2 has a significant contrast at T4, but it has a trivial effect size.

For emotional school engagement the group of students continuously in the lower track declines slightly more than the group of students continuously in the higher track. However, at T4 the contrast is not significant. Only the group of students who changed from the higher to lower track after T1 decline slightly more. Hence, at T4 their contrast is significant but with a trivial effect size.

Table 5
Contrast estimates behavioral engagement and emotional engagement in modern and technical track comparison

	Behavioral engagement		Emotional engagement	
	d	SE	*d*	SE
T1				
High – low	-0.03	0.04	0.01	0.04
T2				
High – low	-0.03	0.05	0.07	0.04
High – T1 Change	0.20*	0.09	0.08	0.07
T3				
High – low	-0.01	0.06	0.08	0.06
High – T1 Change	0.09	0.10	0.11	0.08

	Behavioral engagement		Emotional engagement	
High – T2 Change	0.08	0.08	-0.04	0.07
T4				
High – low	0.03	0.07	0.12	0.06
High – T1 Change	0.19	0.11	0.18*	0.09
High – T2 Change	0.17*	0.08	0.01	0.08
High – T3 Change	0.07	0.11	0.03	0.09

Note: *d* = contrast estimate; high = continuously in the higher track; low = continuously in the lower track; T1 change = changed from higher to lower track after T1; T2 change = changed from higher to lower track after T2; T3 change = changed from higher to lower track after T3
* Significant at $\alpha = 0.05$

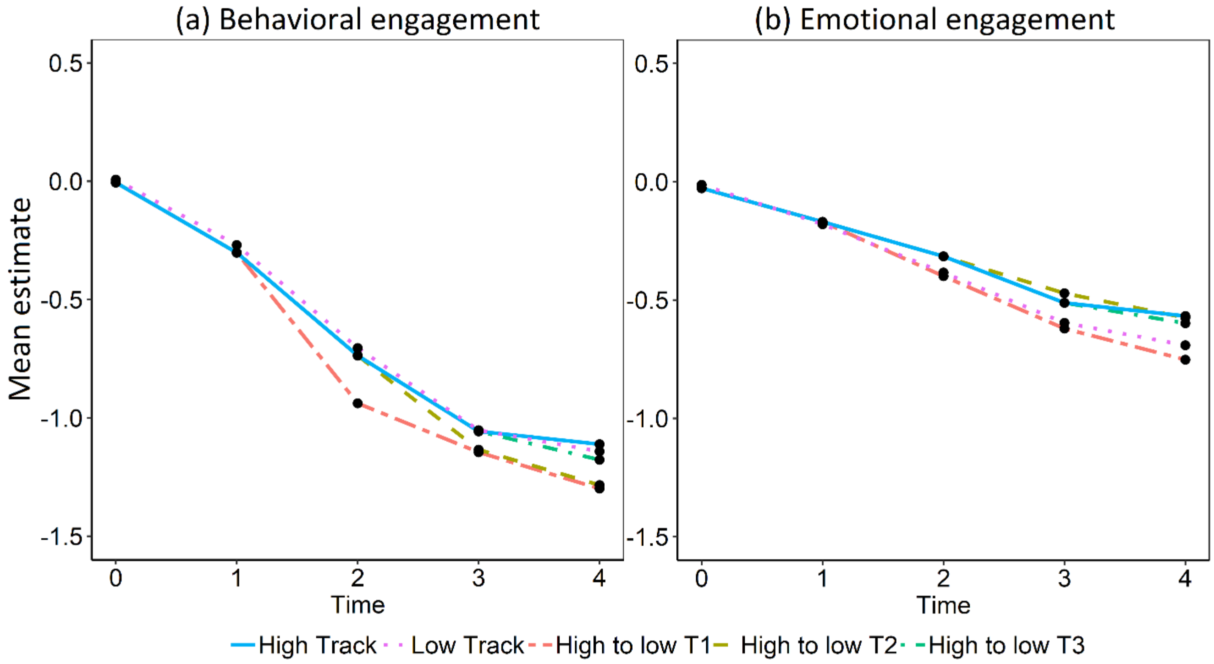

Figure 7. Development behavioral engagement and emotional engagement in modern and technical track comparison.

