SONO
STEUNPUNT
ONDERWIJS ONDERZOEK

EFFECTEN VAN VERANDERING VAN ONDERWIJSVORM OP SCHOOLSE PRESTATIES & ACADEMISCH ZELFCONCEPT

Dockx J, De Fraine B. & Vandecandelaere M.

SONO
STEUNPUNT
ONDERWIJS ONDERZOEK

EFFECTEN VAN VERANDERING VAN ONDERWIJSVORM OP SCHOOLSE PRESTATIES & ACADEMISCH ZELFCONCEPT

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

Promotor: B. De Fraine

Research paper SONO/2019.OL1.1/5

Gent, januari 2019

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

LiSO

Loopbanen in het Secundair Onderwijs

Gelieve naar deze publicatie te verwijzen als volgt:
Dockx J., De Fraine B. & Vandecandelaere M. (2019). Effecten van verandering van onderwijsvorm op schoolse prestaties & 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.

In deze publicatie wordt de mening van de auteur weergegeven en niet die van de Vlaamse overheid. De Vlaamse overheid is niet aansprakelijk voor het gebruik dat kan worden gemaakt van de opgenomen gegevens.

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

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

Beleidssamenvatting

In het Vlaamse secundair onderwijs zijn er vier onderwijsvormen: het algemeen secundair onderwijs (aso), het technisch secundair onderwijs (tso), het beroepssecundair onderwijs (bso) en het kunstsecundair onderwijs (kso). Binnen het aso wordt daarbij vaak een onderscheid gemaakt tussen klassieke talen en moderne studierichtingen. Deze onderwijsvormen worden pas formeel ingericht vanaf de tweede graad van het secundair onderwijs. 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 het keuzegedeelte en de basisopties in de eerste graad ook als voorbereiding op de onderwijsvormen in de bovenbouw. In de eerste graad bereiden het eerste leerjaar B en het beroepsvoorbereidend leerjaar voor op het bso.

Bij beleidsmakers is er discussie over mogelijke effecten van deze onderwijsvormen op schoolse prestaties. Voorstanders argumenteren dat onderwijsvormen die aansluiten op de vaardigheden en interesses van leerlingen de schoolse prestaties van leerlingen verbeteren. Tegenstanders argumenteren echter dat sociale ongelijkheid in schoolse prestaties tussen leerlingen versterkt wordt doordat de onderwijsvormen verschillen in hun mogelijkheden tot leerwinst. Deze discussie wordt verder bemoeilijkt door het hoge aantal leerlingen die doorheen het secundair onderwijs van onderwijsvorm veranderen. Daarom wordt ook de vraag gesteld wat de effecten zijn van het veranderen van onderwijsvorm.

In wetenschappelijk onderzoek wordt het inrichten van verschillende onderwijsvormen *tracking* genoemd. Er zijn diverse studies die onderwijssystemen met *vroege tracking* (categoriale onderwijssystemen) vergelijken met onderwijssystemen met *late tracking* (eerder comprehensieve onderwijssystemen). Deze studies tonen over het algemeen geen helder positief of negatief effect van *tracking* op de gemiddelde schoolse prestaties van onderwijssystemen. Een meerderheid van deze studies toont wel dat *tracking* de sociale ongelijkheid in schoolse prestaties versterkt, maar het effect is doorgaans beperkt. Studies die onderwijssystemen vergelijken beschrijven echter alleen gemiddelde verschillen tussen groepen van onderwijssystemen. De internationaal vergelijkende studies gaan dus niet in op de precieze effecten van de (gepercipieerde) hiërarchie tussen *tracks* binnen een land. Onderzoek naar de effecten van verandering van *track* zijn ook beperkt. In dit rapport willen we daarom inzoomen op de effecten van *tracking* binnen Vlaanderen, waarbij we specifiek aandacht hebben voor leerlingen die van *track* veranderen omdat zulke 'watervalloopbanen' kenmerkend zijn voor het Vlaamse secundair onderwijs.

Dit onderzoek sluit aan op twee eerdere studies naar de effecten van onderwijsvormen op schoolse prestaties (SONO/2017.OL1.1_12) en academisch zelfconcept (SONO/2017.OL1.1_13) tijdens de eerste drie jaar van het secundair onderwijs. In beide studies werd een vergelijking gemaakt tussen leerlingen die in verschillende onderwijsvormen zaten maar gelijke prestaties, sociaaleconomische achtergrond en non-cognitieve uitkomsten hadden aan de start van het secundair onderwijs. In de studie over schoolse prestatie vonden we dat leerlingen die een hoger gepercipieerde onderwijsvorm kiezen hogere prestaties voor wiskunde behaalden. In de studie over academisch zelfconcept vonden we dat leerlingen die een hoger gepercipieerde onderwijsvorm kiezen vaker een lager academisch zelfconcept hebben. De effecten van de tweede studie varieerden echter sterk naargelang de precieze vorm van academisch zelfconcept die werd gemeten en naargelang de onderwijsvormen die vergeleken werden.

Het onderzoek in dit rapport breidt de voorgaande studies uit op drie manieren: (a) in plaats van de eerste drie jaren secundair onderwijs worden de eerste vier jaren secundair onderwijs onderzocht, (b) voor schoolse prestaties wordt ook Nederlands begrijpend lezen onderzocht, en (c) de leerlingen die éénmaal van onderwijsvorm veranderen wordt ook onderzocht. Net zoals bij de voorgaande studies corrigeren we voor de initiële verschillen tussen leerlingen die in verschillende onderwijsvorm schoollopen. Voor de vergelijking tussen leerlingen die in hun onderwijsvorm blijven en leerlingen die van onderwijsvorm veranderen corrigeren we ook voor de verschillen tussen deze groepen net voor de verandering van onderwijsvorm. De concrete onderzoeksvragen zijn:

1. Wat is het effect van een onderwijsvorm op de gemiddelde leerwinst van vergelijkbare leerlingen?
2. Wat is het effect van een onderwijsvorm op de gemiddelde ontwikkeling voor academisch zelfconcept van vergelijkbare leerlingen?
3. Is er een effect van onderwijsvormverandering op de gemiddelde leerwinst van vergelijkbare leerlingen?
4. Is er een effect van onderwijsvormverandering op de gemiddelde ontwikkeling voor academisch zelfconcept van vergelijkbare leerlingen?

Voor dit onderzoek gebruiken we de gegevens van het onderzoek 'Loopbanen in het Secundair Onderwijs' (LiSO-project). De substeekproef bestaat uit 5417 leerlingen die in september 2013 startten in het secundair onderwijs. Er waren vier groepen van *tracks*: (1) klassieke talen (KT), (2) moderne wetenschappen (MW), (3) technisch onderwijs (TO) en (4) beroepsvoorbereidend onderwijs (BV). Hoewel er in het eerste jaar secundair onderwijs nog geen officiële *tracks* onderscheiden worden, sluit de studiekeuze in het eerste jaar SO wel sterk aan bij de onderwijsvormen die in de bovenbouw zullen volgen. In dit Engelstalige rapport wordt daarom wel gesproken over '*tracking*' in het eerste jaar secundair onderwijs, omdat het gaat over het groeperen van leerlingen voor een volledig schooljaar voor (quasi) alle vakken.

De steekproef is verspreid over de vier '*tracks*' in het eerste jaar als volgt: 1419 leerlingen zaten in KT, 2229 leerlingen zaten in MW, 1033 leerlingen zaten in TO en 736 leerlingen zaten in BV. Veel van deze leerlingen veranderden echter van *track* doorheen het secundair onderwijs. LiSO-scholen die kiezen voor een heterogene klassamenstelling in het eerste jaar, werden geschrapt uit de steekproef van deze studie omdat er dus niet aan *tracking* wordt gedaan. Toetsen en vragenlijsten werden afgenomen aan de start van het secundair onderwijs (september 2013), op het einde van het eerste leerjaar van de eerste graad (mei 2014), op het einde van het tweede leerjaar van de eerste graad (mei 2015), op het einde van eerste leerjaar van de tweede graad (mei 2016) en op het einde van tweede leerjaar van de tweede graad (mei 2017). Prestaties voor wiskunde werden gemeten op elk van deze vijf momenten, terwijl we voor prestaties voor Nederlands begrijpend lezen enkel werken met metingen uit september 2013 en mei 2017. Algemeen academisch zelfconcept, zelfconcept in wiskunde en zelfconcept in Nederlands werden ook gemeten op elk van deze momenten. Dit onderzoek beschrijft dus de effecten van *tracks* tijdens de eerste vier jaar van het secundair onderwijs op wiskunde, Nederlands begrijpend lezen, algemeen academisch zelfconcept, zelfconcept in wiskunde en zelfconcept in Nederlands.

