SONO
STEUNPUNT
ONDERWIJS ONDERZOEK

ONDERWIJSVORMEN EN ACADEMISCH ZELFCONCEPT

Dockx J, De Fraine B. & Vandecandelaere M.

SONO
STEUNPUNT
ONDERWIJS ONDERZOEK

ONDERWIJSVORMEN EN ACADEMISCH ZELFCONCEPT

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

Promotor: B. De Fraine

Research paper SONO/2017.OL1.1/13

Gent, september 2017

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 academisch zelfconcept. 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

Leerlingen hebben een perceptie van hun eigen schoolse kunnen door zich te vergelijken met hun omgeving. In de wetenschappelijke literatuur wordt dit *academisch zelfconcept* genoemd. Het academisch zelfconcept van een leerling kan zowel betrekking hebben op het schoolse kunnen in het algemeen als voor specifieke vakken. Voor dit zelfconcept kunnen leerlingen hun schoolse prestaties vergelijken met de medeleerlingen in hun klas. Leerlingen kunnen echter ook de schoolse prestaties van hun klas vergelijken met andere klassen en hun zelfconcept daarop baseren. Voor de vorming van het academisch zelfconcept van leerlingen speelt de klas dus een centrale rol.

In het Vlaamse secundair onderwijs worden klasgroepen doorgaans samengesteld op basis van de onderwijsvorm die de leerling koos. We onderscheiden 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. In wetenschappelijk onderzoek wordt het inrichten van verschillende onderwijsvormen *tracking* genoemd. *Tracks* worden in de literatuur als een belangrijk kenmerk beschouwd van de klas waarin een leerling schoolloopt.

Onderzoek naar de effecten van *tracks* op de ontwikkeling van het academisch zelfconcept is nodig om na te gaan hoe *tracks* academisch zelfconcept beïnvloeden. De onderzoeksresultaten worden geïnterpreteerd volgens twee belangrijke theoretische modellen over de invloed van de omgeving (vaak de klas, maar hier de *track*) op de vorming van het academisch zelfconcept. Het eerste model, het *big-fish-little-pond-effect*, geeft aan dat het academisch zelfconcept van leerlingen samenhangt met de relatieve positie die een leerling inneemt in de klasgroep. Een leerling zal een hoger academisch zelfconcept hebben als hij/zij deel uitmaakt van een eerder zwakke klas, dan wanneer hij/zij deel zou uitmaken van een eerder sterke klas; omdat de relatieve positie van de leerling gunstiger is binnen een zwakke klas. Het tweede model, het *basking-in-reflected-glory effect*, geeft aan dat leerlingen de gepercipieerde waarde van zijn/haar klas ook aan zichzelf toeschrijft. Een leerling zal een hoger academisch zelfconcept hebben als hij/zij deel uitmaakt van een eerder sterke klas, dan wanneer hij/zij deel zou uitmaken van een eerder zwakke klas. Aangezien de *tracks* de sterkte van de klas en de relatieve positie van een leerling binnen de klas mee bepalen verwachten we dat tracks een belangrijke rol spelen in de ontwikkeling van academisch zelfconcept. Er zijn dus twee onderzoeksvragen:

1. Verschillen *tracks* in gemiddelde ontwikkeling voor academisch zelfconcept?
2. Verschillen *tracks* in gemiddelde ontwikkeling voor academisch zelfconcept 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). Op elk van deze momenten werden er drie soorten zelfconcept gemeten: algemeen academisch zelfconcept, zelfconcept voor wiskunde en zelfconcept voor Nederlands. Het onderzoek beschrijft dus de effecten van *tracks* tijdens de eerste drie jaar van het secundair onderwijs.

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. 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 ontwikkeling van academisch zelfconcept 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 zelfconcepten tussen de *tracks*. De zelfconcepten zijn het hoogst in KT, daarna volgen de zelfconcepten in MW, TO en BV. In de loop van de eerste drie jaar secundair onderwijs stellen we vervolgens voor sommige groepen een evolutie vast in het niveau van het zelfconcept. Voor algemeen academisch zelfconcept maakt MW een matige daling en KT een kleine daling terwijl TO en BV stabiel blijven. Voor zelfconcept in wiskunde maken KT en MW een matige daling

terwijl TO en BV stabiel blijven. Voor zelfconcept in Nederlands maakt KT echter een kleine stijging terwijl MW, TO en BV stabiel blijven. In het algemeen worden BV en TO dus gekenmerkt door stabiliteit in zelfconcepten terwijl MW en KT meestal gekenmerkt wordt door een dalend zelfconcept. De daling bij MW is daarbij iets sterker. Toch blijft de rangorde aan de start van het secundair onderwijs grotendeels bewaard op het einde van het derde leerjaar secundair onderwijs.

Voor de tweede onderzoeksvraag beperken we onze vergelijking tussen vergelijkbare leerlingen die alsnog in verschillende *tracks* zitten. We vinden een klein positief effect voor KT vergeleken met MW voor algemeen academisch zelfconcept en zelfconcept in Nederlands. Er is geen verschil voor zelfconcept in wiskunde. We vinden een klein negatief effect voor MW vergeleken met TO voor algemeen academisch zelfconcept en zelfconcept in wiskunde. Er is geen verschil voor zelfconcept in Nederlands. We vinden matig negatieve effecten voor TO vergeleken met BV. Meestal is het dus voordelig voor het zelfconcept om in een *track* te zitten waar de gemiddelde leerling minder hoge prestaties laat optekenen, uitgezonderd bij de vergelijking tussen KT en MW.

Onze resultaten bieden ook geen steun voor de theoretische modellen van het *big-fish-little-pond-effect* en *basking-in-reflected-glory effect*. We vermoeden dat beide modellen té eenvoudig zijn omdat ze er van uitgaan dat het academisch zelfconcept van leerlingen gebaseerd is op slechts twee variabelen: het individuele prestatieniveau van de leerling en het gemiddelde prestatieniveau van de klas. Vermoedelijk zijn er nog heel wat andere invloeden op het academisch zelfconcept, zoals het curriculum, de communicatie door leraren en de mate van uitdaging in de leeromgeving.

We concluderen dat *tracks* de ontwikkeling in academisch zelfconcept kunnen beïnvloeden tijdens de eerste drie jaar van het Vlaamse secundair onderwijs. Het zijn daarbij vooral de *tracks* waar de leerlingen aan de start van het secundair onderwijs een hoger academisch zelfconcept hebben, KT en MW, die nadien gekenmerkt worden door een dalend academisch zelfconcept. Opvallend is echter dat de daling bij MW sterker is dan bij KT. *Tracks* waar de leerlingen aan de start van het secundair onderwijs al een lager academisch zelfconcept hebben, TO en BV, blijven stabiel op hun initieel niveau. Verder is er geen duidelijke ondersteuning voor de hypothese dat de keuze voor een hoger gewaardeerde *track* negatief zou zijn voor het academisch zelfconcept, noch voor de omgekeerde hypothese dat de keuze voor een hoger gewaardeerde *track* positief zou zijn voor het academisch zelfconcept. Als we enkel vergelijken tussen de voorlopers van het aso, tso en bso zonder verder in te gaan op het onderscheid tussen MW en KT, stellen we wel vast dat het voordelig lijkt te zijn voor het zelfconcept van de leerlingen om in een *track* te zitten waar de gemiddelde leerling minder goed presteert.

De resultaten geven hoofdzakelijk weer hoe leerlingen beïnvloed worden door de huidige structuur van het secundair onderwijs. Voor leerlingen in KT en MW is het vooral zorgwekkend dat het initieel hoge academisch zelfconcept daalt, waarbij de daling bij MW merkbaar groter is. Bij leerlingen in TO en BV is het vooral zorgwekkend dat hun initieel lage zelfconcept laag blijft na drie jaar secundair onderwijs.

Inhoud

1. Introduction

Academic self-concept is generally understood as a student's self-perception of academic ability (Marsh & Craven, 2006; Shavelson, Hubner, & Stanton, 1976). This appraisal of academic ability happens within a class or school, providing the students with a frame of reference. Two main theories describe how educational environments influence academic self-concept. The first, the big-fish-little-pond theory, states that students appraise their academic ability relative to their peers in class or school. The second, the basking-in-reflected-glory theory, states that students internalize the value society ascribes to the group they are a member of. Hence, students' self-perception of academic ability largely depends on their educational environments.

Tracking is a practice that shapes educational environments during secondary education. Tracks group students in different schools or classes, tailoring the educational environments to the specific abilities and interests of students (e.g. Hanushek & Wößmann, 2006; OECD, 2012; Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006; Van de Werfhorst & Mijs, 2010). However, how tracking is implemented differs across education systems (OECD, 2012; Trautwein et al., 2006). Generally, different tracks attract students who differ in mean academic ability, whereas the tracks also differ in their status in society. With different tracks offering different educational environments, key in shaping academic self-concept, there is a need to ascertain to what extent tracks influence academic self-concept.

Assessing the effects of being in either a lower or higher track requires data and methods that can control for selection bias that results from differential student intake across tracks. Moreover, data and methods which can describe longitudinal growth in academic self-concept are preferable (Raudenbush, 2001; Robins, 1997). To control for selection bias, we matched comparable students across different tracks and compared their changes in academic self-concept. In the following sections, literature on academic self-concept and tracking is described in more detail.

1.1. What is academic self-concept?

Academic self-concept is defined as a student's self-perception of his/her academic ability, based on personal educational experiences and corresponding inferences (Marsh & Craven, 2006; Shavelson et al., 1976). These experiences are situated in educational environments, typically schools and classrooms, providing frames of reference for interpretation. (Bong & Skaalvik, 2003). Generally, academic self-concept is stable and trait-like, but it is malleable to either positive or negative experiences (Marsh, Trautwein, Lüdtke, Baumert, & Köller, 2007).

Academic self-concept is part of the multidimensional general self-concept, the perception of general ability. This multidimensionality is hierarchically organized, with at the base level the perceptions of specific experiences, at the middle level inferences on specific domains and on top the general self-concept (Bong & Skaalvik, 2003; Shavelson et al., 1976). Academic self-concept

itself is considered one specific domain (Marsh & Craven, 2006), comprised of general academic self-concept and domain-specific academic self-concepts (i.e. mathematics, reading, …). General academic self-concept is typically considered hierarchically higher than the domain-specific academic self-concepts, even though the exact hierarchical structure is still under scrutiny (Brunner et al., 2010; Morin, Arens, & Marsh, 2016). This hierarchical nature of (academic) self-concept provides a framework to study student self-perceptions ranging from general to specific.

Interest in student academic self-concept not only derives from how it describes a cognitive appraisal of ability, but also from how it relates to other outcomes. This encompasses a student's self-efficacy (Bong & Skaalvik, 2003), academic interest (Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005), motivation (Marsh, Hau, Artelt, Baumert, & Peschar, 2006) and anxiety (Goetz, Preckel, Zeidner, & Schleyer, 2008). Academic self-concept is related to long-term outcomes as well, including long-term math interest, school grades, and standardized test scores (Marsh et al., 2005; Trautwein et al., 2006). Accordingly, much attention has been given to the relation between academic self-concept and academic performance. The consensus is that this relation is reciprocal, with academic self-concept being more predictive of academic performance than vice versa (Huang, 2011; Marsh & Craven, 2006; Marsh & Martin, 2011; Marsh et al., 2005; Pinxten, Marsh, De Fraine, Van Den Noortgate, & Van Damme, 2014; Seaton et al., 2014). This is consistent with studies showing that positive self-concepts make people more successfully engage with current challenges, being self-reinforcing (Bong & Skaalvik, 2003). Thus, academic self-concept is situated as both an antecedent and an outcome in a network of variables.

1.2. Frames of reference for academic self-concept

Two empirically supported theories describe how schools and classes function as frames of reference. The first focuses on how students compare their academic ability relative to their peers in class or school, widely known as the big-fish-little-pond effect (BFLPE) or contrast effect (students contrast themselves to their peer group). The second focuses on how students internalize the value society ascribes to the group they are a member of, known as the basking-in-reflected-glory effect (BIRGE) or assimilation effect (students assimilate the perceived value of their peer group).

The BFLPE is a contextual effect on student academic self-concept, due to the peer-group providing a personal frame of reference for a student rating his/her academic performance (Huguet et al., 2009). Students compare themselves with their classmates or other students in the school. This comparative process makes that the academic self-concept of a student with given ability is negatively related to the average ability of his/her class or school.(Ehmke, Drechsel, & Carstensen, 2010; Thijs, Verkuyten, & Helmond, 2010; Wang, 2015). The BFLPE has been found across different research fields (Marsh et al., 2008) and countries (Dai & Rinn, 2008; Marsh & Martin, 2011; Seaton, Marsh, & Craven, 2009), indicating its generalizability. Moreover, the BFLPE has been found across many academic domains, with the main focus on mathematics and reading (see Dai & Rinn, 2008, p.304-313).