3.2.3 Technical and vocational track comparison

For behavioral school engagement there are some differences in the development between the group of students continuously in the higher track (technical track) and the group of students continuously in the lower track (vocational track), but at T4 no difference remains. Hence, at T4 the contrast is not significant. The groups of students who changed from the higher to lower track also show some difference in their development, but at T4 no differences remain.

For emotional school engagement there are some differenced in the development between the group of students continuously in the higher track and the group of students continuously in the lower track, but at T4 no difference remains. Accordingly, at T4 the contrast is not significant. The groups of students who changed from the higher to lower track also show some difference in their development, but at T4 no differences remain.

Table 6
Contrast estimates behavioral engagement and emotional engagement in technical and vocational track comparison

	Behavioral engagement		Emotional engagement	
	d	SE	*d*	SE
T1				
High – low	0.07	0.10	-0.01	0.11
T2				
High – low	0.22*	0.10	0.18*	0.08
High – T1 Change	0.38	0.26	0.18	0.23
T3				
High – low	-0.10	0.12	-0.03	0.10
High – T1 Change	0.17	0.15	0.15	0.13
High – T2 Change	-0.31	0.31	-0.30	0.23
T4				
High – low	-0.11	0.13	0.04	0.12
High – T1 Change	-0.16	0.16	-0.10	0.14
High – T2 Change	-0.16	0.16	-0.13	0.13
High – T3 Change	0.05	0.44	0.07	0.24

Note: *d* = contrast estimate; high = continuously in the higher track; low = continuously in the lower track; T1 change = changed from higher to lower track after T1; T2 change = changed from higher to lower track after T2; T3 change = changed from higher to lower track after T3
* Significant at $\alpha = 0.05$