Om vergelijkbare leerlingen in verschillende *tracks* te vinden gebruiken we *g-methods*. Deze methoden zijn gericht op het schatten van effecten van behandelingen (in dit onderzoek is dat de toewijzing aan een bepaalde *track*), waarbij personen van behandeling kunnen veranderen (in dit onderzoek zijn dat leerlingen die van *track* veranderen). Deze methoden staan toe om onvertekende effecten te schatten wanneer er voldoende over de achtergrond van de leerlingen gekend is. De achtergrond van leerlingen werd beschreven op basis van schoolse prestaties, sociaaleconomische achtergrond en psychosociale variabelen die gemeten waren in september 2013. Ook werd er rekening gehouden met het verschil in de evolutie in schoolse prestaties en non-cognitieve uitkomsten voor leerlingen die van *track* veranderen en leerlingen die in dezelfde *track* blijven. Om onze resultaten methode-onafhankelijk te maken vergeleken we twee *g-methods*: *de marginal structural mean model* en de *structural nested mean model*. 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. Om dezelfde reden was het enkel mogelijk om de effecten van verandering van *track* te onderzoeken voor leerlingen die eenmaal van een zogenaamde hogere *track* naar een zogenaamde lagere *track* veranderen.

Voor de eerste onderzoeksvraag vinden we voor vergelijkbare leerlingen in verschillende *tracks* dat er in *tracks* met een gemiddeld sterkere leerlinginstroom significant meer leerwinst gemaakt wordt. De effecten op wiskunde na vier jaar zijn klein voor de vergelijking KT met MW en de vergelijking MW met TO. Voor de vergelijking TO met BV is het effect groot voor wiskunde na vier jaar. De effecten op Nederlands begrijpend lezen na vier jaar zijn klein voor de vergelijking KT met MW en de vergelijking MW met TO. Voor de vergelijking TO met BV is het effect niet significant voor Nederlands begrijpend lezen na vier jaar. De resultaten zijn in de lijn van de eerdere studie over *tracks* en schoolse prestaties (OL1.1_12). Algemeen genomen bevestigen deze resultaten dat naar een zogenaamd hoger gepercipieerde *track* gaan positief is voor schoolse prestaties.

Voor de tweede onderzoeksvraag vinden we 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 eerder 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. De resultaten zijn in de lijn van de eerdere studie over *tracks* en academisch zelfconcept (OL1.1_13).

Voor de derde onderzoeksvraag vinden we dat leerlingen die veranderen van een hoger gepercipieerde *track* naar een lager gepercipieerde *track* minder leerwinst boeken (wiskunde en

Nederlands) in vergelijking met leerlingen die in de hoger gepercipieerde *track* blijven. Dit vinden we zowel voor de vergelijking KT met MW, MW met TO en TO met BV. In het algemeen presteren de leerlingen die naar een lager gepercipieerde *track* veranderen gelijk aan de leerlingen die heel hun schoolloopbaan in deze *track* waren.

Voor de vierde onderzoeksvraag vinden we dat leerlingen die veranderen van een hoger gepercipieerde *track* naar een lager gepercipieerde *track* een positiever academisch zelfconcept ontwikkelen in vergelijking met leerlingen die in de hoger gepercipieerde *track* blijven. Dit vinden we zowel voor de vergelijking MW met TO en TO met BV. De leerlingen die naar een lager gepercipieerde *track* veranderen evolueren dus naar een academisch zelfconcept dat gelijkt op de leerlingen die heel hun schoolloopbaan in deze *track* waren. Voor de vergelijking KT en MW vinden we echter dat leerlingen die van *track* veranderen eerder dalen in academisch zelfconcept. Dit is wel conform aan de eerdere bevinding dat MW een negatief effect heeft op academisch zelfconcept in vergelijking met KT. Ook hier evolueren de leerlingen dus naar een academisch zelfconcept dat gelijkt op de leerlingen die heel hun schoolloopbaan in deze *track* waren.

Een sterk punt van dit onderzoek is dat met verschillende methodes wordt nagegaan hoe vergelijkbare leerlingen zouden presteren als ze in een andere *track* zouden zitten. Met deze methodes konden we ook de effecten van eenmalige *track*verandering onderzoeken. Dit is vooral mogelijk doordat *tracking* in Vlaanderen een eigenschap heeft die niet kenmerkend is voor de meeste andere onderwijssystemen. In Vlaanderen verloopt het verdelen van leerlingen in *tracks* immers niet op basis van objectieve criteria (bijvoorbeeld een instaptoets). Hierdoor verschillen de *tracks* wel gemiddeld op het vlak van instroomniveau, maar vinden we nog steeds veel vergelijkbare leerlingen terug in verschillende *tracks*. In andere onderwijssystemen zien we dat er minder of nauwelijks vergelijkbare leerlingen zijn in verschillende tracks. Ook tussen leerlingen die eenmaal van track veranderen en leerlingen die in hun track blijven vonden we steeds voldoende vergelijkbare leerlingen.

We concluderen dat hoger gepercipieerde *tracks* doorgaans een positief effect hebben op schoolse prestaties. De effecten zijn meestal klein. Anderzijds concluderen we dat hoger gepercipieerde *tracks* doorgaans een negatief effect hebben op academisch zelfconcept. Dit is echter niet zo bij de vergelijking KT en MW, waar het net KT is dat een positief effect heeft op academisch zelfconcept. Leerlingen die van *track* veranderen verliezen voor een stuk de extra leerwinst die ze maakten in de hoger gepercipieerde *track*. Wanneer ze in de lager gepercipieerde *track* komen presteren ze na verloop van tijd gelijk met de leerlingen die reeds van het begin in deze *track* waren. Naar een lager gepercipieerde *track* gaan is doorgaans positief voor het academisch zelfconcept. Dit is echter niet zo bij verandering van KT naar MW.

De resultaten geven hoofdzakelijk weer hoe leerlingen beïnvloed worden door de huidige structuur van het secundair onderwijs. We tonen dat 'hoog mikken', een strategie die vaak gebruikt wordt bij studiekeuze, een beperkt positief effect heeft op schoolse prestaties voor de vergelijking KT en MW, en de vergelijking in MW en TO. Wanneer we de resultaten voor academisch zelfconcept hiermee vergelijken, dan lijkt er een interessante afweging te zijn. Zo blijkt dat er bij de vergelijking MW en TO, en de vergelijking TO en BV een soort '*trade-off*' te zijn tussen schoolse prestaties en academisch zelfconcept. Verder toont het onderzoek inderdaad dat leerlingen die (eenmaal) van *track* veranderen minder leerwinst maken dan leerlingen die in hun *track* blijven. De leerlingen die van *track* veranderen scoren echter niet lager dan leerlingen die reeds vanaf de start van het secundair onderwijs in dezelfde *track* zaten. Ook bij *track* verandering zien we een gelijkaardige '*trade-off*' tussen schoolse prestaties en academisch zelfconcept.

Inhoud

1 Introduction

Researchers often need to estimate the effect of a treatment on an outcome, usually distinguishing between an active treatment and a control condition. If exposure to the active treatment condition is random, as in a randomized controlled trial (RCT), the average treatment effect is equal to the observed difference between the active treatment and control group (Rosenbaum, 2002, pp. 19-70; Rubin, 1974, pp. 663-695). However, exposing respondents at random to the active treatment condition is often either practically impossible or undesirable due to a lack of external validity (Pearl, 2009, p. 260; Rubin, 1974, pp. 688-689). In this case only observational studies with nonrandom treatment exposure are available (Rosenbaum, 2002, pp. 1-17). Observational studies are usually characterized by pretreatment variables that predict both the active treatment exposure and the outcome. These pretreatment variables are called confounders. If unaccounted for, the estimated average treatment effect is partially attributable to the confounders (i.e., biased, Pearl, 2010, pp. 78-85; VanderWeele & Shpitser, 2013). Several methods can account for confounders, yielding unbiased average treatment effects when certain assumptions are met (Schafer & Kang, 2008). Hence, unbiased average treatment effects can be estimated even in the presence of confounders.