Most discussions on the BFLPE are about the particular conditions under which a student contrasts his/her ability to a specific group. Most studies consider the school or a class as the frame of

reference (e.g. Wouters, Germeijs, Colpin, & Verschueren, 2011). However, whether classes, schools or other groups function as the most salient personal reference group remains a point of discussion. Zell and Alicke (2010) argued that students will rather compare themselves locally (i.e. the class) if multiple frames of reference are present. These authors advised investigating multiple frames of reference simultaneously during research (Alicke, Zell, & Bloom, 2010). There is no consensus on whether the class or school is most important as a frame of reference (Liem, Marsh, Martin, McInerney, & Yeung, 2013; Wouters et al., 2011).

The BIRGE is a contextual effect on self-perception, by a person internalizing the value society ascribes to the group they are a member of (Huguet et al., 2009; Mussweiler, 2003; Preckel & Brüll, 2010). This effect has been mainly studied in social psychology, usually on sports fans and on a wide range of self-perceptions (e.g. Bernache-Assollant, Lacassagne, & Braddock, 2007). It has also been applied in educational research. In this context, when society positively values the educational group the student is a member of, the positive value is assimilated in the academic self-concept, enhancing it (Marsh, Kong, & Hau, 2000; Preckel & Brüll, 2010). The size of this effect depends on the visibility of said group membership (Huguet et al., 2009; Preckel & Brüll, 2010). Although in our view, it remains ambiguous how to determine both which groups are highly valued by society and visibility of group membership.

Comparing the BFLPE and BIRGE, it is tempting to think that both should cancel each other out. However, there is a key difference in which situations they apply. The BFLPE depends on mean group academic performance, whereas the BIRGE depends on how highly this group is valued. Both are typically confounded, but this is not necessarily so. The few studies where the BFLPE was empirically distinguished from the BIRGE showed that both effects exist, with BFLPE's being somewhat larger (Marsh et al., 2000; Seaton et al., 2008). Although both effects may not cancel each other out, they will often (partially) mask each other (Marsh, Köller, & Baumert, 2001). Hence, when hypothesizing on educational environments and academic self-concept, both the differences in group mean academic performance and the status of these groups should be considered.

1.3. Track as a frame of reference

Tracking is the practice of placing students into different groups called tracks (OECD, 2012; Trautwein et al., 2006; Van de Werfhorst & Mijs, 2010), usually differing in mean ability and interest, creating more homogeneous groups. This allows for instructional practices to be tailored to specific student groups through fitting educational programs (e.g. Hanushek & Wößmann, 2006). Most education systems (OECD, 2012, p.57-58) track students during secondary education, however, each system has unique aspects. These differences encompass, but are not limited to, the age of allocation into tracks (e.g. Brunello & Checchi, 2007; OECD, 2012; Van de Werfhorst & Mijs, 2010), the criteria to allocate students (e.g. Brunello & Checchi, 2007; Buser, Niederle, & Oosterbeek, 2014), if between-school tracking or within-school tracking applies (Van de Werfhorst & Mijs, 2010), and the number of tracks offered (e.g. Bol & van de Werfhorst, 2013; Shavit & Müller, 2000). In sum, while tracking is ubiquitous, its characteristics differ across education systems.

Students enrolled in different tracks are subjected to different educational environments. Students with different academic abilities are allocated to different tracks, leading to different mean

academic abilities across tracks (Maaz, Trautwein, Lüdtke, & Baumert, 2008; Trautwein et al., 2006). Furthermore, tracks differ in how they are valued by society, and thus in their status (Kulik & Kulik, 1982; Van Houtte, 2006). Being allocated to a vocational track is perceived as failing the academic requirements of academic tracks, giving it a lower status. This sentiment is echoed by research into how being allocated to a lower track is considered a personal loss of status (Breen & Goldthorpe, 1997). Accordingly, lower tracks are also usually typified as less academically challenging (e.g. Salmela-Aro, Kiuru, & Nurmi, 2008; Stevens & Vermeersch, 2010). Hence, tracks generally differ in both average academic performance and status.

The differences between tracks provide students with different frames of reference in forming their academic self-concept. Specifically, differences in status incentive BIRGE's, whereas differences in average academic performance incentive BFLPE's.

1.4. Studies on tracks and academic self-concept

Hypothesizing that tracking practices matter for academic self-concept, several authors compared education systems. Using PISA 2003 data, Dupriez, Dumay and Vause (2008) found that strongly tracked countries have smaller differences in academic self-concept between lower-achieving and high-achieving students. They attribute this to BFLPE's in strongly tracked countries. Using PISA 2003 data, Chmielewski, Dumont, and Trautwein (2013) investigated how different grouping strategies shape academic self-concept in mathematics by controlling for individual and group mean achievement. In course-by-course grouping students in higher groups had higher academic self-concept for mathematics while students in lower groups had lower academic self-concept. However, in education systems with between-school and within-school tracking the reverse was observed, with students in higher groups having lower self-concept for mathematics while students in lower groups had higher self-concept for mathematics. The authors interpreted the former effect as a BIRGE due to course-by-course grouping having a high visibility, while the latter effect was interpreted as a BFLPE. Salchegger (2016) used both PISA 2003 and TIMSS 2007 data, finding that the negative effect of school-average achievement on academic self-concept was more pronounced in countries with earlier explicit school-level tracking. In short, tracking seems to increase the BFLPE in education systems.

Several authors investigated the effect of being in either a higher or lower track (or a comparable situation) on academic self-concept. A Singaporean study (Liu, Wang, & Parkins, 2005) showed lower-track students declining less in academic self-concept during the first three years of secondary education, compared to higher track students. Accordingly, Mulkey, Catsambis, Steelman and Crain (2005) found that high tracks negatively impact academic self-concept, using three-year panel data from. However, another US study (Chiu et al., 2008) showed no significant effect of track on academic self-concept after controlling for academic performance. Several German studies also investigated the effects of tracks. Trautwein, Lüdtke, Marsh and Nagy (2009) showed that being in a higher track was negative for academic self-concept, given comparable individual achievement. However, perceived status of the class had a positive effect, smaller in absolute size. In contrast, Preckel and Brüll (2010) found that students in a higher track had higher academic self-concept when controlling for student characteristics. Becker et al. (2014) compared students who made an early transition to secondary education (acceleration), based on ability to

regular students and observed a negative effect on academic self-concept. Meanwhile, Arens and Watermann (2015) found the existence of both BIRGE's and BFLPE's when comparing early transition students to regular students, although the BIRGE was smaller. In short, most studies support the BFLPE being stronger than the BIRGE. The few studies that distinguished between a BFLPE and BIRGE also found support for both, with the latter being smaller.

Studies analyzing the effect of being in a higher track on academic self-concept generally have several methodological shortcomings. Most of these studies did not use longitudinal data, preventing any inference on how track enrollment changes academic self-concept over time (Raudenbush, 2001; Robins, 1997). Most studies also used regression-based methods, with few covariates as controls to find a track effect, or did not use controls at all. With this approach, (remaining selection) bias due student selection effects into tracks is likely (see Miller & Chapman, 2001). Moreover, there is a lack of attention for comparability of students across tracks, risking extrapolation (King & Zeng, 2006). Hence, there is a need to investigate the effect of being in higher track on academic self-concept, using methods more suitable for reducing selection bias and describing change over time.

1.5. The current study

The first goal of this study was to investigate if being enrolled in a higher track affects academic self-concept within the Flemish education system. This research question is derived from the observation that different tracks offer different frames of reference, key in shaping academic self-concept. We paid specific attention to how these effects might change over time. Our second goal was to assess whether evidence for both the BFLPE's and BIRGE's existed, with both being plausible when investigating track effects.

In Flemish education, tracking starts at age 12, when students have to choose a secondary school (OECD, 2012, p57). The following tracks are available: classical, modern, technical, and vocational. While the first three tracks share a common core of educational goals during the first two years, each has a unique curriculum. The classical and modern track are academically focused, with students expected to follow tertiary education after these tracks. The technical track offers pathways towards both tertiary education and the labor market. The vocational track primarily prepares for the labor market. Tracks are a class-level variable, with most schools offering multiple tracks, but not all. Each school has a specific profile in attracting different students, based on the tracks they offer. Student track choice is completely free if a student has attained a certificate of primary education. If no certificate has been obtained, the student is obliged to go to the vocational track. The tracks have a known hierarchy in mean academic ability and mean socioeconomic background of students (Van Houtte, 2004). There is some flexibility in changing tracks, with students primarily remaining in their track or going down in the hierarchy of tracks (Boone & Van Houtte, 2013).

In this study, we needed to account for the differential intake of students across tracks. Therefore, we matched comparable students across tracks, hence differential intake of students across tracks could not bias possible track effects on development in academic self-concept. Considering robustness, we used three matching methods: propensity score matching (Schafer & Kang, 2008),

Mahalanobis distance matching (Rosenbaum & Rubin, 1985) and coarsened exact matching (Iacus, King, & Porro, 2012). Furthermore, our view that potential differences in academic self-concept may change over time, warrants the description of growth over time (Raudenbush, 2001; Robins, 1997). Hence, multiple indicator quadratic latent growth curve models were used (Duncan, Duncan, & Strycker, 2013).

Our main hypothesis was that being in a higher track causes a decline in students' academic self-concept, relative to comparable students in a lower track. We had no specific hypothesis on how these track effects change over time. We did hypothesize finding both BFLPE's and BIRGE's, with the former being larger. The following section describes the methodology to test these hypotheses in further depth.

2. Method

2.1. Sample

This study used data from the longitudinal LiSO-project (project Educational Trajectories in Secondary Education). This project follows a cohort of 6158 students in 48 schools who started secondary education in the school year 2013-2014. A regional sampling strategy was used, with almost all the students belonging to the aforementioned cohort in all the classes in all the schools within a certain area part of the study. Due to three schools de-tracking during first grade, 675 students (10.96%) were excluded from the analyses. Furthermore, 2278 students of the remaining subsample of 5483 students (41.55%) were excluded from the analyses as well, for they changed track during the first three years of secondary education. The remaining subsample consisted of 3205 students in 338 classes in 45 schools at the start of secondary education in September 2013 (the first month of school). 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 did not speak Dutch at home, while 21.40% of student parents in this sample were eligible for an educational grant due to low income. Student academic self-concept was measured at four time points: 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

The treatment variable was track, with four tracks in our sample. 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. 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. The hierarchically lower track was always the control track while the hierarchically higher track was the treatment track. Hence, a positive effect would indicate that a higher track predicts higher academic self-concept.

2.3. Measures

2.3.1. Outcomes

The outcome of interest was student academic self-concept. This was operationalized as general academic self-concept and domain-specific self-concepts for mathematics and Dutch. The academic self-concepts were each measured with four items at T0, T1, T2 and T3 and were part of a student questionnaire. The items are Dutch translations of the Self-Description Questionnaire-II (SDQ-II, Marsh et al., 2005), with four items for the domain-specific self-concepts being reductions in length of the original measures. At T3 the second item for academic self-concept in Dutch was not part of the students' questionnaire. Students answered on the items belonging to each measure on a five-point response scale, ranging from "not true" to "true". In the full dataset, the reliabilities of these measures ranged from 0.77 to 0.94 over time using Cronbach's Alpha.

2.3.2. Baseline covariates

Assessing an unbiased effect of track on academic self-concepts requires controlling for confounders that predict both track allocation and academic self-concepts. Controlling for every variable that predicts track allocation will yield an unbiased track effect, but will be inefficient (Golinelli, Ridgeway, Rhoades, Tucker, & Wenzel, 2012; Myers et al., 2011; Pearl, 2010). Hence, matching literature suggests only including those variables that predict the outcome of interest, in this case academic self-concepts (e.g. Brookhart et al., 2006; Myers et al., 2011). Table 1 gives a brief overview of the 24 covariates used during the different matching procedures, how they were measured, their reliability, with what instrument they were measured, and their correlation with the academic self-concepts at T3.