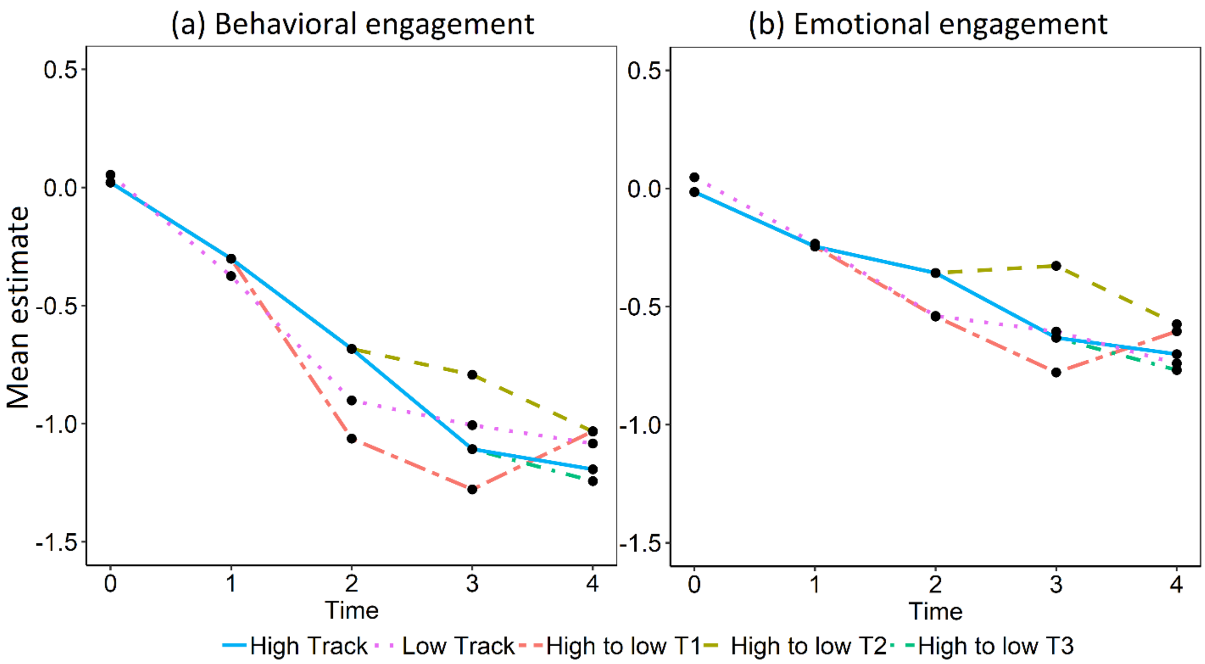

Figure 8. Development behavioral engagement and emotional engagement in technical and vocational track comparison.

3.2.4 Summary

In sum, for the classical and modern track comparison we find that students who changed from the higher to lower track generally decline more in emotional school engagement. There is no substantial difference for behavioral school engagement. Students continuously in the classical track and modern track do not differ for both outcomes. For the modern and technical track comparison we find no substantial differences between student groups. Accordingly, for the technical and vocational track comparison we find no substantial differences between student groups.

4 Discussion

In this study, we investigated whether being assigned to the lower track, either from the start of secondary education or after a track change, affects school engagement. Behavioral and emotional school engagement were compared between four tracks. Students were compared across pairs of tracks that are consecutive in the hierarchy of tracks, which led to three pairwise track comparisons.

We generally do not find support for our hypothesis. Only for one out of three pairwise track comparisons we find evidence that students who start in the higher track but change to the lower track have lower school engagement than comparable students continuously in the higher track. However, there is never evidence that students continuously in the lower track have lower school engagement than comparable students in the higher track. Therefore, the hypothesis is rejected.

This finding is surprising given that an anti-school culture if often used as an explanation why lower tracks negatively affect academic performance (e.g., Abraham, 2008; Müller & Hofmann, 2016; Van Houtte, 2006a). However, it should be noted that we assessed behavioral and emotional school engagement as an operationalization of anti-school norms. Behavioral school engagement describes the level of active involvement in the learning activities and emotional school engagement describes the positive emotions a student has towards teachers, classmates and school in general. Contrarily, Van Houtte and Stevens (2015) measured the sense of futility, which can be defined as not believing it is possible to influence one's own success at school. Hence, a difference in operationalization may be the cause for the rejection of our hypothesis.

However, if anti-school culture cannot explain why lower tracks negatively affect academic performance, alternative explanations become more plausible. For example, higher tracks have more learning opportunities that are both cognitively challenging and supported by adaptive teacher feedback, whereas lower tracks are more characterized by rote memorization and the disciplining of students (e.g., Baumert et al., 2010; Kunter & Baumert, 2006; Retelsdorf, Butler, Streblow, & Schiefele, 2010; Van Houtte, 2004). Accordingly, lower track are considered less academically challenging than higher tracks (e.g., Salmela-Aro, Kiuru, & Nurmi, 2008; Stevens & Vermeersch, 2010). In Germany, teachers were also found to have more pedagogical content knowledge in higher tracks (Baumert et al., 2010; Krauss et al., 2008). Hence, given our results, these alternative explanations of why lower tracks negatively affect academic performance become more plausible.

Note that these findings may only be applicable to Flanders. While earlier, we solely distinguished tracking systems by the age when tracking starts, this is an oversimplification. Many aspects of tracking systems can differ across education systems, such as whether the tracks are strongly specialized or retain general skills development (e.g.,Shavit & Müller, 2000), the number of tracks (e.g., Bol & van de Werfhorst, 2013), whether standardized tests determine assignment of students to tracks (e.g., Bol, Witschge, Van de Werfhorst, & Dronkers, 2014; Tieben, de Graaf, & de Graaf, 2010; Trautwein et al., 2006) and whether tracks are organized on the school-level or class-level (e.g., Trautwein et al., 2006; Van de Werfhorst & Mijs, 2010). Hence, tracking systems can differ substantially, which could affect the effects of a lower track on school engagement.