However, treatment exposure is often not fixed to a single time point, but can occur at multiple time points (Robins, 1997, pp. 69-70; Robins & Hernán, 2008, pp. 560-567). Such time-varying treatments are characterized by different treatment histories across respondents (Robins, Hernan, & Brumback, 2000, p. 151). We illustrate this in Table 1, with treatment exposure Z_{ti} for respondent *i* possible at time *t* = 1 and *t* = 2. Z_{ti} is 0 when respondent *i* is in the control condition and 1 when respondent *i* is in the active treatment condition. In Table 1 respondent 1 was never treated, respondent 2 was treated early, respondent 3 was treated late and respondent 4 was always treated. The treatment history $\bar{Z}_{2i}$ = (Z_{1i}, Z_{2i}) can be expressed as $\bar{Z}_{21}$ = (0,0) for respondent 1, $\bar{Z}_{22}$ = (1,0) for respondent 2, $\bar{Z}_{23}$ = (0,1) for respondent 3 and $\bar{Z}_{24}$ = (1,1) for respondent 4. Hence, a treatment history describes at which time points a respondent was exposed to the time-varying treatment.

Table 1
Illustration treatment histories and covariate values respondents 1, 2, 3 and 4

Respondent	Time	Treatment	Constant confounder	Time-varying confounder
i	*t*	Z_{ti}	X_{0i}	L_{ti}
1	1	0	58	29
1	2	0	58	23
2	1	1	34	19
2	2	0	34	24
3	1	0	67	11
3	2	1	67	18
4	1	1	82	23
4	2	1	82	25

While a treatment history describes a respondent's history of treatment exposure, it does not show how a time-varying treatment affects and is affected by other variables. For that purpose, a directed acyclic graph (DAG) is often used (e.g., Greenland, Pearl, & Robins, 1999). A DAG is a visualization that shows causal relationships between variables over time. In this visualization, an arrow from one variable to another variable represents a causal effect, with the causal variable always preceding the affected variable in time (i.e., directed and acyclic). Note that DAGs do not use subscript *i*, for the effects of variables in DAGs are averages across respondents. Figure 1 shows the example of Table 1 with time-varying treatment Z_t and outcome Y. X_0 predicts both the outcome Y and the treatment exposures Z_1 and Z_2. This makes X_0 a confounder of Z_t and Y. Furthermore, the treatment history Z_t affects and is affected by the time-varying covariate L_t. Because L_t also affects the outcome Y, L_t is called a time-varying confounder. X_0 and L_t are also in Table 1.

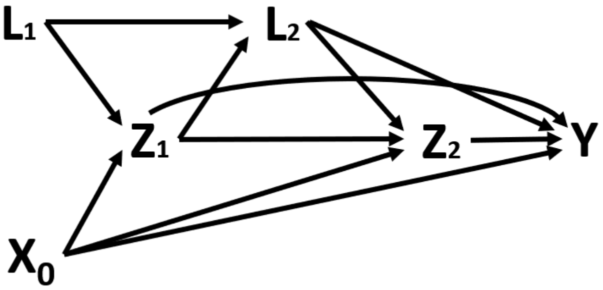

Figure 1. DAG example time-varying treatment Z_t.

A DAG describes time-varying treatments in an abstract way, so practical examples may help comprehension. As a first example, similar to Figure 1, van der Wal et al. (2010) investigated the effect of two competing dialysis treatments (Z_t) on mortality (Y) across six-month intervals (t) in a population of kidney patients. The researchers also controlled for time-varying confounders that described the severity of the kidney disease (L_t). The severity of the kidney disease affected the choice of dialysis treatment. Then, the choice of dialysis treatment affected the severity of the kidney disease at a later stage. As a second example, shown in Figure 2, VanderWeele, Hawkley, Thisted, and Cacioppo (2011) investigated the effects of loneliness (Z_t) on depressive symptoms (Y_t) across several follow-up meetings (t). Here, loneliness at follow-up 1 may affect depressive symptoms at follow-up 2, which in its turn affect both loneliness and depressive symptoms at follow-up 3. This is an example of a repeated measure that is both a time-varying confounder and the outcome. Both examples illustrate that time-varying treatments affect and are affected by time-varying confounders.

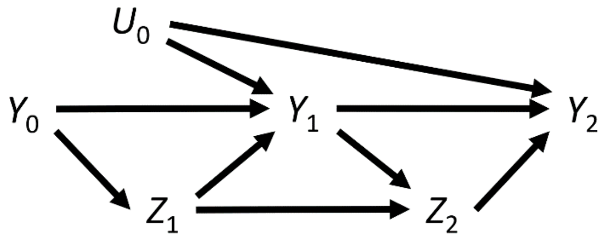

Figure 2. DAG example time-varying treatment Z_t which affects and is affected by Y_t with unmeasured colliding variable U_0.

A time-varying treatment that affects and is affected by a time-varying confounder causes a challenge in procuring unbiased average effect estimates. The main challenge is that it is wrong to simply account for a time-varying confounder as a time-fixed confounder, because this will result in bias due to 'blocking' and 'collider stratification' (Cole et al., 2009; Rosenbaum, 1984). Blocking occurs when controlling for a time-varying confounder value that follows on a treatment exposure

(Rosenbaum, 1984). Using the example of Figure 1, this would happen with a regression model with outcome *Y*, and with Z_1, Z_2, L_1 and L_2 used as covariates. By controlling for L_2, the path of Z_1 through L_2 to *Y* is controlled for, which biases the effect of Z_1 on *Y*. Bias due to collider stratification happens when conditioning on a variable which is a common effect of two independent variables (Cole et al., 2009; Whitcomb, Schisterman, Perkins, & Platt, 2009). In Figure 2, Y_1 is the common effect of variables Y_1 and U_0. When estimating the average treatment effect of Z_2, researchers may think that only controlling for Y_1 is enough, for it is the only confounder. However, conditioning on Y_1 creates a backdoor path between Z_1 and Y_2 through U_0. Through this backdoor path Z_1 now confounds Y_2 and Z_2. Accordingly, Z_1 needs to be controlled for when estimating the average treatment effect of Z_2 (If unfamiliar with colliders, Appendix A further illustrates collider stratification bias). Hence, apt methods are required when assessing time-varying treatments without introducing bias by blocking or collider stratification.

In the following sections we compare two models that can estimate unbiased average effects of time-varying treatments in the presence of time-varying confounding. Both models are derived from the g-formula, and together they are known as the G-methods. The first model, the marginal structural mean model (MSMM), is not new for educational research (Vandecandelaere, Vansteelandt, De Fraine, & Van Damme, 2016), but has been rarely used. The second, the structural nested mean model (SNMM) has to our knowledge never been used in psychological research. Both models share common goals, and the differences are not apparent. Hence, we first introduce the potential outcomes framework, wherein both models are situated, and introduce the g-formula. Afterwards both the MSMM and the SNMM are introduced and compared. We then apply both models in a simulation study and an empirical study.

1.1 Potential outcomes framework

1.1.1 The fundamental problem of causal inference

To introduce the potential outcomes framework, we use a simple example with treatment exposure being only possible at one time point and having two treatment conditions. Treatment exposure for respondent *i* is described by an indicator Z_i, with $Z_i = 0$ for the control condition and $Z_i = 1$ for the active treatment condition. The central idea (Hernán et al., 2004; Imbens & Rubin, 2015; Rubin, 1974 pp. 689- 690) is that there exist two potential outcomes for a respondent *i*: the potential outcome of being in the control condition, $Y_i(0)$, and the potential outcome of being in the active treatment condition, $Y_i(1)$. However, in practice respondent *i* only receives one of both treatments, either $Z_i = 0$ or $Z_i = 1$. The potential outcome of being in the control condition is never observed for someone in the active treatment condition, while the potential outcome of being in the active treatment condition is never observed for someone in the control condition. Formally put, $Y_i(0)|Z_i = 1$ and $Y_i(1)|Z_i = 0$ are never observed. Therefore, the individual treatment effect for the potential outcomes, $\Delta_i = Y_i(1) - Y_i(0)$, cannot be observed. Holland (1986, p. 947) calls this the fundamental problem of causal inference.