Table 1
Baseline covariates at T0

Variable	Description	Rel.	Info	r_{gen}	r_{mat}	r_{dut}	Mis
ASC General	Factor score general academic self-concept based on 4 items	.77	SQ	.39	.27	.29	.04
ASC Math.	Factor score academic self-concept mathematics based on 4 items	.91	SQ	.28	.43	.10	.04
ASC Dutch	Factor score academic self-concept Dutch based on 4 items	.86	SQ	.23	.01	.39	.04
ASC French	Factor score academic self-concept French based on 4 items	.92	SQ	.26	.09	.25	.04
Math. T0	IRT-score achievement in mathematics T0	.85	AT	.22	.30	.20	.03
Dutch T0	IRT-score achievement in Dutch T0	.82	AT	.22	.09	.29	.02
French T0	IRT-score achievement in French T0	.79	AT	.20	.09	.22	.05
Gender	Binary variable gender student, reference category is		OR	.05	.13	-.06	.00

Variable	Description	Rel.	Info	r_{gen}	r_{mat}	r_{dut}	Mis
	girl						
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	.12	.13	.10	.11
Allowance	Binary variable whether family is eligible for an allowance due to low income		OR	-.07	-.03	-.05	.00
Ed. mother	Binary variable whether mother is lowly educated		OR	-.07	-.06	-.06	.00
Other lang.	Binary variable whether the home language is not Dutch		OR	.01	.00	-.11	.00
Wellbeing	Factor score wellbeing based on 9 items	.82	SQ	.17	.10	.14	.04
Mindset	Factor score mindset (i.e. if intelligence in considered as static or flexible) based on 3 items	.55	SQ	-.05	-.03	-.06	.04
Aut. Mot.	Factor score autonomous based on 4 items	.77	SQ	.17	.09	.09	.04
Beh. Eng.	Factor score behavioral engagement based on 5 items	.78	SQ	.23	.13	.15	.04
Em. Eng.	Factor score emotional engagement based on 4 items	.77	SQ	.19	.08	.15	.04
Beh. Dis.	Factor score behavioral disengagement based on 5 items	.68	SQ	-.24	-.16	-.20	.04
Em. Dis.	Factor score emotional disengagement based on 6 items	.63	SQ	-.16	-.12	-.17	.04
Int. Math.	Sum score interest in mathematics based on 2 items		SQ	.18	.32	.02	.05
Int. Dutch	Sum score interest in Dutch based on 2 items		SQ	.09	-.02	.20	.05
Int. French	Sum score interest in French based on 2 items		SQ	.16	.05	.16	.05
Int. Tech.	Sum score interest in technology based on 2 items		SQ	.03	.06	-.06	.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. Matching

For each matching procedure, the goal was always to find comparable students across two tracks for each combination of confounder values (Schafer & Kang, 2008). This way, a matched dataset of students across tracks with equal confounder distributions could be constructed. What constitutes a comparable student and how the matched datasets are constructed differs across the matching methods (Stuart, 2010). After constructing a matched dataset, any effect of the track found should be unbiased given the ignorable treatment assumption (Rubin, 1978; Winship & Morgan, 2007). In practice, this means that in a matched dataset track allocation should be uncorrelated with the outcome of interest (Rubin, 1978; Winship & Morgan, 2007). In this study, three main matching procedures were used: propensity score matching (e.g. Caliendo & Kopeinig, 2008), Mahalanobis distance matching, and coarsened exact matching (e.g. Iacus et al., 2012). We chose for applying different matching methods due to the lack of consensus on which matching method is optimal under which conditions (Stuart, 2010). The different matching methods will be discussed in following paragraphs, while an overview of the different matching methods can be found in Table 3, which also displays the results of the matching procedure.

Matching students across different tracks necessitates a common support region. Using propensity score matching, this equals the overlap in propensity scores between treatment groups (Steiner & Cook, 2014). We therefore assessed the overlap in the density plots of propensity scores of both tracks for each comparison.

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 should not exceed 0.25 as a rule of thumb (Caliendo & Kopeinig, 2008).

2.4.1. Propensity score matching

The first step in propensity score matching (PSM), was to estimate for every student the propensity of being allocated to the higher track were estimated. 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 is random (Imbens & Rubin, 2015). We used logit models to estimate these propensities, with higher track allocation as outcome and the aforementioned baseline covariates as predictors. (Austin, 2011; Austin, Grootendorst, Normand, & Anderson, 2007; Myers et al., 2011) (see Table 1). Next, two propensity score procedures were applied: nearest neighbor caliper matching and full matching, in which students of both tracks were matched based on their propensity scores (e.g. Caliendo & Kopeinig, 2008).

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. 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, nearest neighbor caliper matching with caliper 0.05SD was conducted as one-to-one (1:1) matching, one-to-three (1:3) matching and one-to-three (1:3) matching with replacement. A modification had to be made for the technical and vocational track comparison concerning caliper 1:3 matching (with replacement). Students with propensities above 0.95 and under 0.05 had to be excluded from the matching procedures to prevent extreme weights causing unstable estimates.

Using full matching, lower track students are 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 this matching procedure by minimum and maximum propensity score. We applied full matching with propensity scores between 0.05 and 0.95 and full matching with propensity scores between 0.10 and 0.90.

2.4.2. 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 use a specific variation of this method whereby only students within a 0.25 propensity score caliper are considered for MDM (see Rosenbaum & Rubin, 1985, p35). In this implementation, we use 1:1 nearest neighbor matching without replacement.

2.4.3. 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. There are no clear guidelines on how to coarsen highly multivariate data with continuous scales. We chose to coarsen each of the continuous variables for general academic self-concept, self-concepts for mathematics, Dutch and French, academic performance in mathematics, Dutch and French, and socio-economic status in five bins. The dichotomous variables educational grant for poverty, other language than Dutch spoken at home and lowly educated mother all consisted of two bins. Hence, 3 125 000 strata were constructed, with most having no observations. Subsequently, strata with students from both tracks were reweighted so that a m

2.5. Outcome analyses with multiple indicator latent growth curves

In this study, multiple indicator latent growth curves (MILGC's) were used to describe growth in academic self-concept (Ferrer, Balluerka, & Widaman, 2008; Geiser, Keller, & Lockhart, 2013; McArdle, 1988). Specifically, a growth curve of latent factors approach was used (McArdle, 1988). This entails that the indicators for academic self-concept per time point were used in a confirmatory factor analysis per time point. Given the four time points, there are four factor analyses, each modeling a latent factor for academic self-concept. This was done for each outcome separately. Initially, the intercepts and factor loadings of these items on the latent factors were held equal across time points and tracks. Subsequently, based on these latent factors across time points a quadratic growth curve was specified (Duncan et al., 2013). This growth curve consists of a latent intercept (IC), a latent linear slope (LSL) and a latent quadratic slope (QSL). The means of the IC, LSL and QSL were freely estimated for each track, while the IC and LSL each had a freely estimated error term for each track. The error term of the QSL was always constrained to zero. We let the residuals of the same indicators on adjacent time points freely correlate, as is typical for these models. Furthermore, the residual variances of the latent factors were held equal across time points, as is typical in multilevel growth curve models (Duncan et al., 2013). For parameter estimation we used maximum likelihood with bootstrapped standard errors (using the resampling

method for clustered data, Asparouhov & Muthén, 2010). These models were specified in Mplus 8 (Muthén & Muthén, 2015). A simplified model for the MILGC-model for general academic self-concept is shown in Figure 1.

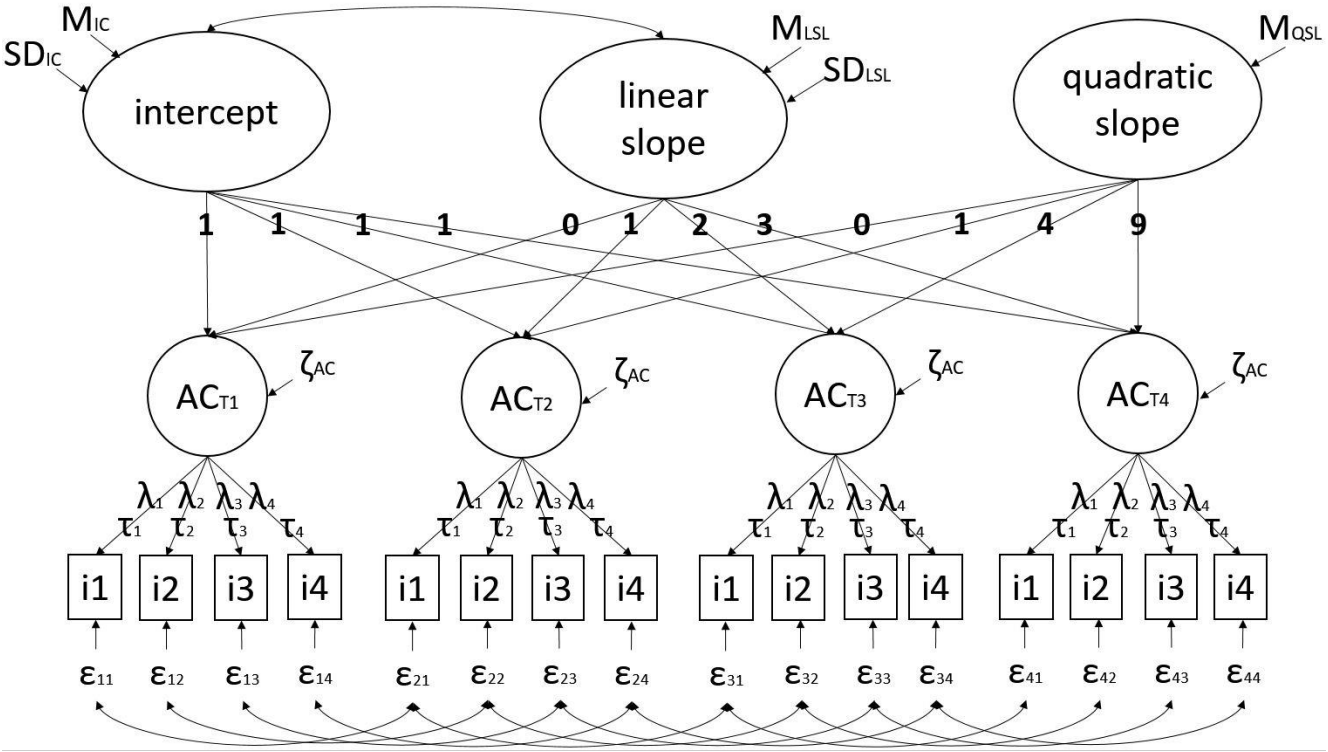

Figure 1. *Multiple indicator latent growth curve model general academic self-concept*

The benefit of MILGC's is the partitioning of variance between measurement error per time point, variance in the latent factor per time point and variance in the slopes across time points. Accordingly, these models are also more efficient in detecting differences between groups (Oertzen, Hertzog, Lindenberger, & Ghisletta, 2010). Furthermore, in recent literature on academic self-concept, there is a preference for using latent factors instead of estimated factor scores (Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Marsh et al., 2012).

There are two main prerequisites when using MILGC's to compare tracks: good model fit of the factor model per time point, and measurement invariance across time points and tracks (Baumgartner & Steenkamp, 2006; Cheung & Rensvold, 2002). The former means assessing how well the factor model reproduces the observed covariance matrix per time point using fit indices. The latter means assessing whether factor loadings and means of the indicators can be held equal over time and across tracks with sufficient model fit (i.e. metric and scalar invariance). Three fit indices were used in examining these conditions, the Comparative Fit Index (CFI), the Tucker–Lewis Index (TLI), and the root mean square error of approximation (RMSEA). We followed Hu and Bentler's (1999) cutoff criteria, that state that a model fits the data if the CFI is equal to or greater than .95, the TLI is equal to or greater than .95 and the RMSEA is equal to or less than .08. This was

tested for general academic self-concept and the self-concepts for Mathematics and Dutch in the unmatched sample across the four tracks. We used the modern track as reference, setting its latent mean to zero and its variance to one. When the cut-offs were attained, the resulting factor loadings and intercepts of the items were subsequently used to constrain the outcome analysis models. This way, all outcome analyses are on the same scale, while limiting the number of parameters estimated per outcome model. The final fit-indices and alterations made for achieving model fit per academic self-concept were reported prior to the analyses.

Assessing track effects on academic self-concept after matching, two different latent growth curves were estimated and compared using a multigroup model with the two tracks. The differences in self-concept development at T1, T2 and T3 were estimated based on the estimated LSL and QSL for each track. Significance was tested by assessing whether the estimated difference between both tracks differed significantly from zero at each time point. Differences in academic self-concepts between tracks at the end of each of the first three years of secondary education were reported as dT1, dT2 and dT3. Interpretation of the effect sizes is by Cohen's d (Cohen, 1977). Because negative estimates of LSL variance occurred in the technical and vocational track comparison, the LSL variance and covariance between the IC and LSL was set to zero. The choice for modeling a quadratic growth curve instead of other functional forms was of practical nature. A latent basis model (Grimm, Ram, & Hamagami, 2011) with free functional form yielded similar results, but had difficulties of convergence in matched datasets with larger weights. Autoregressive models per time point also yielded similar results, but were less efficient.

2.6. Assessing track effects controlling for class-mean achievement

We assessed whether track effects on academic self-concept encompass both a BFLPE and a BIRGE by modifying the matching procedure. Hence, we matched students across tracks as before, but only between classes with a maximum 0.5SD difference in class-mean achievements at T0 for academic performance in mathematics, Dutch reading comprehension and French reading comprehension. Reducing the variance in class-mean achievement should reduce the BFLPE and give way for showing the BIRGE. A more positive effect in these matched datasets compared to the former matched datasets we considered an indication of controlling for group composition and thus for the BFLPE. An effect swinging from negative to positive would be an indication of the BIRGE. 1:1 matching with caliper 0.05SD was used for this analysis. This could only be estimated for the both the classical and modern track comparison, and the modern and technical track comparison. Between the technical and vocational track there were not enough classes with comparable mean achievements.

2.7. Missing data

In our sample, on average 3.43% of the covariate data was missing 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 24 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 (the differences between tracks in self-concepts) ranged from 94.24% to 99.58% with an average of 97.55%. Hence, the results were unlikely to notably differ in precision from the perfect efficiency case. The imputations were estimated using the packages mice 2.30 (van Buuren & Groothuis-Oudshoorn, 2011) and pan 1.4 (Zhao & Schafer, 2016) in R 3.3.2.

Regarding the outcomes of interest, some students were censored due to missingness. Across the 12 items (11 items at T3) on average 14.77% were censored at T1, 7.38% at T2 and 7.45% at T3. To obtain unbiased and efficient estimates, full information maximum likelihood (FIML; Enders & Bandalos, 2001) was incorporated into the estimation of the parameters.