5 Strengths and limitations

A limitation of this study is that school engagement is only assessed for the behavioral and emotional component. This is a somewhat narrow operationalization of school engagement (e.g., Fredricks, 2011), for the cognitive component and disaffection are not included. However, keen readers will have noted that we have used behavioral disaffection and emotional disaffection as controls (Table 1). These measures were found to be unreliable and had only partial measurement invariance, and we decided not to present their results here. However, we did estimate the effect of lower tack allocation on these disaffection measures at T4, which gave highly similar results as the school engagement measures. In our view, this indicated that the inclusion of disaffection measures would be of too limited value. The lack of a measure for cognitive school engagement is more problematic, given that it can behave differently than emotional and behavioral school engagement (e.g., Wang & Eccles, 2012).

Any estimate in this study is based on the assumption the no covariates remain that predict both lower track allocation and the school engagement (e.g., Rosenbaum & Rubin, 1983). This assumption can never be tested (Steiner & Cook, 2013). We argue though that by accounting for indicators of academic performance, socioeconomic background and psychosocial variables, most bias should have been removed. Furthermore, unobserved variables that are correlated with the accounted for variables (Stuart, 2010) are accounted for as well.

6 Conclusion

In sum, we do not find that the assignment of students to a lower track affects school engagement. In addition, for students who change from the higher to the lower track, we only find limited evidence of a negative effect on school engagement. Therefore, we consider the hypothesis that an anti-school culture causes lower tracks to have a negative effect on academic performance as implausible.

Bibliografie

Abraham, J. (2007). Differentiating between and synthesizing quantitative, qualitative, and longitudinal research on polarized school cultures: a comment on Van Houtte (2006). *Journal of Curriculum Studies*, *39*(5), 597–602. https://doi.org/10.1080/00220270701487624

Abraham, J. (2008). Pupils' perceptions of setting and beyond—a response to Hallam and Ireson. *British Educational Research Journal*, *34*(6), 855–863. https://doi.org/10.1080/01411920802044511

Ainsworth, J. W., & Roscigno, V. J. (2005). Stratification, school-work linkages and vocational education. *Social Forces*, *84*(1), 257–284. https://doi.org/10.1353/sof.2005.0087

Ammermüller, A. (2005). Educational Opportunities and the Role of Institutions. *ZEW Discussion Paper*, 5–44.

Archambault, I., Janosz, M., Fallu, J.-S., & Pagani, L. S. (2009). Student engagement and its relationship with early high school dropout. *Journal of Adolescence*, *32*(3), 651–670. https://doi.org/10.1016/j.adolescence.2008.06.007

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. https://doi.org/10.1002/sim.3150

Bauer, P., & Riphahn, R. T. (2006). Timing of school tracking as a determinant of intergenerational transmission of education. *Economics Letters*, *91*(1), 90–97. https://doi.org/10.1016/j.econlet.2005.11.003

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. https://doi.org/10.3102/0002831209345157

Baumgartner, H., & Steenkamp, J.-B. E. M. (2006). An extended paradigm for measurement analysis of marketing constructs applicable to panel data. *Journal of Marketing Research*, *43*(3), 431–442. https://doi.org/10.1509/jmkr.43.3.431

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. https://doi.org/10.1037/a0027608

Becker, M., Neumann, M., & Dumont, H. (2016). Recent developments in school tracking practices in Germany: An overview and outlook on future trends. *Orbis Scholae*, *10*(3), 9–25. https://doi.org/10.14712/23363177.2017.8

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51. https://doi.org/10.1007/BF02291411

Bol, T., & van de Werfhorst, H. G. (2013). Educational systems and the trade-off between labor larket allocation and equality of educational opportunity. *Comparative Education Review*, *57*(2), 285–