To overcome the fundamental problem of causal inference statistical theory has been developed, which allows for the estimation of an average treatment effect. Rosenbaum and Rubin (1983) showed that an average treatment effect can be estimated as $E[Y(1)|Z = 1] - E[Y(0)|Z = 0]$, if the exchangeability assumption is true. This assumption means that the average potential outcome of either the active condition or the control condition is equal across the populations in different treatment conditions, with E[Y(1)|Z = 1] = E[Y(1)] and E[Y(0)|Z = 0] = E[Y(0)]. Colloquially, this means that the respondents of one treatment condition are representative for the entire population. There are no differences in pretreatment variables between treatment conditions. However, this assumption only holds if respondents were assigned at random to the treatment conditions, as in an RCT. Put otherwise, treatment exposure is in this case independent from the potential outcomes, with $Y(1),Y(0) \perp\!\!\!\perp Z$. Hence, it is possible to overcome the fundamental problem of causal inference by estimating an average treatment effect if the exchangeability assumption is tenable.

1.1.2 The conditional exchangeability assumption

In observational studies, the exchangeability assumption is usually untenable, for the treatment conditions likely differ in pretreatment variables that affect the outcome Y. These pretreatment variables, referred to as L, are the confounders. However, Rosenbaum and Rubin (1983) showed that an average treatment effect can still be estimated, if the conditional exchangeability assumption is tenable. This assumption means that the average potential outcome of either the active treatment condition or the control condition is equal across the populations in different treatment conditions, when controlling for confounders L, with $Y(1),Y(0) \perp\!\!\!\perp Z|L$. For clarity, we further refer to L as a single variable. Colloquially, by using confounder L we can estimate what the average outcome would be if the entire population were in one treatment condition. Hence, we can estimate an average treatment effect by using the conditional exchangeability assumption.

When estimating an average treatment effect a distinction is usually made between the average treatment effect for the entire population, the ATE, and the average treatment effect for the treated population, the ATT (Imbens, 2004). The ATE is defined as E[Y(1)] - E[Y(0)] whereas the ATT is defined as E[Y(1)|Z = 1] - E[Y(0) |Z = 1]. Colloquially, the ATE is the average effect of a treatment on the entire population. The ATT is the average effect of a treatment on the treated population. For the ATT, the conditional exchangeability assumption can be slightly relaxed into the weak conditional exchangeability assumption, with $Y(0) \perp\!\!\!\perp Z|L$. Hence, for the ATT, it is enough that the average potential outcome of being in the control condition can be estimated by using confounder L. When we hereafter discuss ‘average treatment effects’, this includes both the ATE and ATT. We will specifically refer to the ATE and ATT when necessary.

The potential outcomes framework describes which assumptions should be fulfilled for estimating average treatment effects in the presence of confounders. However, an estimator that can incorporate these assumptions is required. Several estimators exist, including the Horvitz-Thompson type estimator with inverse probability treatment weights and the g-estimator. Both estimators are considered applications of the g-formula (e.g., Snowden, Rose, & Mortimer, 2011, p. 732; Vansteelandt & Keiding, 2011, p. 739). The g-formula is an unbiased estimator for the population average of a potential outcome and was introduced by Robins (1986). Applied to a time-fixed treatment the g-formula is identical to standardization. Equation 1 illustrates standardization for estimating the ATE with treatment Z, discrete confounder L and outcome Y:

$$\mathrm{E}\big(Y(Z=z)\big) = \textstyle\sum_l \ \mathrm{E}(Y|Z=z\,, L=l\,)P(L=l\,) \ (1)$$

Equation 1 consists of two main components to the right of the equality sign. The first component is the expected value of the outcome Y given treatment condition z and confounder level l. Given that this is simply the mean of observed outcomes for a stratum, it is easily estimated. The second component is the probability of confounder *L* having value *l*, which is also easily estimated. Subsequently, the two components are simply multiplied and summed across all confounder levels of *L*. This results in an unbiased estimate of the potential outcome for treatment *Z* being *z* for the population. Another way to understand standardization is that the first component is an unbiased estimate of the outcome after treatment exposure for a specific confounder level. The second component then simply weights each of these unbiased estimates according to the probabilities of the confounder distribution in the population. Hence, standardization allows us to estimate the population average potential outcome for each treatment condition and accordingly estimate the average treatment effect.

We illustrate how ATE estimation with standardization works on a simple example. In Table 2 there are three respondents in the control condition (Z=0) and three respondents in the active treatment condition (Z=1). For the former respondents we know $Y|Z$=0, whereas for the latter respondents we know $Y|Z$=1. Naïvely, we may think that subtracting the averages of both (56-52) yields the average treatment effect, 4. This is incorrect, for we note that confounder *L* is unequally distributed across the active treatment condition and control condition, with P(L=1|Z=1) =2/3 and P(L=1|Z=0) =1/3. However, in the total sample P(L=1) =1/2. With this information we can estimate E(Y(0)), which is the mean of the potential outcome of being in the control condition for the entire population, and E(Y(1)), which is the mean of the potential outcome of being in the active treatment condition. When we use standardization, E(Y(0)) = ((50+50)/2)*1/2+56*1/2 results in 53, whereas E(Y(1)) = 52*1/2+((58+58)/2)*1/2 results in 55. The difference between E(Y(1)) and E(Y(0)) is 2, hence the ATE is 2.

Table 2
Example dataset estimation of ATE and ATT using standardization

Respondent	Z	Y|Z=0	Y|Z=1	L
1	0	50	?	0
2	0	50	?	0
3	0	56	?	1
4	1	?	52	0
5	1	?	58	1
6	1	?	58	1

1.1.3 Assumptions of consistency, stable unit treatment value and positivity

We note that three, often implicit, assumptions precede the conditional exchangeability assumption: the consistency assumption, stable unit treatment value assumption and positivity assumption.

The consistency assumption entails that the observed outcome of a respondent *i* in a treatment condition should be equal to the potential outcome under that treatment condition, with Y_i^{obs} = $Y_i(z_i)$ if Z_i = z_i (Cole & Frangakis, 2009). Colloquially, this assumption connects potential outcomes to the observed outcomes, because it states that when potential outcomes are observed, they should be equal to the observed outcomes.

The stable unit treatment value assumption (Rubin, 1990) entails that the potential outcome for a respondent does not change according to the treatment assignment of another respondent. Colloquially, treatment exposure of one respondent should not interfere with another respondent.

The positivity assumption (Cole & Hernán, 2008) entails that for all observed strata of confounder combinations, both the treatment and control condition occur at least once. A stratum of a confounder combination refers to the unique combination of confounder values. This assumption is alternatively described as the experimental treatment assignment assumption (ETA). The ETA is that there should be no combination of confounder values which perfectly predicts treatment assignment, because then treatment conditions are simply incomparable due to a unique pretreatment difference. The positivity assumption in a dataset is tested by assessing the area of common support. The area of common support can either be complete (i.e., overlap for all observed strata), limited (i.e., overlap for some observed strata) or non-existing (i.e., no overlap). Causal inference is impossible in the latter case.

For brevity, during the following sections we consider the consistency and stable unit treatment value assumptions to be fulfilled. The positivity assumption will be reinterpreted for the g-formula, MSMMs and SNMMs.

1.2 Potential outcomes for time-varying treatments

The potential outcomes framework has been adapted for time time-varying treatments as well (e.g., Robins, 1997; Robins & Hernán, 2008; Robins et al., 2000). We continue using the example of the introduction (Figure 1) with $\bar{Z}_t$ describing the treatment history until time t and $\bar{L}_t$ describing the time-varying confounders histories until time t. For brevity, we drop the vector of confounders with constant values over time X_0. Now we are no longer estimating population average potential outcomes for treatment conditions but population average potential outcomes for treatment histories. We have four possible treatment histories at t = 2, and $\bar{Z}_2$ is either (0,0), (0,1), (1,0) or (1,1). In this case there are four potential outcomes, one for each treatment history, expressed as $Y(\bar{Z}_2)$.