3. Results

3.1. Testing measurement invariance

Before matching and outcome analyses, multigroup factor models were used to test measurement invariance across tracks and time points. The goal was to find a final multigroup factor model for each self-concept which attained the cutoff values of the fit indices. The estimated item factor loadings and item intercepts were than used as fixed values in subsequent MILGC's. In what follows model fit of the final MILGC's is reported and which modifications were made to attain satisfactory model fit. Satisfactory model fit was achieved for general academic self-concept (RMSEA = 0.05, CFI = 0.96, TLI = 0.97). To achieve this model fit it was necessary to let the residuals of item 1 and item 2 covary freely, although the covariance was still held equal over time. Satisfactory model fit was achieved for self-concept in mathematics (RMSEA = 0.04, CFI = 0.95, TLI = 0.96), with no model modifications required. Thus Figure 1 accurately describes this model. Satisfactory model fit was achieved for self-concept in Dutch (RMSEA = 0.04, CFI = 0.96, TLI = 0.96). To achieve this model fit, it was necessary to let the factor loadings and intercept of item 4 be free over time, but still constrained to equality between tracks per time point. Furthermore, it was necessary to let the residuals of item 2 and item 3 covary freely, although the covariance was still held equal over time. Hence, for our three outcomes of interest there was (partial) measurement invariance. At T0 composite reliability (Raykov, 1997) in the reference group was 0.78 for general academic self-concept, 0.86 for academic self-concept in mathematics, and 0.86 for academic self-concept in Dutch. Appendix A shows the (estimated) parameter values of these MILGC's. The factor loadings and intercepts in this table were used as constraints in all subsequent analyses.

3.2. Development in academic self-concepts prior to matching

Table 2 describes the mean academic self-concepts at each time point per track estimated by the MILGC's. Generally, we see that the hierarchy in tracks is reflected in the academic self-concept at the start of secondary education. The general trend is that between T0 and T1 self-concepts are rather stable or even increase. At T3 most self-concepts have gone downward since T0 for the classical, modern and technical track. For the vocational track self-concept either increases between T0 and T3 or remains stable. However, the initial hierarchy between tracks in self-concepts largely remains in place, with the classical track having substantially larger average self-concepts. The differences in development between classical and modern track students are in favor of the classical track for general academic self-concept ($dT3 = 0.21$, $p < 0.05$) and self-concept in Dutch ($dT3 = .34$, $p < 0.05$), but trivial for self-concept in mathematics ($dT3 = -0.02$, $p = 0.76$). The differences in development between modern and technical track students are in favor of the modern track for general academic self-concept ($dT3 = -0.34$, $p < 0.05$), self-concept in mathematics ($dT3 = -0.36$, $p < 0.05$), but trivial for self-concept in Dutch ($dT3 = -0.14$, $p = 0.12$). The differences in development between technical and vocational track students are trivial for general academic self-concept ($dT3 = -0.10$, $p = 0.35$), self-concept in mathematics ($dT3 = -0.06$, $p = 0.54$) and self-concept in Dutch ($dT3 = -0.13$, $p = 0.22$).

Table 2
Predicted general academic self-concept and self-concepts for Mathematics and Dutch using multiple indicator multilevel latent growth curve models

Track	GASC				SCMATH				SCDUT			
	M_{T0} (*SE*)	M_{T1} (*SE*)	M_{T2} (*SE*)	M_{T3} (*SE*)	M_{T0} (*SE*)	M_{T1} (*SE*)	M_{T2} (*SE*)	M_{T3} (*SE*)	M_{T0} (*SE*)	M_{T1} (*SE*)	M_{T2} (*SE*)	M_{T3} (*SE*)
classical	.61 (.04)	.82 (.04)	.75 (.04)	.38 (.04)	.53 (.04)	.55 (.04)	.38 (.04)	.02 (.04)	.34 (.05)	.49 (.05)	.58 (.05)	.60 (.05)
modern	.00 (.04)	.09 (.04)	-.06 (.04)	-.44 (.04)	-.01 (.04)	.05 (.04)	-.11 (.04)	-.50 (.04)	-.02 (.04)	.07 (.04)	.04 (.04)	-.10 (.04)
technical	-.44 (.06)	-.36 (.05)	-.40 (.05)	-.53 (.05)	-.39 (.08)	-.36 (.07)	-.41 (.07)	-.52 (.07)	-.31 (.06)	-.27 (.06)	-.25 (.06)	-.27 (.06)
vocational	-.44 (.07)	-.23 (.07)	-.23 (.07)	-.43 (.07)	-.59 (.08)	-.32 (.07)	-.34 (.06)	-.66 (.07)	-.41 (.06)	-.10 (.05)	-.04 (.05)	-.24 (.06)

Note: M_{T0} - M_{T3} = Estimated mean achievement at T0 – T3 according to model; GASC = general academic self-concept, SCMATH = self-concept in mathematics, SCDUT = self-concept in Dutch

3.3. Track differences in propensity scores

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 propensity score model. These show the overlap between the propensity scores distributions before matching. The smaller the overlap, the less matches can be made. The plots show that there are differences in student selection across tracks. However, they also show a substantial area of common support between tracks, a required condition for any matching procedure. Although the overlap between the technical and vocational track is smaller, procuring matched datasets is possible.

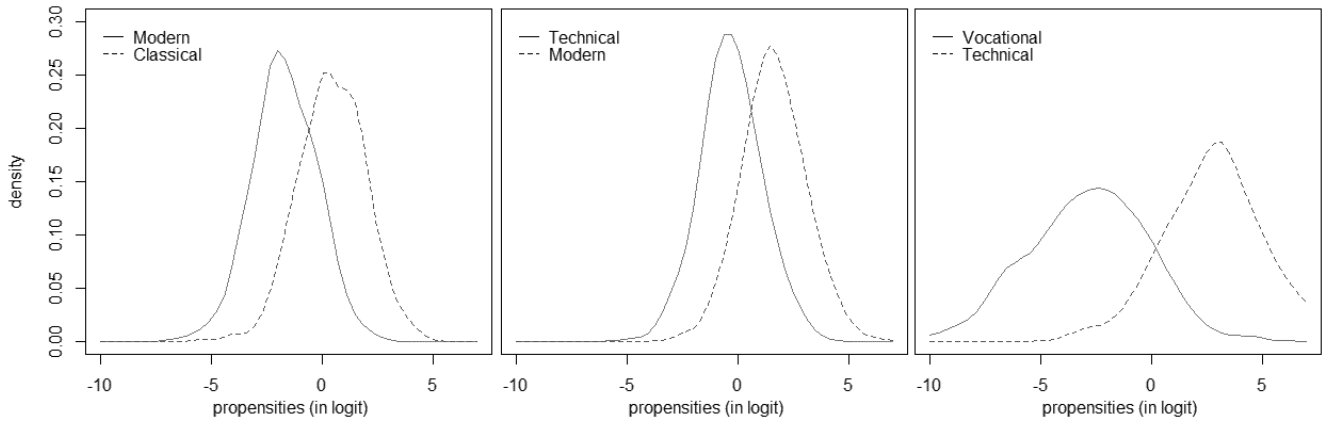

Figure 2. *Overlap propensity scores in pairwise comparisons of tracks*

3.4. 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 3 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 3
Indicators of remaining selection bias after application of matching 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
PSM Cal. 0.05 1:1	-.02	.01	-.11	.05	.02	.01	-.07	.07	.04	.01	-.07	.14
PSM Cal. 0.05 rep.	-.01	.00	-.09	.14	-.05	.00	-.23	.16	-.01	.00	-.39	.41
PSM Cal. 0.05 1:3	-.01	.00	-.09	.12	-.06	.00	-.22	.15	-.02	.00	-.33	.35
PSM Full .05-.95	.00	.00	-.15	.12	-.01	.00	-.10	.10	.00	-.04	.00	-.33
PSM Full .10-.90	-.02	.00	-.14	.16	.01	.00	-.11	.16	-.01	.00	-.2	.15
MDM	.01	.05	-.10	.14	.05	.06	-.08	.25	.03	.08	-.26	.26
CEM	.06	NA	.00	.21	.08	NA	.00	.28	.13	NA	-.10	.86

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 exceeded when using CEM. Using caliper matching and full matching, the mean propensity score difference between tracks was close to zero for each comparison. Thus, generally satisfactory balance was achieved between tracks per comparison.

Concerning the SMD's for individual covariates, the classical and modern track comparison SMD's were all under 0.25 across matching procedures. For the modern and technical track comparison, all SMD's were under 0.25, except when using CEM. For the technical and vocational comparison only for 1:1 matching with caliper 0.05SD and full matching between propensities 0.10 and 0.90 were all SMD's under 0.25. The larger SMD's in the other matching was likely due to strong differences between these tracks impeding the success of the matching procedure (Steiner & Cook,

2013). Hence, caution was needed when making inferences for the technical and vocational track comparison.

Although generally all matching procedures reached balance in mean SMD, the resulting matched sets had different mean propensities of being in a higher track. Tables 4, 5 and 6 show the mean propensities per track per matched sample. Across the three comparisons 1:1 matching with caliper 0.05SD yielded the lowest propensities. Adding replacement heightened the propensities, while allowing for multiple matches caused no change. Full matching had higher mean propensities, trending higher when allowing more extreme propensities into the weighting scheme. MDM yielded propensities somewhat comparable to 1:1 matching with caliper 0.05SD.

Another difference was the number of students in the matched samples, also shown in Tables 4, 5 and 6. 1:1 matching with caliper 0.05SD produced the smallest matched sets for propensity score matching. Allowing for replacement increased the number of students in the higher track, but lowered the number of students in the lower track. Allowing multiple matches increased the number of students in the lower tracks. Full matching with students between 0.05 and 0.95 propensity score yielded the largest sample sizes. Full matching with students between 0.10 and 0.90 propensity score reduced the number of matches. MDM attained datasets comparable to 1:1 matching with caliper 0.05SD. CEM generally yielded the smallest number of matched students.

3.5. Analyses 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. An example of the input in Mplus is given in Appendix B. For each comparison, the differences in mean value between both tracks at T1, T2 and T3 were estimated using the MILGC'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 4, 5 and 6 respectively. Figures 3, 4 and 5 show the growth curves of these comparisons for general academic self-concept, self-concept in Mathematics and self-concept in Dutch using 1:1 matching with caliper 0.05SD. In the following paragraphs, we discuss the general trends for each pairwise comparison.

For the classical and modern track comparison, the effects of being in a higher track on general academic self-concept ranged from d = 0.15 to d = 0.21 at T1, from d = 0.20 to d = 0.30 at T2 and from d = 0.15 to d =0.35 at T3. Effect sizes pointed to a trivial difference at T1, a small difference at T2 and a trivial to small difference at T3. The small differences were in favor of the higher track. Assessing Figure 3a reveals that the classical track first increased in general academic self-concept and then decreased after T2. The modern track seemed initially stable, but showed an accelerating downward trend over time. The effects of being in a higher track on self-concept in Mathematics ranged from d = -0.06 to d = 0.03 at T1, from d = -0.08 to d = 0.04 at T2 and from d = -0.08 to d =0.07 at T3. Effect sizes pointed to trivial differences across all time points. Figure 3b shows that both tracks were initially stable for self-concept in mathematics, but after T1 there was an accelerating downward trend. The effects of being in a higher track on self-concept in Dutch ranged from d = 0.11 to d = 0.20 at T1, from d = 0.20 to d = 0.35 at T2 and from d = 0.27 to d =0.45 at T3. Effect sizes pointed to a trivial to small difference at T1, a small difference at T2 and a small difference at T3.

The small differences were in favor of the higher track. Figure 3c shows that the modern track was stable with a slightly downward trend. The classical track however showed a decelerating upward trend.