308. https://doi.org/10.1086/669122

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. https://doi.org/10.1093/sf/sou003

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. https://doi.org/10.1080/01425692.2012.704720

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. https://doi.org/10.1093/aje/kwj149

Brunello, G., & Checchi, D. (2007). Does school tracking affect equality of opportunity? New international evidence. *Economic Policy*, *22*(52), 781–861. https://doi.org/10.1111/j.1468-0327.2007.00189.x

Brunello, G., & Rocco, L. (2017). The labor market effects of academic and vocational education over the life cycle: Evidence based on a British cohort. *Journal of Human Capital*, *11*(1), 106–166. https://doi.org/10.1086/690234

Caliendo, M., & Kopeinig, S. (2008). Some practicial guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

Carbonaro, W. (2005). Tracking, students' effort, and academic achievement. *Sociology of Education*, *78*(1), 27–49. https://doi.org/10.1177/003804070507800102

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Christenson, S. L., Reschly, A. L., & Wylie, C. (2012). *Handbook of research on student engagement*. Boston, MA: Springer Science & Business Media.

Cohen, J. (1977). *Statistical power analysis for the behavioural sciences*. New York City, NY: Academic press.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*(1), 187–199. https://doi.org/10.1093/biomet/asn055

Denissen, J. J. A., Zarrett, N. R., & Eccles, J. S. (2007). I like to do it, I'm able, and I know I am: Longitudinal couplings between domain-specific achievement, self-concept, and interest. *Child Development*, *78*(2), 430–447. https://doi.org/10.1111/j.1467-8624.2007.01007.x

Dweck, C. S., Chiu, C., & Hong, Y. (1995). Implicit theories and their role in judgments and reactions: A word from two perspectives. *Psychological Inquiry*, *6*(4), 267–285. https://doi.org/10.1207/s15327965pli0604_1

Eccles, J. S. (2004). Schools, academic motivation, and stage-environment fit. In R. M. Lerner & L. Steinberg (Eds.), *Handbook of adolescent psychology, 2nd edition* (pp. 125–153). Hoboken, NJ: John Wiley & Sons.

Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., & Mac Iver, D. (1993). Development during adolescence: The impact of stage-environment fit on young adolescents’ experiences in schools and in families. *American Psychologist*, *48*(2), 90–101. https://doi.org/10.1037/0003-066X.48.2.90

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. https://doi.org/10.1073/pnas.0502433102

Fall, A.-M., & Roberts, G. (2012). High school dropouts: Interactions between social context, self-perceptions, school engagement, and student dropout. *Journal of Adolescence*, *35*(4), 787–798. https://doi.org/10.1016/j.adolescence.2011.11.004

Fredricks, J. A. (2011). Engagement in school and out-of-school contexts: A multidimensional view of engagement. *Theory into Practice*, *50*(4), 327–335. https://doi.org/10.1080/00405841.2011.607401

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Green, J., Liem, G. A. D., Martin, A. J., Colmar, S., Marsh, H. W., & McInerney, D. (2012). Academic motivation, self-concept, engagement, and performance in high school: Key processes from a longitudinal perspective. *Journal of Adolescence*, *35*(5), 1111–1122. https://doi.org/10.1016/j.adolescence.2012.02.016

Guill, K., Lüdtke, O., & Köller, O. (2017). Academic tracking is related to gains in students’ intelligence over four years: Evidence from a propensity score matching study. *Learning and Instruction*, *47*, 43–52. https://doi.org/10.1016/j.learninstruc.2016.10.001

Gustafsson, J. E. (2008). Schooling and intelligence: Effects of track of study on level and profile of cognitive abilities. In P. C. Kyllonen, R. D. Roberts, & L. Stankov (Eds.), *Extending intelligence: Enhancement and new constructs* (pp. 31–49). London, United Kingdom: Routledge.