Robins et al. (2000) showed that to estimate average treatment effects of a time-varying treatment in the presence of a time-varying confounder, a sequential conditional exchangeability assumption should be true (Robins & Hernán, 2009). This assumption entails that the average potential outcomes of being in one of either treatment histories does not differ between respondents across the compared treatment histories when controlling for $\bar{L}_t$ and $\bar{Z}_{t-1}$, with $Y(\bar{Z}_t) \perp\!\!\!\perp \bar{Z}_t | \bar{Z}_{t-1}, \bar{L}_t$. This is analogous to the conditional exchangeability assumption for a time-fixed treatment. Hence, we can estimate the average effect of a treatment history when the sequential conditional exchangeability assumption is tenable.

Just as for a time-fixed treatment, a distinction can be made between the ATE and ATT for time-varying treatments. The ATE is now defined as E[$Y(\bar{z}_t)$] - E[$Y(\overline{z'}_t)$], the average effect of a treatment history on the entire population, whereas the ATT is now defined as E[$Y(\bar{z}_t) | \bar{Z}_t = \bar{z}_t$] - E[$Y(\overline{z'}_t)$ $| \bar{Z}_t = \bar{z}_t$], the average effect of a treatment history on those in the population having had that treatment history. It should also be noted that the average effect of a treatment history is always relative to a treatment history that is considered a reference treatment history. This is usually the never treated treatment history.

Having situated time-varying treatments in the potential outcomes framework allows us to describe the g-formula (Robins & Hernán, 2008, pp. 571-572). This formula is an unbiased estimator for the population average of potential outcomes for a treatment history under the sequential conditional exchangeability assumption. For our treatment history $\bar{Z}_t$ with confounder history $\bar{L}_t$ and outcome *Y*, the g-formula for a potential value $Y(\bar{Z}_t = \bar{z}_t)$ is given in equation 2.

$$E(Y(\bar{Z}_t = \bar{z}_t)) = \sum_{\bar{l}} \left[E\left(Y \middle| \bar{Z}_t = \bar{z}_t, \bar{L}_t = \bar{l}_t\right) \prod_{t=0}^{T} P\left(\bar{L}_t = \bar{l}_t \middle| \bar{Z}_{t-1} = \bar{z}_{t-1}, \bar{L}_{t-1} = \bar{l}_{t-1}\right)\right] \text{ (2)}$$

Just as standardization for a time-fixed treatment, equation 2 consists of two main components to the right of the equation sign. The first component is now the expected value of the observed outcome *Y* given treatment history $\bar{z}_t$ and confounder history $\bar{l}_t$. Given that this is simply the mean of the observed outcomes for a stratum, it is easily estimated. The second component is a product of probabilities of confounder history $\bar{l}_t$ from time *t*=0 until *t*=*T*, given prior treatment history $\bar{z}_{t-1}$ and prior confounder history $\bar{l}_{t-1}$. The two components are estimated for all possible confounder histories of $\bar{L}_t$ and subsequently summed. The sum provides an unbiased estimate of the population average potential outcome for treatment history $\bar{z}_t$. Hence, the g-formula permits the estimation of the population average potential outcome for each treatment history and to subsequently estimate the average treatment effect.

When using the g-formula, the positivity assumption (Cole & Hernán, 2008) requires that, prior to each treatment exposure of a treatment history, all observed combinations of confounder values and prior treatment histories occur at least once.

The introduction of the g-formula makes it seem that MSMMs and SNMMs are superfluous. However, the g-formula is not well suited to finite samples. The formula requires stratification according to both treatment histories and confounder histories, and the amount of unique combinations will quickly exceed any dataset's ability to deliver stable estimates (Daniel et al., 2013, pp. 1614-1615). Hence, the positivity assumption is usually untenable.

1.3 Marginal structural mean model

The MSMM is a model for estimating the population average of the potential outcome for a treatment history, which is called the marginal mean (Hernán, Brumback, & Robins, 2000; Robins et al., 2000). The marginal mean is formally expressed as E[$Y(\bar{Z}_t)$]. Conceptually, the marginal mean of each treatment history is central in the MSMM, for the difference between two marginal means is an ATE estimate. The estimation of the marginal means consists of three steps. First, we define a structural model that links the marginal mean with a set of treatment history indicators. Second, inverse probability treatment weights are estimated, which are based on the confounder history and treatment history. Third, the resulting weights are used in a Horvitz-Thompson type estimator for estimating the parameters of the structural model. The resulting marginal means can then be used to estimate the ATE, which is unbiased if the sequential conditional exchangeability assumption holds (Robins & Hernán, 2009). We describe each of these steps in more detail in the following sections.

1.3.1 Structural model

The first step in estimating a marginal mean is to define a structural model. For treatment history $\bar{z}_2$ in our example, its structural model is described in equation 3:

$$E[Y(\overline{Z_2} = \bar{z}_2)] = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_1 z_2 \text{ (3)}$$

β_0 is the marginal mean when the population was never treated, treatment history $\bar{Z}_2$ = (0,0). z_1 and z_2 are binary indicators for allocation to the active treatment condition when *t* = 1 and *t* = 2 respectively. β_1 corresponds to the average change in $\mathrm{E}[Y(\overline{Z_2} = \bar{z}_2)]$ under treatment history $\bar{Z}_2$ = (1,0) compared to β_0. β_2 corresponds to the average change in $\mathrm{E}[Y(\overline{Z_2} = \bar{z}_2)]$ under treatment history $\bar{Z}_2$ = (0,1) compared to β_0. β_3 corresponds to the average change in $\mathrm{E}[Y(\overline{Z_2} = \bar{z}_2)]$ under treatment history $\bar{Z}_2$ = (1,1) compared to β_0+β_1+β_2. If needed, the model can be easily simplified by placing constraints. For example, β_3 can be constrained to zero if the average change resulting from treatment history $\bar{z}_2$ = (1,1) is thought to be equal to β_1+β_2. Hence, the model describes the population average potential outcome per treatment history. To procure unbiased estimates of this structural model we us a Horvitz-Thompson type estimator with inverse probability treatment weights.

1.3.2 Inverse probability treatment weighting

The main rationale of inverse probability treatment weighting (IPTW; Austin, 2011) is to correct for the unequal selection probabilities into different treatment conditions (Imbens, 2000, p. 708; Rosenbaum & Rubin, 1983). Because when these probabilities are based on the confounder and treatment histories, they summarize the pretreatment differences in confounder and treatment histories between treatment conditions. An estimator that accounts for these unequal treatment probabilities will also account for pretreatment differences between treatment conditions. Therefore, the inverse of these probabilities can be used as weights, for these weights make each treatment history resemble the total population. Equation 4 describes the formula for estimating the weights based on treatment and covariates histories:

$$W_{total} = \prod_t P[Z_t = z_t | \bar{L}_t, \bar{Z}_{t-1}]^{-1} \text{ (4)}$$

The main component of this formula is the probability of a treatment exposure Z_t, conditional on the confounder history $\bar{L}_t$ and treatment history $\bar{Z}_{t-1}$. The inverse of this probability is then computed to procure a time-specific weight. Followingly, the product of the time-specific weights for each treatment exposure for a treatment history until time *t* are computed, resulting in the total weight W_{total}. These estimated weights can get very large though, which leads to inefficiency in the parameter estimates. Therefore, we use a stabilized total weight SW_{total} by multiplying each time-specific weight with the probability of a treatment exposure z_t for treatment history $\bar{z}_{t-1}$.

$$SW_{total} = \prod_t \frac{\mathrm{P}[Z_t = z_t | \bar{Z}_{t-1}]}{\mathrm{P}[Z_t = z_t | \bar{L}_t, \bar{Z}_{t-1}]} \text{ (5)}$$

The variability in weights is reduced by using equation 5 (Hernán et al., 2000; Robins et al., 2000), which benefits efficiency. We note that when using stabilized weights the entire treatment history needs to be part of the structural model (Robins & Hernán, 2009; Talbot, Atherton, Rossi, Bacon, & Lefebvre, 2015).

Compared to the g-formula, MSMMs also weight each treatment history to resemble the total population. However, MSMMs do not need to estimate $P(\bar{L}_t = \bar{l}_t \,| \bar{Z}_{t-1} = \bar{z}_{t-1}, \bar{L}_{t-1} = \bar{l}_{t-1})$ from the g-formula, bypassing the data sparseness caused by stratification on both treatment and confounder histories (Daniel et al., 2013, pp. 1614-1615). Rather, IPTW makes each treatment history resemble the total population and achieve equal confounder distributions across treatment histories on average across samples. Accordingly, the positivity assumption when using MSMMs is less stringent compared to the g-formula. It suffices to have respondents with equal propensity scores across treatment conditions for an area of common support.