Table 4
Differences classical and modern track in matched sample at T1, T2 and T3

Match	Track	N	M_{ps}	GASC			SCMAT			SCDUT		
				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*)
PSM Cal. .05	clas.	424	.43	.19*	.30*	.35*	-.01	.02	.07	.20*	.35*	.47*
	mod.	424	.42	(.08)	(.10)	(.11)	(.07)	(.10)	(.11)	(.09)	(.11)	(.10)
PSM Cal. .05 rep.	clas.	656	.56	.16	.22*	.17	-.06	-.08	-.07	.14	.23	.29*
	mod.	339	.56	(.08)	(.11)	(.11)	(.08)	(.10)	(.11)	(.11)	(.13	(.13)
PSM Cal. .05 1 to 3	clas.	656	.56	.15	.20	.15	-.04	-.07	-.08	.11	.20	.27*
	mod.	616	.56	(.08)	(.10)	(.10)	(.09)	(.10)	(.10)	(.11)	(.12)	(.12)
PSM Full .05 .95	clas.	656	.57	.15	.21	.16	-.03	-.05	-.06	.12	.21	.28
	mod.	1057	.57	(.09)	(.12)	(.13)	(.10)	(.11)	(.12)	(.11)	(.14)	(.15)
PSM Full .10 .90	clas.	592	.56	.21*	.29*	.24	.00	-.03	-.08	.14	.26	.36*
	mod.	810	.56	(.09	(.12)	(.13)	(.11)	(.13)	(.14)	(.11)	(.14)	(.14)
MDM	clas.	472	.46	.19*	.29*	.32*	.02	.04	.06	.20*	.35*	.44*
	mod.	472	.41	(.07)	(.09)	(.09)	(.07)	(.09)	(.10)	(.09)	(.11)	(.10)
CEM	clas.	392	NA	.15	.22*	.22	.03	.03	.01	.12	.27*	.45*
	mod.	463	NA	(.08)	(.11)	(.13)	(.11)	(.12)	(.11)	(.10)	(.12)	(.12)

Note. N = Number of students in matched set per track; 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 modern track at T0; NA = Not applicable, GASC = General academic self-concept, SCMAT = Self-concept mathematics, SCDUT = Self-concept Dutch. * Significant at α = 0.05

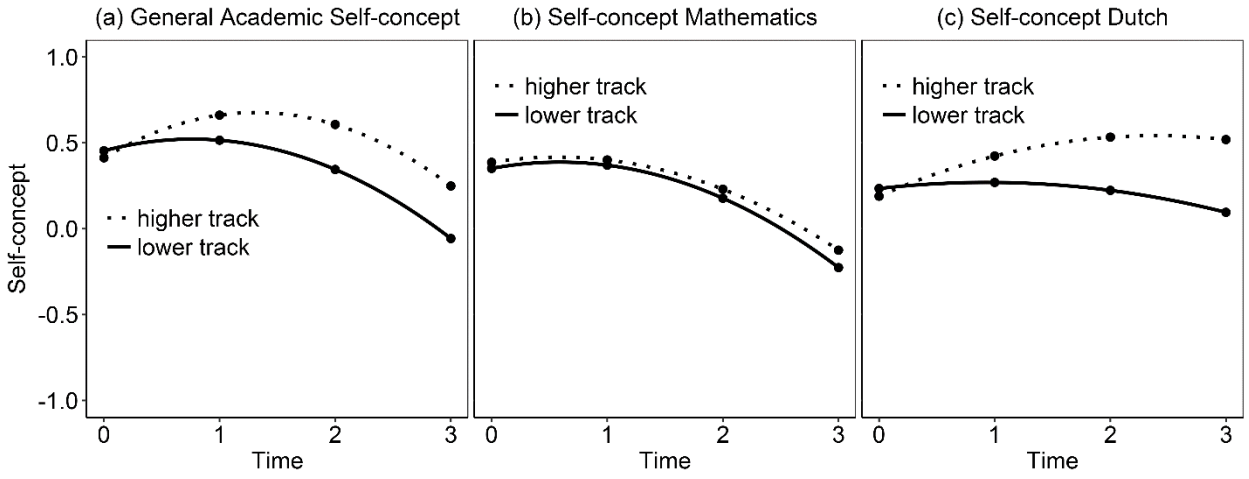

Figure 3. *Development self-concept in matched dataset classical/modern comparison.*

For the modern and technical track comparison, the effects of being in a higher track on general academic self-concept ranged from d = -0.04 to d = 0.01 at T1, from d = -0.13 to d = -0.08 at T2 and from d = -0.32 to d = -0.26 at T3. Effect sizes pointed to a trivial difference at T1 and T2, and a small difference at T3. The small differences were in favor of the lower track. Assessing Figure 4a reveals that both tracks were initially relatively stable. However, the modern track showed an accelerating downward trend while the technical track remained at the same level. The effects of being in a higher track on self-concept in Mathematics ranged from d = -0.04 to d = 0.11 at T1, from d = -0.15 to

d = 0.02 at T2, and from d = -0.31 to d = -0.25 at T3. Effect sizes pointed to trivial differences at T1 and T2, and a small difference at T3 in favor of the lower track. Figure 4b showed that the modern track initially rises, but thereafter had an accelerating downward trend. The technical track remained relatively stable over time. The effects of being in a higher track on self-concept in Dutch ranged from d = 0.04 to d = 0.16 at T1, from d = 0.03 to d = 0.16 at T2 and from d = -0.12 to d = 0.01 at T3. Effect sizes pointed to trivial differences across all time points. Figure 4c shows that both tracks remained at the same level over time.

Table 5
Differences modern and technical track in matched samples at T1, T2 and T3

Match	Track	N	M_{ps}	GASC			SCMAT			SCDUT		
				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*)
PSM Cal. .05	mod.	422	.58	-.02	-.12	-.30*	.10	.02	-.26	.06	.05	-.05
	tech.	422	.57	(.08)	(.09)	(.10)	(.09)	(.12)	(.14)	(.09)	(.11)	(.12)
PSM Cal. .05 rep.	mod.	1256	.77	.01	-.08	-.26	.05	-.05	-.29*	.16	.16	.00
	tech.	348	.76	(.10)	(.13)	(.15)	(.10)	(.13)	(.14)	(.14)	(.17)	(.14)
PSM Cal. .05	mod.	1256	.77	.00	-.09	-.26	.05	-.04	-.30*	.15	.15	-.01
1 to 3	tech.	507	.76	(.09)	(.12)	(.14)	(.10)	(.13)	(.13)	(.14)	(.17	(.14)
PSM Full	mod.	1052	.73	-.04	-.13	-.28*	.05	-.04	-.25	.07	.05	-.05
.05 .95	tech.	640	.73	(.09)	(.12)	(.14)	(.11)	(.14)	(.17)	(.14)	(.18)	(.16)
PSM Full	mod.	848	.68	.00	-.10	-.29*	.05	-.05	-.30	.04	.03	-.03
.10 .90	tech.	586	.68	(.10)	(.13)	(.15)	(.12)	(.16)	(.18)	(.15)	(.19)	(.16)
Maha.	mod.	464	.61	-.01	-.11	-.31*	.11	.02	-.27*	.07	.08	.01
	tech.	464	.55	(.07)	(.10)	(.11)	(.08)	(.11)	(.13)	(.09)	(.11)	(.11)
CEM	mod.	427	NA	-.01	-.12	-.32*	-.04	-.15	-.31*	.08	.04	-.12
	tech.	254	NA	(.11)	(.15)	(.16)	(.14)	(.17)	(.15)	(.12)	(.16)	(.16)

Note. N = number of students in matched set per track; 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 modern track at T0; NA = Not applicable, GASC = General academic self-concept, SCMAT = Self-concept mathematics, SCDUT = Self-concept Dutch. * Significant at α = 0.05

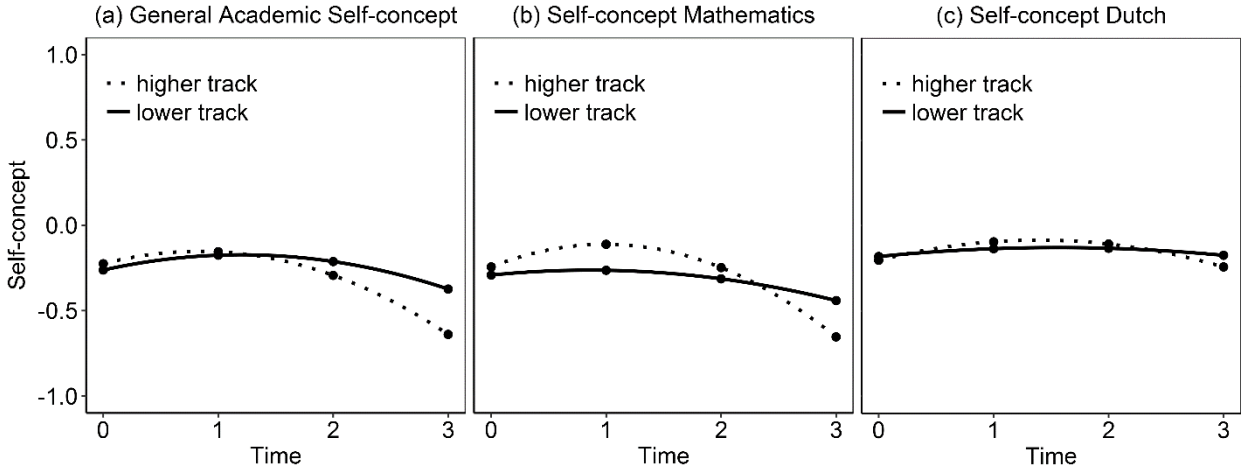

Figure 4. *Development self-concept in matched dataset modern/technical comparison.*

For the technical and vocational track comparison, the effects of being in a higher track on general academic self-concept ranged from d = -0.61 to d =-0.27 at T1, from d = -0.80 to d = -0.41 at T2 and from d = -0.61 to d = -0.40 at T3. Effect sizes pointed to a small to moderate difference at T1, a small to large difference at T2 and a small to moderate difference at T3. The differences were in favor of the lower track. Assessing Figure 5a reveals that the vocational track first sharply increased, while decreasing again after T2. The technical track seemed mostly stable, with a small decline between T0 and T3. The effects of being in a higher track on self-concept in Mathematics ranged from d = -0.71 to d = -0.34 at T1, from d = -0.90 to d = -0.42 at T2 and from d = -0.57 to d = -0.22 at T3. Effect sizes revealed a small to moderate difference at T1, a small to large difference at T2 and a small to moderate difference at T3. The differences were in favor of the lower track. Assessing Figure 5b reveals that the vocational track first sharply increased, while declining strongly again after T2. The technical track seemed mostly stable, with an overall small decline between T0 and T3. The effects of being in a higher track on self-concept in Dutch ranged from d = -0.56 to d = -0.26 at T1, from d = -0.69 to d = -0.35 at T2 and from d = -0.45 to d = -0.27 at T3. Effect sizes pointed to a small to moderate difference at T1 and T2, and a small difference at T3. These were in favor of the lower track. Assessing Figure 5c reveals that the vocational track first rises, while declining again after T2. The technical track seemed mostly stable.

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

Match	Track	N	M_{ps}	GASC			SCMAT			SCDUT		
				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*)
PSM Cal. .05	tech.	143	.52	-.39*	-.54*	-.45*	-.54*	-.68*	-.41*	-.32*	-.41*	-.27
	voc.	143	.51	(.16)	(.20)	(.21)	(.15)	(.19)	(.20)	(.16)	(.20)	(.21)
PSM Cal. .05 rep.	tech.	331	.72	-.45	-.64*	-.56*	-.65*	-.84*	-.57	-.52*	-.65*	-.41
	voc.	105	.72	(.24)	(.29)	(.25)	(.29)	(.37)	(.31)	(.19)	(.24)	(.27)
PSM Cal. .05 1 to 3	tech.	331	.72	-.45*	-.64*	-.58*	-.69*	-.88*	-.57*	-.53*	-.67*	-.43
	voc.	176	.72	(.20)	(.24)	(.22)	(.27)	(.34)	(.27)	(.19)	(.23)	(.25)
PSM Full .05 .95	tech.	331	.72	-.49	-.69*	-.61*	-.71*	-.90*	-.57	-.56*	-.69*	-.41
	voc.	295	.72	(.26)	(.32)	(.30)	(.32)	(.41)	(.33)	(.23)	(.30)	(.30)
PSM Full .10 .90	tech.	232	.66	-.61*	-.80*	-.58	-.69*	-.87*	-.52	-.51*	-.66*	-.45
	voc.	227	.66	(.22)	(.30)	(.31)	(.22)	(.28)	(.27)	(.25)	(.31)	(.28)
Maha.	tech.	162	.56	-.41*	-.57*	-.48*	-.59*	-.72*	-.39*	-.26*	-.35*	-.29
	voc.	162	.49	(.18)	(.22)	(.20)	(.14)	(.18)	(.17)	(.13)	(.16)	(.17)
CEM	tech.	94	NA	-.27	-.41	-.40	-.34	-.42	-.22	-.43*	-.55*	-.36
	voc.	66	NA	(.18)	(.24)	(.27)	(.19)	(.24)	(.28)	(.19)	(.26)	(.29)

Note. N = number of students in matched set per track; 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 modern track at T0; NA = Not applicable, GASC = General academic self-concept, SCMAT = Self-concept mathematics, SCDUT = Self-concept Dutch. * Significant at α = 0.05

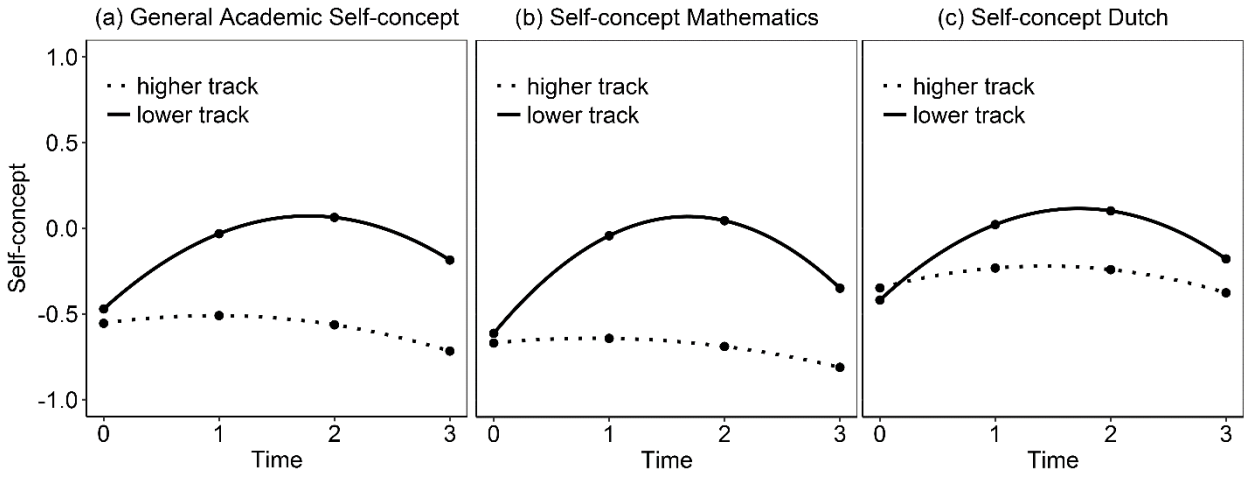

Figure 5. Development self-concept in matched dataset technical/vocational comparison.