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. https://doi.org/10.3368/jhr.47.1.237

Hallam, S., & Ireson, J. (2006). Secondary school pupils’ preferences for different types of structured grouping practices. *British Educational Research Journal*, *32*(4), 583–599. https://doi.org/10.1080/01411920600775274

Hallam, S., & Ireson, J. (2007). Secondary school pupils’ satisfaction with their ability grouping placements. *British Educational Research Journal*, *33*(1), 27–45. https://doi.org/10.1080/01411920601104342

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

Hargreaves, D. H. (1967). *Social Relations in a Secondary School*. London, United Kingdom: Routledge & Kegan Paul.

Højsgaard, S., Halekoh, U., & Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software*, *15*(2), 1–11. https://doi.org/10.18637/jss.v015.i02

Horn, D. (2009). Age of selection counts: a cross-country analysis of educational institutions. *Educational Research and Evaluation*, *15*(4), 343–366. https://doi.org/10.1080/13803610903087011

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706–710. https://doi.org/10.1093/biomet/87.3.706

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. https://doi.org/10.1177/0001699307080926

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. https://doi.org/10.1080/09645292.2016.1149548

Jimerson, S. R., Campos, E., & Greif, J. L. (2003). Toward an understanding of definitions and measures of school engagement and related terms. *The California School Psychologist*, *8*(1), 7–27. https://doi.org/10.1007/BF03340893

Joffe, M. M., Ten Have, T. R., Feldman, H. I., & Kimmel, S. E. (2004). Model selection, confounder control, and marginal structural models. *The American Statistician*, *58*(4), 272–279. https://doi.org/10.1198/000313004X5824

Johnson, M. K., Crosnoe, R., & Elder Jr, G. H. (2001). Students' attachment and academic engagement: The role of race and ethnicity. *Sociology of Education*, 318–340. https://doi.org/10.2307/2673138

Kelly, S., & Price, H. (2011). The correlates of tracking policy: Opportunity hoarding, status competition, or a technical-functional explanation? *American Educational Research Journal*, *48*(3), 560–585. https://doi.org/10.3102/0002831210395927

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

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. https://doi.org/10.1111/j.1468-4446.2009.01235.x

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

Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M., & Jordan, A. (2008). Pedagogical content knowledge and content knowledge of secondary mathematics teachers. *Journal of Educational Psychology*, *100*(3), 716–725. https://doi.org/10.1037/0022-0663.100.3.716

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. https://doi.org/10.1007/s10984-006-9015-7

Lacey, C. (1966). Some sociological concomitants of academic streaming in a grammar school. *The British Journal of Sociology*, *17*(3), 245–262. https://doi.org/10.2307/588773

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. https://doi.org/10.1177/1474904115589039

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS ONE*, *6*(3), 1–6. https://doi.org/10.1371/journal.pone.0018174

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22. https://doi.org/10.1093/biomet/73.1.13

LiSO-project. (2018). Het LiSO-project kort samengevat *[The LiSO-project summarized]*. Retrieved March 30, 2018, from https://lisoproject.be/

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. https://doi.org/j.jpubeco.2011.03.006

Marks, G. N. (2005). Cross-national differences and accounting for social class inequalities in education. *International Sociology*, *20*(4), 483–505. https://doi.org/10.1177/0268580905058328

Marsh, H. W., Ellis, L. A., Parada, R. H., Richards, G., & Heubeck, B. G. (2005). A short version of the Self Description Questionnaire II: Operationalizing criteria for short-form evaluation with new applications of confirmatory factor analyses. *Psychological Assessment*, *17*(1), 81–102. https://doi.org/10.1037/1040-3590.17.1.81

Marsh, H. W., & O'Neill, R. (1984). Self Description Questionnaire III: The construct validity of multidimensional self-concept ratings by late adolescents. *Journal of Educational Measurement*, *21*(2), 153–174. https://doi.org/10.1111/j.1745-3984.1984.tb00227.x