1.3.3 Application of marginal structural mean models

The inverse probability treatment weights are procured with logistic regression models. These models estimate propensity scores (PS) based on the treatment histories and confounder histories, as described in equations 4 and 5. The inverse of the PS estimates are then used weights (e.g., Robins et al., 2000, pp. 52-53). This is also the reason why propensity scores are sometimes called balancing scores, for IPTW gives balanced confounder distributions on average (Caliendo & Kopeinig, 2008, p. 1; Rosenbaum & Rubin, 1983, pp. 42-43).

Balance in the confounders must be assessed after IPTW to evaluate its success in a single sample (Austin, 2011, pp. 411-414). If balance is reached, the Horvitz-Thompson type estimator will give unbiased parameter estimates of the structural model. Accordingly, the marginal means of different treatment histories can be used to estimate ATEs.

However, while the MSMM gives unbiased ATE estimates, it has a limited efficiency for multiple reasons. A first reason is that inverse probability treatment weights achieve confounder balance on average across infinite samples. However, in a single sample imbalance exists by chance (e.g., Imai, King, & Stuart, 2008), which decreases efficiency. Second, inverse probability treatment weights are very large for respondents who have a probability of a treatment exposure close to either one or zero (e.g., Vansteelandt et al., 2014, pp. 12-13). Given that time-varying treatments have multiple treatment exposures, very large weights are likely to occur. Third, variance in weights of an earlier treatment exposure of a time-varying treatment is transmitted to all later treatment exposures (e.g., Imai & Ratkovic, 2015, p. 1013). Hence, variance in the weights increases per time point of a time-varying treatment. This decreases efficiency per time point of a time-varying treatment.

Several analysis strategies have been developed to ameliorate the limited efficiency of MSMMs. First, to reduce imbalance in a single sample, both time-fixed confounders and baseline measures of the time-varying confounders can be included as covariates in the Horvitz-Thompson type estimator (e.g., Robins et al., 2000). Any imbalance in these confounders will then be reduced, which improves efficiency. However, it will not directly improve imbalance in time-varying confounder values measured after the first treatment exposure. Second, to prevent extreme weights caused by probabilities close to either one or zero, respondents with these probabilities can be removed from the sample. Often a minimum of 0.05 and a maximum of 0.95 are used as cutoff values (e.g., Crump, Hotz, Imbens, & Mitnik, 2009). While this removal can improve efficiency, it also a causes selection bias. Third, extreme weights can be prevented by setting a limit to the weights. This is known as truncation or trimming (e.g., Lee, Lessler, & Stuart, 2011). Typically, a percentile limit is used based on the distribution of the weights before truncation. Accordingly, weights above a certain percentile are changed to the weight of the percentile limit. Typical percentile limits are 0.99, 0.98 or 0.95. While truncation indeed improves efficiency, the weight estimates are now biased.

In conclusion, MSMs give unbiased estimates of average treatment effects, but they are inefficient. Hence, several analysis strategies have been developed to improve efficiency, but they often achieve this at the cost of increased bias.

1.4 Structural nested mean model

SNMMs are a family of models for estimating how much each treatment exposure of a treatment history adds to the average effect of a treatment history (Hernán et al., 2004; Vansteelandt et al., 2014). The effect of a treatment exposure is called a 'blip'. Conceptually, the effect of a treatment history results from these blips. We emphasize that each blip is estimated as the effect of a treatment exposure for those who were exposed to the treatment, which is an ATT. The estimation of the blips consists of three steps. First, we define a structural model that links the average differences in potential outcomes between treatment histories to the blips. Second, g-estimation is used to estimate the blips based on the confounder and treatment histories. These estimates will be unbiased if the (weak) sequential conditional exchangeability assumption is true (Robins & Hernán, 2009). We describe the idea of a blip, a structural model and g-estimation in more detail in the following paragraphs.

1.4.1 Structural model

We first define a structural model where the differences between the potential outcomes of two treatment histories are linked to the blips of treatment exposures (Vansteelandt et al., 2014, p. 714). We continue with our example from the introduction, which has four possible treatment histories ($\bar{Z}_2$ = (0,0), $\bar{Z}_2$ = (1,1), $\bar{Z}_2$ = (1,0) or $\bar{Z}_2$ = (0,1)). How the two treatment exposures z_1 and z_2 of treatment history $\bar{z}_2$ affect the outcome Y can be formally described in a structural model with equations 6 and 7.

$$E\left[Y\left(\bar{Z}_2 = (z_1, 0)\right) - Y\left(\bar{Z}_2 = (0{,}0)\right) | Z_1 = z_1\right] = \psi_0 z_1 \text{ (6)}$$

$$E\left[Y\left(\bar{Z}_2 = (z_1, z_2)\right) - Y\left(\bar{Z}_2 = (z_1, 0)\right) |\ Z_1 = z_1, Z_2 = z_2\ \right] = \psi_1 z_2 + \psi_2 z_1 z_2 \text{ (7)}$$

In equation 6 blip ψ_0 describes the average change in Y if z_1=1 for the population with z_1=1. In equation 7 blip ψ_1 describes the average change in Y if z_2=1 for the population with z_2=1. The sum of blips ψ_0, ψ_1 and ψ_2 describes the average change in Y if z_1=1 and z_2=1 for the population with z_1=1 and z_2=1. The inclusion of ψ_2 allows for the treatment exposures z_1 and z_2 to have an interaction effect (Daniel et al., 2013, p. 1615). ψ_2 can be set to zero if the average change resulting from treatment history $\bar{z}_2$ = (1,1) is thought to be equal to ψ_0+ψ_1.

1.4.2 g-estimation

g-estimation was developed as an estimator for SNMMs and consists of three steps. First, equations are defined for estimating each respondent's potential outcome as if he or she is always in the control condition (Robins, Mark, & Newey, 1992; Vansteelandt et al., 2014, p. 708). Second, each respondent's potential outcome is then used in the equations that predict the treatment exposures. Under the (weak) sequential conditional exchangeability assumption, the true potential outcomes of respondents should not contribute to the prediction of treatment exposures. Third, based on the equations from step 2, we use a search grid algorithm to procure unbiased blip estimates and accordingly each respondent's unbiased potential outcome estimate. Appendix B shows a worked-example of g-estimation of a time-fixed treatment.

We continue with our example described in equation 6, equation 7 and Figure 3. For the first step, equations 6 and 7 are rearranged as if we were estimating the potential outcome of being in the control condition (i.e., not treated) for each respondent *i*. The rearrangement of equation 6 is shown in equation 8.

$$Y_{2i}^{*}\,(Z_{1i}=0, Z_{2i}=0) = y_{2i} - \psi_0 z_{1i} - \psi_1 z_{2i} - \psi_2 z_{1i} z_{2i} \quad (8)$$

The left-hand side of equation 8 is the potential outcome of respondent *i* of being in the control condition at *t*=1 and *t*=2. The right-hand side is the observed outcome Y_{2i} minus the blip ψ_0 of treatment exposure $z_{1i,}$ the blip ψ_1 of treatment exposure $z_{2i,}$ and the blip ψ_2 of having been exposed to both treatments, z_{1i} and z_{2i}. These blips are subtracted from y_{2i} based on treatment exposures at *t*=1 and *t*=2. If respondent *i* has always been in the control condition, the potential outcome is equal to the observed outcome y_{2i}. Equation 8 is used for estimating blip ψ_0. Blip ψ_1 and ψ_2 are estimated with equation 9, the rearrangement of equation 7.

$$Y_{2i}^{*}(Z_{1i}=z_{1i}, Z_{2i}=0) = y_{2i} - \psi_1 z_{2i} - \psi_2 z_{1i} z_{2i} \quad (9)$$

The left-hand side of equation 9 is the potential outcome of respondent *i* being in the control condition for the second treatment, z_{2i}=0. The right-hand side is the observed outcome y_{2i} minus the blip ψ_1 of treatment exposure z_{2i} and the blip ψ_2 of having been exposed to both treatments, z_{1i} and z_{2i}. These blips are subtracted from y_{2i} based on treatment exposure at *t*=1 and *t*=2. If respondent *i* was in the control condition at *t*=2, the potential outcome is equal to the observed outcome y_{2i}. This equation is used for estimating blip ψ_1 and ψ_2.