3.6. 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 and 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 a small effect size (r = 0.2), a moderate effect size (r = 0.4) or a large effect size (r = 0.6), performed for each track effect. For brevity, only those for nontrivial results using 1:1 matching with caliper 0.05SD at T3 are reported here. Concerning the classical and modern track comparison for general academic self-concept, an unobserved confounder with a moderate (small/large) relation to general academic self-concept at T3 needed to differ between both tracks with a SD of 0.9 (1.8/0.6) to invalidate the effects. For self-concept in Dutch, an unobserved confounder with a moderate (small/large) relation self-concept in Dutch at T3 needed to differ between both tracks with a SD of 1.2 (2.4/0.8). Concerning the modern and technical track comparison for general academic self-concept, an unobserved confounder with a moderate (small/large) relation to general academic self-concept at T3 needed to differ between both tracks with a SD of 0.8 (1.5/0.5). For self-concept in mathematics, an unobserved confounder with a moderate (small/large) relation to reading comprehension at T3 needed to differ between both tracks with a SD of 0.7 (1.3/0.4). Concerning the technical and vocational track comparison for general academic self-concepts, an unobserved confounder with a moderate (small/large) relation to general academic self-concept at T3 needed to differ between both tracks with a SD of 1.1 (2.3/0.8). For self-concept in mathematics, an unobserved confounder with a moderate (small/large) relation to self-concept in mathematics at T3 needed to differ between both tracks with a SD of 1.3 (2.5/0.8). For self-concept in Dutch, an unobserved confounder with a moderate (small/large) relation to self-concept in Dutch at T3 needed to differ between both tracks with a SD of 0.7 (1.4/0.5).

Furthermore, some authors suggest including the variables used during propensity score estimation as covariates in the outcome analysis model (i.e. double robustness; Schafer & Kang,

2008). Hence, if the propensity score estimation did not achieve the necessary balance, these covariates can still allow for the estimation of an unbiased estimate. We added the variables as predictors for the IC, LSL and QSL when using 1:1 matching with caliper 0.05SD. For the classical and modern track comparison, the effects at T3 were d = 0.33, d = 0.07 and d = 0.46 for respectively the general academic self-concept and self-concept in mathematics and Dutch. For the modern and technical track comparison, the effects at T3 were d = -0.32, d = -0.31 and d = -0.01 for respectively the general academic self-concept and self-concept in mathematics and Dutch. For the technical and vocational track comparison, the effects at T3 were d = -0.46, d = -0.42 and d = -0.30 for respectively the general academic self-concept and self-concept in mathematics and Dutch. These results almost equal the results in Tables 4, 5 and 6.

3.7. Differences in track effects for class-mean achievement

To assess if any positive effect of being allocated to the higher track disappears if controlling for class-mean achievement, we also matched students in different tracks but with comparable class-mean achievements. Subsequently, we estimated the effect of being allocated to the higher track and compared them to the original matched datasets. We found that for the classical and modern track comparison, the effect of being allocated to a higher track at T3 is more positive for general academic self-concept (d= 0.53, t= 2.30, p< 0.05), more positive for academic self-concept in mathematics (d= 0.28, t= 1.17, p= 0.24) and less positive for academic self-concept in Dutch (d= 0.22, t= 1.01, p= 0.31). For the modern and technical track comparison, the effect of being allocated to a higher track at T3 is equal for general academic self-concept (d= -0.29, t= -1.53, p= 0.13), more positive for academic self-concept in mathematics (d= -0.06, t= -0.24, p= 0.81) and equal for academic self-concept in Dutch (d= 0.00, t= 0.02, p= 0.98).

4. Discussion

This study addressed the question whether being allocated to a higher track matters for the development of academic self-concept during the first three years of Flemish secondary education. It expands on prior research by using a quasi-experimental approach through matching students across tracks and describing how development in academic self-concepts changes over time. It not only looks at general academic self-concept but also domain-specific academic self-concepts of Mathematics and Dutch. Furthermore, we assessed whether there is evidence for both BFLPE's and BIRGE's.

Our results show that for five out of nine comparisons, being allocated to the higher track is detrimental for academic self-concept at the end of the third year. This is in line with the BFLPE hypothesis, expecting that being surrounded by academically stronger students is detrimental for academic self-concept. It also agrees with most of prior studies on track effects or similar grouping

methods (Arens & Watermann, 2015; Becker et al., 2014; Liu et al., 2005; Mulkey et al., 2005; Trautwein et al., 2009). However, two results show opposite findings, with the allocation to the higher track being beneficial for self-concept at the end of the third year. This is in line with the BIRGE hypothesis, expecting that belonging to a more valued group (i.e. a higher track) is beneficial for self-concept. Again, this is not an unprecedented result (Preckel & Brüll, 2010). For two comparisons, the differences are trivial, but again, this is not a unique result (Chiu et al., 2008). Thus, our findings seem generally in line with prior research, pointing to BFLPE's in most comparisons but not all.

Exploring how these track effects develop over time shows that the effects are most pronounced after two or three years in secondary education. All differences at the end of the first year remain trivial for both the classical and modern track comparison, and the modern and technical track comparison. Only for the technical and vocational track comparison the differences are already clear at the end of the first year. These results are in line with prior research on BFLPE's, showing that they often only become clear after some time has passed (Marsh et al., 2000, 2007). Hence, when investigating effects of grouping strategies on academic self-concepts, a long time span wherein these effects can be assessed should be considered.

Further assessing how academic self-concept develops in our (matched) dataset(s) reveals that there is not a consistent pattern of growth and acceleration across tracks and self-concepts. The two trivial results are also for different academic self-concepts in different track comparisons, showing that different academic self-concepts do not yield exchangeable results. Previously, few attention has been given if track effects differ across academic self-concepts, with most studies focusing solely on either general academic self-concept (Arens & Watermann, 2015; Becker et al., 2014; Liu et al., 2005) or self-concept for mathematics (Mulkey et al., 2005; Preckel & Brüll, 2010; Trautwein et al., 2009). In our view, this corroborates studies on the validity of academic self-concepts wherein higher order factor models show unique variances for domains-specific self-concepts and rather low correlations between them (Arens, Yeung, Craven, & Hasselhorn, 2011; Möller, Pohlmann, Köller, & Marsh, 2009). We do notice that nontrivial effects always go in the same direction, both in the matched and complete datasets. Hence, we conclude that track effects for general-academic self-concepts and domain-specific self-concepts are distinct, but not in opposite directions.

Of substantive interest in interpreting the effects of being in a higher track on academic self-concept in matched datasets is that most already occur in the complete dataset. Hence, for the classical and modern track comparison, small effects at the end of the third year are also present in the complete dataset. Accordingly, for the modern and technical track comparison the small effects found in the matched datasets are already present in the complete dataset. For both comparisons, trivial effects in the matched dataset are also trivial in the complete dataset. In our view, these findings argue against the BFLPE hypothesis. We expected that for students with equal achievement (and other characteristics held equal) the higher mean academic achievement of the higher track should negatively affect academic self-concepts compared to the lower tracks (Ehmke et al., 2010; Thijs et al., 2010; Wang, 2015). However, considering that the effect of being in a higher track does not differ between matched dataset and the complete dataset, explaining them as BFLPE's seem implausible. Only for the technical and vocational track comparison are the effects

in the matched datasets not present in the complete dataset. If interpreting that for two out of three track comparisons the matched students simply follow the average track trend of academic self-concept development, the BFLPE hypothesis must be rejected.

Further doubt is cast on both the BFLPE and BIRGE hypotheses when matching students across tracks, but in classes with comparable mean achievement. If the BFLPE hypothesis is true, than the effect of being in a higher track on academic self-concepts should disappear when holding class-mean achievement equal (Ehmke et al., 2010; Thijs et al., 2010; Wang, 2015). The remaining track effect should also be positive when the BIRGE hypothesis is true, for the higher track is still a more valued group (Huguet et al., 2009; Mussweiler, 2003; Preckel & Brüll, 2010). However, only in three out of six comparisons, the track effects slightly decrease. None of the effects turns positive. Hence, we find no strong evidence for the BFLPE's and BIRGE's being meaningfully involved in the track effects.

How come that neither BFLPE's and BIRGE's seem to occur when assessing our results? Concerning the former, in the context of gifted education it has been argued by Marsh et al. (2008) that class-mean achievement and individual achievement are not the only variables that impact academic self-concept development. They argue that differences in curriculum, higher training of teachers and the more stimulating environments provided gifted programs may more strongly influence academic self-concept development. We think the same argument can be made for tracks. Hence, tracks in Flemish education differ in curriculum, higher tracks are characterized by higher levels of problem solving and cognitive activating instructions, whereas in low tracks memorization and disciplining students are emphasized (Kunter & Baumert, 2006; Retelsdorf, Butler, Streblow, & Schiefele, 2010; Van Houtte, 2004), lower tracks have an anti-school culture (Van de gaer, Pustjens, Van Damme, & De Munter, 2006), tracks differ in teacher beliefs about their classrooms (Hallam & Ireson, 2003), higher tracks have teachers with more pedagogical content knowledge (Baumert et al., 2010; Krauss et al., 2008) and lower tracks are considered less academically challenging (e.g. Salmela-Aro, Kiuru, & Nurmi, 2008; Stevens & Vermeersch, 2010). Thus, different tracks offer educational environments which differ so much, that BFLPE's may not meaningfully apply.

If accepting the former explanation, it begs the question why other studies did find BFLPE's between tracks. Unfortunately, most studies did not report enough to assess whether BFLPE's are a more plausible explanation than general differences in academic self-concept development between tracks. One exception is the study by Becker et al. (2014) who compared students who made an early transition to secondary education to regular students, using a comparable matching approach as our study. They concluded finding a BFLPE, but when we assess their results we find that the effect in matched dataset mirrors the effect in the complete dataset. In fact, in the matched dataset the negative effect of being transitioned early is less pronounced, whereas it should be more pronounced given the BFLPE hypothesis. Liu et al. (2005) did also conclude finding BFLPE's when investigating tracks. However, they did not control for individual student characteristics and rather described differences in mean academic self-concept development per track. In our view differences in mean academic self-concept development per track could lead to erroneously perceiving BFLPE's.

Another explanation for the absence of BFLPE's could be that consecutive pairs of tracks did not differ enough in class-mean achievement to elicit these effects. Accordingly, it is striking that the

overlap in propensity scores is larger in Flemish education as compared to studies in Germany (e.g. Becker et al., 2014; Becker, Lüdtke, Trautwein, Köller, & Baumert, 2012; Guill, Lüdtke, & Köller, 2016; Retelsdorf, Becker, Köller, & Möller, 2012). It may therefore not be by chance that we did find BFLPE's when comparing the technical and vocational tracks, for these are much more distinct in class-mean achievement. Unfortunately, we could not find comparative studies on differences in academic performances between tracks across education systems to further substantiate this reasoning.

Concerning BIRGE's, it should be noted that we have no insight whether students truly deem higher tracks to be more highly valued. However, the selection process of track allocation at the start of Flemish secondary education is highly visible. Given that this is the main condition for the perceived value (Marsh et al., 2000), we argue that track valuation in Flanders has a high saliency. Yet, we did not find any evidence for BIRGE's. This finding is not out of step with prior studies, which are inconsistent in finding such effects (Marsh et al., 2008).

Although not the main purpose of this study, the MILGC's reveal that the quadratic growth curves do not explain all variance in the academic self-concepts (see Appendix A). At the start of secondary education, the variance in academic self-concepts that was unexplained by the quadratic growth curves ranged from 30.41% to 46.53%. This is not measurement error, the MILGC's account for this, but true variance in the academic self-concepts unique to the start of secondary education (Ferrer et al., 2008). Recently, the application of MILGC's and its separation of time-specific variance and latent growth curve variance has been connected to the latent state-trait theory (Steyer, Schmitt, & Eid, 1999). It is suggested that the variance in the growth curves constitutes variance in the latent trait of interest, whereas the remaining variance per time point constitutes variance in the temporary state (Geiser et al., 2015; Steyer, Mayer, Geiser, & David, 2015). Using this logic, evaluating effect sizes should be based on the variance of the latent trait, not the variance of the temporary state. Therefore, rather than using the standard deviation of the modern track at the start of secondary education (always 1 due to setting the measurement scale) as a base to describe effect size, we can also use the estimated standard deviation of the intercept. In this study, this entails that the effect sizes should be 15.47% larger for general academic self-concept, 16.25% larger for self-concept in Mathematics and 36.08% larger for self-concept in Dutch. This does not alter our conclusion, but does make the effects slightly more pronounced.