Martin, A. J., Anderson, J., Bobis, J., Way, J., & Vellar, R. (2012). Switching on and switching off in mathematics: An ecological study of future intent and disengagement among middle school students. *Journal of Educational Psychology*, *104*(1), 1–18. https://doi.org/10.1037/a0025988

McCaffrey, D. F., Ridgeway, G., & Morral, A. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*(4), 403–425. https://doi.org/10.1037/1082-989X.9.4.403

Meece, J. L., Blumenfeld, P. C., & Hoyle, R. H. (1988). Students' goal orientations and cognitive engagement in classroom activities. *Journal of Educational Psychology*, *80*(4), 514–523.

Moodie, E. E. M., Delaney, J. A. C., Lefebvre, G., & Platt, R. W. (2008). Missing confounding data in marginal structural models: A comparison of inverse probability weighting and multiple imputation. *The International Journal of Biostatistics*, *4*(1), 1–23. https://doi.org/10.2202/1557-4679.1106

Müller, C. M., & Hofmann, V. (2016). Does being assigned to a low school track negatively affect psychological adjustment? A longitudinal study in the first year of secondary school. *School*

Effectiveness and School Improvement, *27*(2), 95–115. https://doi.org/10.1080/09243453.2014.980277

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. https://doi.org/10.1093/aje/kwr364

OECD. (2012). Equity and Quality in Education. Paris, France: OECD. https://doi.org/10.1787/9789264130852-en

OECD. (2016). *Netherlands 2016: Foundations for the Future, Reviews of National Policies for Education*. Paris, France: OECD Publishing. https://doi.org/10.1787/9789264257658-en

Piopiunik, M. (2014). The effects of early tracking on student performance: Evidence from a school reform in Bavaria. *Economics of Education Review*, *42*, 12–33. https://doi.org/10.1016/j.econedurev.2014.06.002

Reschly, A. L., Huebner, E. S., Appleton, J. J., & Antaramian, S. (2008). Engagement as flourishing: The contribution of positive emotions and coping to adolescents' engagement at school and with learning. *Psychology in the Schools*, *45*(5), 419–431. https://doi.org/10.1002/pits.20306

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. https://doi.org/10.1111/j.2044-8279.2011.02051.x

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. https://doi.org/10.1016/j.learninstruc.2009.01.001

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

Robins, J. M., & Hernán, M. A. (2009). Estimation of the causal effects of time-varying exposures. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 553–599). Boca Raton, FL: Chapman and Hall/CRC.

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*(5), 550–560. https://doi.org/10.1097/00001648-200009000-00011

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York City, NY: Wiley.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*(3–4), 169–188. https://doi.org/10.1023/A:1020363010465

Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: examining reasons for acting in two domains. *Journal of Personality and Social Psychology*, *57*(5), 749–761. https://doi.org/10.1037/0022-3514.57.5.749

Salmela-Aro, K., Kiuru, N., & Nurmi, J.-E. (2008). The role of educational track in adolescents' school burnout: A longitudinal study. *British Journal of Educational Psychology*, *78*(4), 663–689.

https://doi.org/10.1348/000709908X281628

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Schütz, G., Ursprung, H. W., & Wößmann, L. (2008). Education policy and equality of opportunity. *Kyklos*, *61*(2), 279–308. https://doi.org/10.1002/pits.20306

Shavit, Y., & Müller, W. (2000). Vocational secondary education, tracking, and social stratification. In M. Hallinan (Ed.), *Handbook of the sociology of education* (pp. 437–452). New York City, NY: Plenum.

Skinner, E. A., Furrer, C., Marchand, G., & Kindermann, T. (2008). Engagement and disaffection in the classroom: Part of a larger motivational dynamic? *Journal of Educational Psychology*, *100*(4), 765–781. https://doi.org/10.1037/a0012840

Skinner, E. A., Kindermann, T. A., & Furrer, C. J. (2009). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and Psychological Measurement*, *69*(3), 493–525. https://doi.org/10.1177/0013164408323233

Skinner, E. A., & Pitzer, J. R. (2012). Developmental dynamics of student engagement, coping, and everyday resilience. In *Handbook of research on student engagement* (pp. 21–44). Boston, MA: Springer.