As equations 8 and 9 show, g-estimation starts from each respondent's potential outcome as if he or she was not treated, which is either observed or estimated. We note that this approach assumes that the blip values are constant on average across all respondents. This assumption is required to estimate each respondent's potential outcome of being in the control condition and is called the constant treatment effect assumption (see Appendix C).

In the second step of g-estimation, each respondent's potential outcome of being in the control condition is used to predict treatment exposures (Robins et al. 2000). The blips ψ_0, ψ_1 and ψ_2 in equations 8 and 9 are therefore replaced with candidate values $\psi_0^{\dagger}$, $\psi_1^{\dagger}$ and $\psi_2^{\dagger}$. When using the candidate values $Y_{2i}^{*}\,(Z_{1i}=0, Z_{2i}=0)$ becomes $H_i(\psi_0^{\dagger}, \psi_1^{\dagger}, \psi_2^{\dagger}\,)$, whereas $Y_{2i}^{*}\,(Z_{1i}, Z_{2i}=0)$ becomes $H_i(\psi_1^{\dagger}, \psi_2^{\dagger}\,)$. These respondents' potential outcomes of being in the control condition, based on the candidate values, are used to predict treatment exposure at *t*=1 and *t*=2.

$$P[Z_{1i}|H_i(\psi_0^{\dagger}, \psi_1^{\dagger}, \psi_2^{\dagger}), l_{1i}] = \alpha_0 + \alpha_1 H_i(\psi_0^{\dagger}, \psi_1^{\dagger}, \psi_2^{\dagger}) + \alpha_2 l_{1i} \quad (10)$$

$$P[Z_{2i}|H_i(\psi_1^{\dagger}, \psi_2^{\dagger}), l_{1i}, l_{2i}, z_{1i}] = \alpha_3 + \alpha_4 H_i(\psi_1^{\dagger}, \psi_2^{\dagger}) + \alpha_5 l_{1i} + \alpha_6 l_{2i} + \alpha_7 z_{1i} \quad (11)$$

Equation 10 links the probability of treatment exposure Z_{1i} to an intercept with parameter α_0, $H_i(\psi_0^{\dagger}, \psi_1^{\dagger}, \psi_2^{\dagger})$ with parameter α_1, and confounder L_i with parameter α_2. Of main interest is that α_1 is zero under the sequential conditional exchangeability assumption if $H_i(\psi_0^{\dagger}, \psi_1^{\dagger}, \psi_2^{\dagger})$ is an unbiased estimate of $Y_{2i}^{*}\,(Z_{1i}=0, Z_{1i}=0)$. Accordingly, α_1 is zero for unbiased estimates of blips ψ_0, ψ_1 and ψ_2. For the (weak) sequential conditional exchangeability assumption states that the potential outcome of being in the control condition should not predict treatment exposure after conditioning on treatment and confounder histories. Equation 11 is set up in the same way as equation 10. Hence, g-estimation uses the (weak) sequential conditional exchangeability assumption in equations predicting treatment exposures to procure unbiased estimates of respondents' potential outcomes of being in the control condition.

But how to find the values for $\psi_0^\dagger$, $\psi_1^\dagger$ and $\psi_2^\dagger$ that are unbiased estimates of ψ_0, ψ_1 and ψ_2? In the third step g-estimation uses a search grid algorithm (Robins & Hernán, 2008, p. 581), which implies that different candidate values should simply be tried until $\alpha_1 = 0$ and $\alpha_4 = 0$. While this approach is certainly possible, closed form solutions have been developed with Generalized Estimating Equations (GEEs).

Compared to the g-formula, SNMMs do not need to estimate $P(\bar{L}_t = \bar{l}_t \,|\bar{Z}_{t-1} = \bar{z}_{t-1}, \bar{L}_{t-1} = \bar{l}_{t-1})$ from the g-formula, bypassing the data sparseness caused by stratification on both treatment and confounder histories (Daniel et al., 2013, pp. 1614-1615). Rather, g-estimation directly uses the prediction of treatment probabilities to estimate average treatment effects. Accordingly, to satisfy the positivity assumption it suffices to have respondents with equal probabilities across treatment conditions for an area of common support. It is important to note that in g-estimation the closer respondents' treatment probabilities are to the extremes (0 or 1), the less they contribute to the blip estimates. Rather, when respondents have a probability of 0.50 they contribute most to the estimation of blips (Vansteelandt et al., 2014, pp. 716-718). However, under the constant treatment effect assumption, the blip estimates should be unbiased, for the blips are constant on average across all respondents. As a last note, to increase efficiency, it is possible to include the confounders as predictors in equations 8 and 9 of the first step (a form of double robustness, e.g., Moodie, Richardson, & Stephens, 2007).

1.5 Comparison marginal structural mean model and structural nested mean model

Both the MSMM and SNMM are models for estimating average treatment effects of time-varying treatments. At first glance, both models are highly similar, for the first step of each model is to define a structural model, whereas the second step of each model is the application of an estimator. However, it is precisely in their structural model and estimator that both models differ.

For the structural model, the parameters of the MSMM describe the marginal means of different treatment histories (Robins & Hernán, 2009). It is only after estimating the marginal means that the average treatment effect is estimated by subtracting two marginal means. However, the parameters of the SNMM directly estimate differences between treatment conditions at each time point of a treatment history (Vansteelandt et al., 2014, p. 714). No marginal means are estimated with the SNMM. Hence, a researcher should think carefully on whether marginal means are of interest when choosing between MSMMs and SNMMs.

The structural models of the MSMM and SNMM also differ for which population the average treatment effect is estimated. The MSMM estimates the ATE, an average effect for the entire population (Hernán et al., 2000; Robins et al., 2000). Unbiasedness depends on the (regular) sequential conditional exchangeability assumption. The SNMM estimates the ATT, an average effect for that part of the population exposed to a treatment (Hernán et al., 2004; Vansteelandt et al., 2014). Unbiasedness depends on the weak sequential conditional exchangeability assumption.

Therefore, both models have different estimands and assumptions, and, dependent on the research question, either the MSMM or the SNMM may be more appropriate.

The differences in estimators of the MSMM and SNMM also cause differences in applicability. The MSMM uses IPTW (Imbens, 2000, p. 708; Rosenbaum & Rubin, 1983), which can be done with most statistical software for estimating propensity scores and weighting. Many researchers are also familiar with sampling weights (e.g., Pfeffermann, 1993), which makes IPTW seem familiar. The reverse is true for SNMM, which uses g-estimation. Software for g-estimation is relatively rare and its application is unfamiliar (e.g., Joffe, 2012; Vansteelandt et al., 2014, p. 12). Hence, based on the ubiquity of available software and familiarity of weighting, the MSMM is more enticing to use.

However, while both estimators give unbiased average treatment effects, the MSMM lacks efficiency. This lack of efficiency is caused by single sample imbalance by chance, extremely large weights for respondents with extreme propensity scores, and weight variance of earlier treatment exposures being transmitted to later treatment exposures. While several analysis strategies can be used to improve efficiency, these often come at the cost of bias. Because SNMMs do not use IPTW, they are not subject to the efficiency problems that plague MSMMs. However, to achieve this efficiency, the SNMM leans heavily on the constant treatment effect assumption (Vansteelandt et al., 2014, pp. 716-718), which assumes that the average effect would be the same for all respondents on average.

Despite the differences in the structural model and estimators, both models account for confounder histories and treatment histories. Accordingly, it is tenable to assume that many paths resulting from colliding variables in a confounder history are accounted for in both models (Hernán et al., 2004; Pearl, 2009, pp. 16-18). Furthermore, they will not cause blocking of a treatment effect, for they only condition on confounders preceding the last treatment exposure of a treatment history (Vansteelandt, Joffe, & others, 2014, p. 728). Both are also less stringent in the positivity assumption than the g-formula, only requiring overlap in treatment probabilities. In conclusion, both the SNMMs and MSMM are appropriate methods for estimating average treatment effects of time-varying treatments. However, the differences in estimands, assumptions, efficiencies and bias between both models should be considered when applying these models.