5. Limitations and suggestions for future research

This study starts from the hypothesis that tracks could constitute a frame of reference for student academic self-concept development. Rejecting the BFLPE and BIRGE hypotheses, it may seem plausible to reject tracks as important frames of references altogether. However, our results show that, although our hypotheses should be rejected, the effects of tracks on academic self-concepts are substantial. Specifically, we find that there are general trends in academic self-concept

development in each track. Our study does not assess how these general trends in academic self-concepts come to be, warranting further research.

Although this study paid attention to both general academic self-concept and domain-specific self-concepts in mathematics and Dutch, it cannot assess more aspects of the multidimensional nature of academic self-concept (Marsh & Craven, 2006). Furthermore, there is much discussion on how general academic self-concept is best measured (i.e. via a specific scale or through higher order factor analyses) and how domain-specific self-concepts are best operationalized (Brunner et al., 2010; Morin et al., 2016). Operationalizing academic self-concept with (four items of) the SDQ-II may be a prevalent approach, it is not necessarily the most valid way to assess the underlying concepts. Whether differences in operationalization of these concepts could influence results should remain under scrutiny.

This study only investigates students who remain in their tracks for three years. Many students change tracks in our study, yet they were removed from the sample. Although this issue is not unique to this study (e.g. Guill et al., 2016), it immediately raises the question how changing tracks influences academic self-concepts. We could find only one study which investigated track changes, showing that track change has a tangible effect on academic self-concept (Wouters, De Fraine, Colpin, Van Damme, & Verschueren, 2012). Furthermore, it could be that students who change tracks have a substantial role in how other students who remain in their track develop their academic self-concept. Indeed, perhaps a student who remains in his/her track compares oneself favorably to students who change tracks and vice versa, constituting a salient frame of reference. Although investigating how track changes influence academic self-concept through quasi-experimental methods is possible, it requires hefty assumptions and large samples (Robins, Hernan, & Brumback, 2000).

The success of a matching approach rests on the ignorable treatment assumption, i.e. there should be no confounder that predicts both track assignment and the outcome (Rosenbaum & Rubin, 1983; Steiner & Cook, 2014). Applying different matching methods generally yielded comparable effect sizes. Only for the classical and modern track comparison the choice for matching method matters. We found that this is due to the track effect being slightly stronger for students with a low propensity to be allocated to the higher track. We also conducted sensitivity tests of the treatment effect (Caliendo & Kopeinig, 2008; Vanderweele & Arah, 2011) across track comparisons and applied a double robustness approach (Schafer & Kang, 2008). In our view, it seems that small effects could still be somewhat plausibly explained by an unobserved confounder, but track effects of moderate size or larger cannot. The doubly robustness approach does not change the results in any meaningful way. Hence, while the ignorable treatment assumption cannot be truly tested, it at least seems a tenable position in this study.

Any estimate on a matched dataset is limited in inference to the area of common support for which enough statistical power exists in compared groups (Stuart, 2010). This limits the extent to which the results of our study generalize to the entire population of a track. Although it should be noted that this problem is not unique to quasi-experimental methods (such as matching), but makes it more visible (King & Zeng, 2006).

6. Conclusion

This study shows that tracks meaningfully influence academic self-concept development during the first three years of Flemish secondary education. For one out of three comparisons, it is beneficial to be allocated to the higher track, while for two out three comparisons it detrimental to be allocated to the higher track. Investigating these effects shows that both BFLPE's and BIRGE's are not plausible explanations for these trends. Rather, it seems more plausible that each track has its own unique trend in the developments of academic self-concepts, which explains why comparable students across tracks develop differently. Furthermore, our results show that these effects only clearly reveal themselves over a longer time span.

BIJLAGEN

Appendix A

Parameters multiple group multiple indicator quadratic latent growth curve across all tracks

	GASC	SCMAT	SCDUT
Factor loading			
λ_{i1T0}			
λ_{i1T1}	0.47	0.67	0.54
λ_{i1T2}			
λ_{i1T3}			NA
λ_{i2T0}			
λ_{i2T1}	0.60	0.71	0.58
λ_{i2T2}			
λ_{i2T3}			
λ_{i3T0}			
λ_{i3T1}	0.56	0.76	0.79
λ_{i3T2}			
λ_{i3T3}			
λ_{i4T0}	0.47		0.83
λ_{i4T1}	0.45	0.86	0.87
λ_{i4T2}	0.50		0.87
λ_{i4T3}	0.55		0.85
Intercept			
τ_{i1T0}			
τ_{i1T1}	3.61	3.85	4.05
τ_{i1T2}			
τ_{i1T3}			NA
τ_{i2T0}			
τ_{i2T1}	3.85	3.70	3.81
τ_{i2T2}			
τ_{i2T3}			
τ_{i3T0}			
τ_{i3T1}	3.76	3.85	3.59
τ_{i3T2}			
τ_{i3T3}			
τ_{i4T0}	4.28		3.45
τ_{i4T1}	4.27	3.70	3.40
τ_{i4T2}	4.26		3.34
τ_{i4T3}	4.28		3.18
Item covariance			
$\theta_{i1T0\text{-}i1T1}$			0.03
$\theta_{i1T1\text{-}i1T2}$	0.09	0.01	
$\theta_{i1T2\text{-}i1T3}$			NA
$\theta_{i2T0\text{-}i2T1}$			
$\theta_{i2T1\text{-}i2T2}$	0.02	0.03	0.01
$\theta_{i2T2\text{-}i2T3}$			
$\theta_{i3T0\text{-}i3T1}$			
$\theta_{i3T1\text{-}i3T2}$	0.02	0.04	0.04
$\theta_{i3T2\text{-}i3T3}$			
$\theta_{i4T0\text{-}i4T1}$			
$\theta_{i4T1\text{-}i4T2}$	0.05	0.07	0.04
$\theta_{i4T2\text{-}i4T3}$			

	GASC				SCMAT				SCDUT			
$\theta_{i1T0\text{-}i2T0}$ $\theta_{i1T1\text{-}i2T1}$ $\theta_{i1T2\text{-}i2T2}$ $\theta_{i1T3\text{-}i2T3}$	NA				0.16				NA			
$\theta_{i2T0\text{-}i3T0}$ $\theta_{i2T1\text{-}i3T1}$ $\theta_{i2T2\text{-}i3T2}$ $\theta_{i2T3\text{-}i3T3}$	NA				NA				0.09			
	Clas.	Mod.	Tec.	Voc.	Clas.	Mod.	Tec.	Voc.	Clas.	Mod.	Tec.	Voc.
Mean												
IC	0.61	0.00	-0.44	-0.44	0.53	-0.01	-0.39	-0.59	0.34	-0.02	-0.31	-0.41
LSL	0.36	0.20	0.12	0.32	0.12	0.17	0.06	0.42	0.19	0.14	0.06	0.44
QSL	-0.15	-0.12	-0.05	-0.11	-0.10	-0.11	-0.04	-0.15	-0.03	-0.06	-0.02	-0.13
(Co-)variance												
IC	0.41	0.75	0.76	0.97	0.40	0.74	0.84	1.03	0.48	0.54	0.71	0.80
LSL	0.05	0.06	0.06	0.12	0.04	0.05	0.05	0.07	0.02	0.04	0.04	0.06
IC - LSL	0.00	-0.04	-0.04	-0.14	0.02	-0.04	-0.03	-0.05	-0.04	-0.04	-0.07	-0.08
Residual variance												
$\theta_{\varepsilon i1T0}$	0.50	0.52	0.66	0.82	0.32	0.43	0.48	0.86	0.40	0.40	0.50	0.82
$\theta_{\varepsilon i1T1}$	0.43	0.51	0.67	0.78	0.29	0.33	0.48	0.64	0.27	0.32	0.44	0.60
$\theta_{\varepsilon i1T2}$	0.45	0.47	0.58	0.73	0.31	0.37	0.40	0.62	0.34	0.33	0.40	0.61
$\theta_{\varepsilon i1T3}$	0.38	0.45	0.52	0.62	0.32	0.35	0.39	0.62	NA	NA	NA	NA
$\theta_{\varepsilon i2T0}$	0.21	0.25	0.32	0.45	0.29	0.34	0.40	0.71	0.23	0.25	0.27	0.39
$\theta_{\varepsilon i2T1}$	0.18	0.22	0.32	0.42	0.34	0.34	0.40	0.48	0.25	0.25	0.28	0.38
$\theta_{\varepsilon i2T2}$	0.20	0.21	0.30	0.34	0.31	0.34	0.41	0.59	0.19	0.24	0.29	0.39
$\theta_{\varepsilon i2T3}$	0.15	0.18	0.22	0.32	0.31	0.36	0.44	0.66	0.21	0.23	0.26	0.43
$\theta_{\varepsilon i3T0}$	0.18	0.22	0.22	0.43	0.21	0.34	0.36	0.64	0.14	0.22	0.26	0.39
$\theta_{\varepsilon i3T1}$	0.18	0.22	0.23	0.35	0.23	0.27	0.41	0.70	0.28	0.36	0.32	0.43
$\theta_{\varepsilon i3T2}$	0.18	0.18	0.25	0.31	0.24	0.23	0.34	0.52	0.24	0.30	0.31	0.43
$\theta_{\varepsilon i3T3}$	0.20	0.20	0.23	0.33	0.27	0.30	0.40	0.68	0.23	0.35	0.40	0.46
$\theta_{\varepsilon i4T0}$	0.25	0.32	0.40	0.72	0.30	0.32	0.48	0.72	0.36	0.30	0.30	0.54
$\theta_{\varepsilon i4T1}$	0.21	0.32	0.37	0.61	0.28	0.40	0.44	0.50	0.39	0.36	0.27	0.55
$\theta_{\varepsilon i4T2}$	0.24	0.28	0.37	0.64	0.30	0.35	0.40	0.49	0.36	0.28	0.39	0.43
$\theta_{\varepsilon i4T3}$	0.22	0.33	0.37	0.55	0.33	0.42	0.39	0.53	0.48	0.42	0.43	0.55
ψ_{T0} ψ_{T1} ψ_{T2} ψ_{T3}	0.25	0.36	0.38	0.50	0.25	0.37	0.46	0.45	0.36	0.47	0.50	0.49

Note: GASC = general academic self-concept; SCMAT = self-concept mathematics ; SCDUT = self-concept Dutch; *clas.* = classical track; mod = modern track; tech. = technical track; voc. = vocational track; i1 – i4 = item 1 – item4 *; T0 – T3 = time 0 – time 3; IC = intercept parameter growth curve; LSL = linear slope parameter growth curve; QSL = quadratic slope parameter growth curve*

Appendix B

Input Mplus for outcome analyses general academic self-concept classical/modern comparison.

```
DATA:
FILE = dataset.csv;

VARIABLE:
NAMES=
LISOID
AC1301 AC1302 AC1303 AC1304
AC1401 AC1402 AC1403 AC1404
AC1501 AC1502 AC1503 AC1504
AC1601 AC1602 AC1603 AC1604
TREAT WEIGHT;
IDVARIABLE=LISOID;
MISSING ARE ALL (9999999);
USEVARIABLES=
AC1301 AC1302 AC1303 AC1304
AC1401 AC1402 AC1403 AC1404
AC1501 AC1502 AC1503 AC1504
AC1601 AC1602 AC1603 AC1604;
GROUPING IS TREAT (0=Modern 1=Classical);
WEIGHT=WEIGHT;
CLUSTER=C13SEP;

Analysis:
TYPE=COMPLEX;
ESTIMATOR=ML;
REPSE=BOOTSTRAP;
BOOTSTRAP=500;

Model:
AC13 by
AC1301@0.47109(b1)
AC1302@0.60345(b2)
AC1303@0.56071(b3)
AC1304@0.47109;
AC14 by
AC1401@0.47109(b1)
AC1402@0.60345(b2)
AC1403@0.56071(b3)
AC1404@0.45431;
AC15 by
AC1501@0.47109(b1)
AC1502@0.60345(b2)
AC1503@0.56071(b3)
AC1504@0.50426;
AC16 by
AC1601@0.47109(b1)
AC1602@0.60345(b2)
AC1603@0.56071(b3)
AC1604@0.54465;

[ AC1301@3.60495 AC1401@3.60495 AC1501@3.60495 AC1601@3.60495 ](IC01);
```

[AC1302@3.84737 AC1402@3.84737 AC1502@3.84737 AC1602@3.84737](IC02);
[AC1303@3.76220 AC1403@3.76220 AC1503@3.76220 AC1603@3.76220](IC03);
[AC1304@4.27465 AC1404@4.26592 AC1504@4.25802 AC1604@4.28070];

[AC13@0 AC14@0 AC15@0 AC16@0];

AC1301 with AC1401(cor1);
AC1401 with AC1501(cor1);
AC1501 with AC1601(cor1);

AC1302 with AC1402(cor2);
AC1402 with AC1502(cor2);
AC1502 with AC1602(cor2);

AC1303 with AC1403(cor3);
AC1403 with AC1503(cor3);
AC1503 with AC1603(cor3);

AC1304 with AC1404(cor4);
AC1404 with AC1504(cor4);
AC1504 with AC1604(cor4);

AC1301 with AC1302(cor12);
AC1401 with AC1402(cor12);
AC1501 with AC1502(cor12);
AC1601 with AC1602(cor12);

IC by AC13@1
AC14@1
AC15@1
AC16@1;

LSL by AC13@0
AC14@1
AC15@2
AC16@3;

QSL by AC13@0
AC14@1
AC15@4
AC16@9;

IC;
LSL;
QSL@0;

[IC LSL QSL];

IC with LSL;
IC with QSL@0;
LSL with QSL@0;

MODEL Modern:
AC13(V0);
AC14(V0);
AC15(V0);
AC16(V0);

IC;

LSL;

[IC LSL QSL](IC0 LSL0 QSL0);

MODEL Classical:
AC13(V1);
AC14(V1);
AC15(V1);
AC16(V1);

IC;
LSL;

[IC LSL QSL](IC1 LSL1 QSL1);

Model constraint:
new(T1_0 T2_0 T3_0
T1_1 T2_1 T3_1
T0D T1D T2D T3D);
T1_0=1*LSL0+1*QSL0;
T2_0=2*LSL0+4*QSL0;
T3_0=3*LSL0+9*QSL0;
T1_1=1*LSL1+1*QSL1;
T2_1=2*LSL1+4*QSL1;
T3_1=3*LSL1+9*QSL1;
T1D=T1_1-T1_0;
T2D=T2_1-T2_0;
T3D=T3_1-T3_0;

Bibliografie

Alicke, M. D., Zell, E., & Bloom, D. L. (2010). Mere categorization and the frog-pond effect. *Psychological Science*, *21*(2), 174–177.