Steiner, P. M., & Cook, T. D. (2013). Matching and propensity scores. In T. D. Little (Ed.), *The oxford handbook of quantitative methods* (pp. 237–259). Oxford, United Kingdom: University Press.

Stevens, P. A. J., & Vermeersch, H. (2010). Streaming in Flemish secondary schools: Exploring teachers' perceptions of and adaptations to students in different streams. *Oxford Review of Education*, *36*(3), 267–284. https://doi.org/10.1080/03054981003629862

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313

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. https://doi.org/10.1016/j.rssm.2009.12.004

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology*, *98*(4), 788–806. https://doi.org/10.1037/0022-0663.98.4.788

Turner, J. C., Midgley, C., Meyer, D. K., Gheen, M., Anderman, E. M., Kang, Y., & Patrick, H. (2002). The classroom environment and students' reports of avoidance strategies in mathematics: A multimethod study. *Journal of Educational Psychology*, *94*(1), 88. https://doi.org/10.1037//0022-0663.94.1.88

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

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. https://doi.org/10.1146/annurev.soc.012809.102538

Van den Branden, N. (2016). *De evolutie van non-cognitieve uitkomsten in de eerste graad secundair onderwijs. Analyse van de leerlingvragenlijst in het LiSO-project*. Ghent, Belgium.

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. https://doi.org/10.1086/422790

Van Houtte, M. (2006a). School type and academic culture: evidence for the differentiation--polarization theory. *Journal of Curriculum Studies*, *38*(3), 273–292. https://doi.org/10.1080/00220270500363661

Van Houtte, M. (2006b). Tracking and teacher satisfaction: Role of study culture and trust. *The Journal of Educational Research*, *99*(4), 247–256. https://doi.org/10.3200/JOER.99.4.247-256

Van Houtte, M. (2009). De verschillen tussen onderwijsvormen: een kwestie van culturen. *Caleidoscoop*, *21*(6), 25–30.

Van Houtte, M. (2016). Lower-track students' sense of academic futility: Selection or effect? *Journal of Sociology*, *52*(4), 874–889. https://doi.org/10.1177/1440783315600802

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. https://doi.org/10.1002/berj.3172

Vandecandelaere, M., Vansteelandt, S., De Fraine, B., & Van Damme, J. (2016). Time-varying treatments in observational studies: Marginal structural models of the effects of early grade retention on math achievement. *Multivariate Behavioral Research*, *51*(6), 843–864. https://doi.org/10.1080/00273171.2016.1155146

Vorst, H. C. M., & Smits, J. A. E. (2010). *Schoolvragenlijst voor basisonderwijs en voortgezet onderwijs [School questionnaire for primary education and secondary education]*. Nijmegen, the Netherlands: Berkhout.

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

Wang, M.-T., & Eccles, J. S. (2012). Adolescent behavioral, emotional, and cognitive engagement trajectories in school and their differential relations to educational success. *Journal of Research on Adolescence*, *22*(1), 31–39. https://doi.org/10.1111/j.1532-7795.2011.00753.x

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627

Wolbers, M. H. J. (2007). Patterns of labour market entry: a comparative perspective on school-to-work transitions in 11 European countries. *Acta Sociologica*, *50*(3), 189–210. https://doi.org/10.1177/0001699307080924

Wößmann, L. (2008). Efficiency and equity of European education and training policies. *International Tax and Public Finance*, *15*(2), 199–230. https://doi.org/10.1177/0001699307080924

Wylie, C., & Hodgen, E. (2012). Trajectories and patterns of student engagement: Evidence from a longitudinal study. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 585–599). Boston, MA: Springer US.