Arens, A. K., & Watermann, R. (2015). How an early transition to high-ability secondary schools affects students' academic self-concept: Contrast effects, assimilation effects, and differential stability. *Learning and Individual Differences*, *37*, 64–71.

Arens, A. K., Yeung, A. S., Craven, R. G., & Hasselhorn, M. (2011). The twofold multidimensionality of academic self-concept: Domain specificity and separation between competence and affect components. *Journal of Educational Psychology*, *103*(4), 970–981.

Asparouhov, T., & Muthén, B. (2010). Resampling methods in Mplus for complex survey data. *Structural Equation Modeling*, *14*(4), 535–569.

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.

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.

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.

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.

Becker, M., Neumann, M., Tetzner, J., Böse, S., Knoppick, H., Maaz, K., … Lehmann, R. (2014). Is early ability grouping good for high-achieving students' psychosocial development? Effects of the transition into academically selective schools. *Journal of Educational Psychology*, *106*(2), 555.

Bernache-Assollant, I., Lacassagne, M.-F., & Braddock, J. H. (2007). Basking in reflected glory and blasting: Differences in identity-management strategies between two groups of highly identified soccer fans. *Journal of Language and Social Psychology*, *26*(4), 381–388.

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.

Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review*, *15*(1), 1–40.

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.

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.

Brunner, M., Keller, U., Dierendonck, C., Reichert, M., Ugen, S., Fischbach, A., & Martin, R. (2010). The structure of academic self-concepts revisited: The nested Marsh/Shavelson model. *Journal of Educational Psychology*, *102*(4), 964.

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.

Chanal, J. P., Marsh, H. W., Sarrazin, P. G., & Bois, J. E. (2005). Big-fish-little-pond effects on gymnastics self-concept: Social comparison processes in a physical setting. *Journal of Sport and Exercise Psychology*, *27*(1), 53–70.

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

Chiu, D., Beru, Y., Watley, E., Wubu, S., Simson, E., Kessinger, R., … Wigfield, A. (2008). Influences of math tracking on seventh-grade students' self-beliefs and social comparisons. *The Journal of Educational Research*, *102*(2), 125–136.

Chmielewski, A. K., Dumont, H., & Trautwein, U. (2013). Tracking effects depend on tracking type: An international comparison of students' mathematics self-concept. *American Educational Research Journal*, *50*(5), 925–957.

Cialdini, R. B., & Richardson, K. D. (1980). Two indirect tactics of image management: Basking and blasting. *Journal of Personality and Social Psychology*, *39*(3), 406–415.

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

Dai, D. Y., & Rinn, A. N. (2008). The big-fish-little-pond effect: What do we know and where do we go from here? *Educational Psychology Review, 20*(3), 283–317.

De Fraine, B., Van Damme, J., & Onghena, P. (2007). A longitudinal analysis of gender differences in academic self-concept and language achievement: A multivariate multilevel latent growth approach. *Contemporary Educational Psychology*, *32*(1), 132–150.

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.

Ehmke, T., Drechsel, B., & Carstensen, C. H. (2010). Effects of grade retention on achievement and self-concept in science and mathematics. *Studies in Educational Evaluation, 36*(1), 27–35.

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.

Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order latent growth models. *Methodology*, *4*(1), 22–36.

Geiser, C., Keller, B. T., & Lockhart, G. (2013). First-versus second-order latent growth curve models: some insights from latent state-trait theory. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(3), 479–503.

Geiser, C., Keller, B. T., Lockhart, G., Eid, M., Cole, A. C., & Koch, T. (2015). Distinguishing State Variability From Trait Change in Longitudinal Data: The Role of Measurement (Non)Invariance in Latent State-Trait Analyses. *Behavior Research Methods, 47*(1), 172–203.

Goetz, T., Preckel, F., Zeidner, M., & Schleyer, E. (2008). Big fish in big ponds: A multilevel analysis of test anxiety and achievement in special gifted classes. *Anxiety, Stress, & Coping, 21*(2), 185–198.

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.

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.

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.

Huang. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology*, *49*(5), 505–528.

Huguet, P., Dumas, F., Marsh, H., Régner, I., Wheeler, L., Suls, J., … Nezlek, J. (2009). Clarifying the role of social comparison in the big-fish--little-pond effect (BFLPE): An integrative study. *Journal of Personality and Social Psychology*, *97*(1), 156.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, *20*(1), 1–24.

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

King, G., & Zeng, L. (2006). The dangers of extreme counterfactuals. *Political Analysis*, *14*(2), 131–159.

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.

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

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.

Liem, G. A. D., Marsh, H. W., Martin, A. J., McInerney, D. M., & Yeung, A. S. (2013). The big-fish-little-pond effect and a national policy of within-school ability streaming: Alternative frames of reference. *American Educational Research Journal*, *50*(2), 326–370.

Liu, W. C., Wang, C. K. J., & Parkins, E. J. (2005). A longitudinal study of students' academic self-concept in a streamed setting: The Singapore context. *British Journal of Educational Psychology*, *75*(4), 567–586.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2x2 taxonomy of multilevel latent contextual models: Accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods*, *16*(4), 444.

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.

Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Educational transitions and differential learning environments: How explicit between-school tracking contributes to social inequality in educational outcomes. *Child Development Perspectives*, *2*(2), 99–106.

Marsh, H. W., Craven, R., & Debus, R. (1999). Separation of competency and affect components of

multiple dimensions of academic self-concept: A developmental perspective. *Merrill-Palmer Quarterly, 45*(1), 567–601.

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1*(2), 133–163.

Marsh, H. W., Hau, K.-T., Artelt, C., Baumert, J., & Peschar, J. L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing, 6*(4), 311–360.

Marsh, H. W., Köller, O., & Baumert, J. (2001). Reunification of East and West German school systems: Longitudinal multilevel modeling study of the big-fish-little-pond effect on academic self-concept. *American Educational Research Journal, 38*(2), 321–350.

Marsh, H. W., Kong, C.-K., & Hau, K.-T. (2000). Longitudinal multilevel models of the big-fish-little-pond effect on academic self-concept: counterbalancing contrast and reflected-glory effects in Hong Kong schools. *Journal of Personality and Social Psychology, 78*(2), 337.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*(2), 106–124.

Marsh, H. W., & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology, 81*(1), 59–77.

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K.-T., O'Mara, A. J., & Craven, R. G. (2008). The big-fish--little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review, 20*(3), 319–350.

Marsh, H. W., Trautwein, U., Lüdtke, O., Baumert, J., & Köller, O. (2007). The big-fish-little-pond effect: Persistent negative effects of selective high schools on self-concept after graduation. *American Educational Research Journal, 44*(3), 631–669.

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development, 76*(2), 397–416.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 561–614). Boston, Massachusetts: Springer.

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110*(1), 40.

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research, 79*(3), 1129–1167.

Morin, A. J. S., Arens, A. K., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric

multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 116–139.

Mulkey, L. M., Catsambis, S., Steelman, L. C., & Crain, R. L. (2005). The long-term effects of ability grouping in mathematics: A national investigation. *Social Psychology of Education*, *8*(2), 137–177.

Mussweiler, T. (2003). "Everything is relative": Comparison processes in social judgment The 2002 Jaspars Lecture. *European Journal of Social Psychology*, *33*(6), 719–733.

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, France: OECD.

Oertzen, T., Hertzog, C., Lindenberger, U., & Ghisletta, P. (2010). The effect of multiple indicators on the power to detect inter-individual differences in change. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 627–646.

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

Pinxten, M., Marsh, H. W., De Fraine, B., Van Den Noortgate, W., & Van Damme, J. (2014). Enjoying mathematics or feeling competent in mathematics? Reciprocal effects on mathematics achievement and perceived math effort expenditure. *British Journal of Educational Psychology*, *84*(1), 152–174.

Preckel, F., & Brüll, M. (2010). The benefit of being a big fish in a big pond: Contrast and assimilation effects on academic self-concept. *Learning and Individual Differences*, *20*(5), 522–531.

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

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*(2), 173–184.

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.

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 City, 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*, *70*(1), 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33–38.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, *6*(1), 34–58.

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

Salchegger, S. (2016). Selective school systems and academic self-concept: How explicit and implicit school-level tracking relate to the big-fish---little-pond effect across cultures. *Journal of Educational Psychology*, *108*(3), 405–423.

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.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods*, *13*(4), 279–313.

Seaton, M., Marsh, H. W., & Craven, R. G. (2009). Earning its place as a pan-human theory: Universality of the big-fish-little-pond effect across 41 culturally and economically diverse countries. *Journal of Educational Psychology*, *101*(2), 403–419.

Seaton, M., Marsh, H. W., Dumas, F., Huguet, P., Monteil, J.-M., Régner, I., … others. (2008). In search of the big fish: Investigating the coexistence of the big-fish-little-pond effect with the positive effects of upward comparisons. *British Journal of Social Psychology*, *47*(1), 73–103.

Seaton, M., Parker, P., Marsh, H. W., Craven, R. G., & Yeung, A. S. (2014). The reciprocal relations between self-concept, motivation and achievement: juxtaposing academic self-concept and achievement goal orientations for mathematics success. *Educational Psychology*, *34*(1), 49–72.

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research*, *46*(3), 407–441.

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

Steiner, P. M., & Cook, T. D. (2014). Matching and propensity scores. In T. D. Litle (Ed.), *The oxford handbook of quantitative methods* (pp. 237–259). Oxford, United Kingdom: Oxford University Press.

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.

Steyer, R., Mayer, A., Geiser, C., & David, A. C. (2015). A Theory of States and Traits—Revised. *Annual Review of Clinical Psychology*, *11*, 71–98.

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state–trait theory and research in personality and individual differences. *European Journal of Personality*, *13*(5), 389–408.

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

Stuart, E. A., & Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*, *44*(2), 395–406.

Thijs, J., Verkuyten, M., & Helmond, P. (2010). A further examination of the big-fish--little-pond effect perceived position in class, class size, and gender comparisons. *Sociology of Education*, *83*(4), 333–345.

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.

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.

Trautwein, U., Lüdtke, O., Marsh, H. W., & Nagy, G. (2009). Within-school social comparison: How students perceive the standing of their class predicts academic self-concept. *Journal of Educational Psychology*, *101*(4), 853–866.

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 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. (2006). Tracking and teacher satisfaction: Role of study culture and trust. *The Journal of Educational Research*, *99*(4), 247–256.

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.

Wang, Z. (2015). Examining big-fish-little-pond-effects across 49 countries: a multilevel latent variable modelling approach. *Educational Psychology*, *35*(2), 228–251.

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.

Wouters, S., De Fraine, B., Colpin, H., Van Damme, J., & Verschueren, K. (2012). The effect of track changes on the development of academic self-concept in high school: A dynamic test of the big-fish--little-pond effect. *Journal of Educational Psychology*, *104*(3), 793.

Wouters, S., Germeijs, V., Colpin, H., & Verschueren, K. (2011). Academic self-concept in high school: Predictors and effects on adjustment in higher education. *Scandinavian Journal of Psychology*, *52*(6), 586–594.

Zell, E., & Alicke, M. D. (2010). The local dominance effect in self-evaluation: Evidence and explanations. *Personality and Social Psychology Review*, *14*(4), 368–384.

Zhao, J. H., & Schafer, J. L. (2016). pan: Multiple imputation for multivariate panel or clustered data. *CRAN Package Repository*, 1–15.