2 Simulation study

The simulation study had three objectives. The first was to show how regression models give biased average effect estimates of time-varying treatments, and how MSMMs and SNMMs give unbiased average effect estimates. The second objective was to show how MSMMs and SNMMs differ in efficiency and bias when different analysis strategies are used. The third objective was to show how MSMMs and SNMMs are applied, this includes software usage and analysis strategies.

2.1 Data generation

As a first step, four scenarios with a time-varying treatment were simulated. Each scenario stepwise added a time-fixed confounder (X_0), a time-varying confounder (Y_0 and Y_1) and a collider (U_0) to the former scenario. There were 1.000 simulations per scenario. All continuous variables were generated with a standard normal distribution, whereas all dichotomous variables were generated with a logistic distribution and a variance of π²/3. In the following paragraphs the four scenarios are further explained and supported by DAGs in Figure 3.

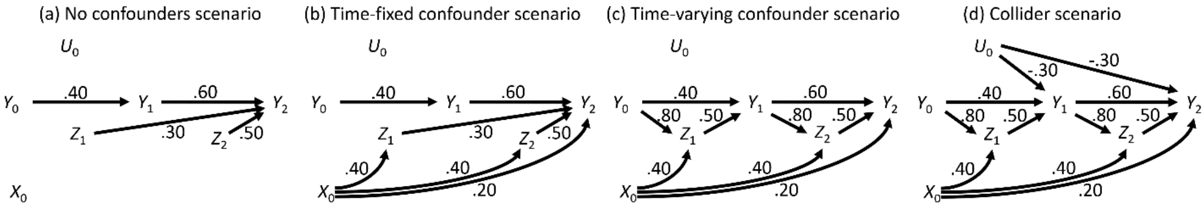
(a) No confounders scenario
(b) Time-fixed confounder scenario
(c) Time-varying confounder scenario
(d) Collider scenario
U_0
Y_0
Y_1
Y_2
Z_1
Z_2
X_0
.40
.60
.30
.50
.80
.20
-.30

Figure 3. Four scenarios simulated datasets time-varying treatment.

In the first dataset, a time-varying treatment (Z_t) with two time-points (t=1 and t=2) was generated with an average effect of 0.30 for early treatment (Δ(1,0)) and 0.50 for late treatment (Δ(0,1)). The other variables (Y_0 Y_1, X_0, and U_0) had no confounding relationship with the outcome (Y_2) in this dataset. Hence, the first dataset represented a scenario where Z_t was unconfounded with Y_2.

In the second dataset, each unit change in a time-fixed variable (X_0) changed the logit probability of the first treatment exposure (Z_1) and second treatment exposure (Z_2) with 0.40. Each unit change in the same time-fixed variable (X_0) also changed the outcome (Y_2) with 0.20. Hence, the second dataset represented a scenario where Z_t was confounded with Y_2 by time-fixed confounder X_0.

In the third dataset, we changed how the time-varying treatment (Z_t) affected the outcome (Y_2). Z_t was made to affect and to be affected by the time-varying outcome measure (Y_t). Accordingly, a unit change in $Y_\text{t-1}$ changed the logit probability of the treatment exposure (Z_t) with 0.80. A unit change in Z_t also changed Y_t with 0.50. Furthermore, a unit change in Y_0 changed Y_1 with 0.40, whereas a unit change in Y_1 changed Y_2 with 0.60. These relations caused the effect of Z_1 on Y_2 to be mediated through Y_1, which makes the average effects equal to the first and second datasets. Hence, the third dataset represented a scenario where Z_t was confounded with Y_2 by a time-fixed confounder X_0 and a time-varying confounder Y_t.

In the fourth dataset, a unit change in an unmeasured variable (U_0) changed the intermediate measure (Y_1) and the final measure (Y_2) of the outcome with 0.30. However, the unmeasured variable (U_0) was unrelated to any other variable. Hence, the fourth dataset represented the same scenario as the third dataset, but Y_1 was now a collider.

2.2 Data analyses

Five models were applied to the simulated datasets: three linear regression models, a MSMM and a SNMM. The first regression model had no covariates and served as a reference for what would happen when not accounting for confounders. The second regression model illustrated how estimates are biased when only controlling for covariates that precede the time-varying treatment. The third regression model illustrated how estimates are biased when controlling for all variables measured during the time-varying treatment. Accordingly, a MSMM and a SNMM were also applied to the simulated datasets.

Two points need mentioning before we describe these models any further. First, average treatment effects of MSMMs, SNMMs and linear regression models are defined differently. A MSMM estimates the ATE, a SNMM estimates the ATT, whereas regression models estimate conditional effects. However, because there was no interaction between the treatment effect and the confounders in any scenario, the ATE, ATT and conditional effects were equal. Second, in this simulation the average treatment effect Δ(1,1), was equal to summing Δ(1,0) and Δ(0,1). Accordingly, we decided not to discuss Δ(1,1) for brevity.

2.2.1 Linear regression models

The three linear regression models related the outcome (Y_2) with an identity link function to the treatment history ($\bar{Z}_2$) and to a linear combination of the time-fixed confounder (X_0) and the time-varying confounder (Y_t).

The first linear regression model was the no covariates model (reg. 1):

$$y_{2i} = \beta_0 + \beta_1 z_{1i} \ + \beta_2 z_{2i} \ + \varepsilon_i \ \ (12)$$

The second linear regression model was the time-fixed covariates model (reg. 2):

$$y_{2i} = \beta_0 + \beta_1 z_{1i} \ + \beta_2 z_{2i} \ + \beta_3 x_{0i} + \beta_4 y_{0i} + \varepsilon_i \ \ (13)$$

The third linear regression model was the time-varying covariates model (reg. 3):

$$y_{2i} = \beta_0 + \beta_1 z_{1i} \ + \beta_2 z_{2i} + \beta_3 x_{0i} + \beta_4 y_{0i} + \beta_5 y_{1i} + \varepsilon_i \ \ (14)$$

For estimating the parameters of the linear regression models, maximum likelihood was used in R 3.4.3. An independent and normally distributed error distribution was specified for each model.

2.2.2 Marginal structural mean model

The first step in applying the MSMM was linking the marginal mean of each treatment history to a structural model with the following equation:

$$\mathrm{E}[Y_2(z_1, z_2)] = \beta_0 \ + \beta_1 z_1 \ + \beta_2 z_2 \ (15)$$

In this equation $\mathrm{E}[Y_2(z_1, z_2)]$ was the marginal mean, whereby parameters β_1 and β_2 described the ATEs of early treatment Δ(1,0) and late treatment Δ(0,1). β_0 was equal to the never treated treatment history $\bar{Z}_2 = (0{,}0)$.

The second step was weight estimation with the following equation:

$$\overline{SW}_2 = SW_1 * SW_2 = \frac{\mathrm{P}[Z_1{=}1]}{\mathrm{P}[Z_1 = 1|y_0, x_0]} * \frac{\mathrm{P}[Z_2 = 1|z_1]}{\mathrm{P}[Z_2 = 1|y_0, y_1, x_0, z_1]} \ (16)$$

In this equation $\overline{SW}_2$ was the total weight at *t*=2, whereas SW_1 and SW_2 were the time-specific stabilized weights a *t*=1 and *t*=2. SW_1. The probabilities were estimated with logistic regression models.

For estimating the parameters of the structural model, a linear regression model with maximum likelihood estimation was used. The total weights $\overline{SW}_2$ were incorporated in the estimation procedure. The models were estimated in R 3.4.3.

2.2.3 Structural nested mean model

The first step in applying the SNMM was linking the blips of each treatment exposure to a structural model with the following equations:

$$\mathrm{E}[Y_2(z_1,0) - Y_2(0,0)|z_1] = \psi_0 z_1 \text{ (17)}$$

$$\mathrm{E}\left[Y_2(z_1,z_2) - Y_2(z_1,0)|z_1,z_2\ \right] = \psi_1 z_2 \text{ (18)}$$

Equation 17 was the ATT of early treatment Δ(1,0), described by blip ψ_0, whereas equation 18 was the ATT of late treatment Δ(0,1), described by blip ψ_1.

The second step in applying the SNMM was g-estimation with the following equations:

$$\mathrm{logit}(\mathrm{P}\left[Z_1 = 1\middle|H\left(\psi_0^\dagger,\psi_1^\dagger\right),x_0,y_0\right]) = \alpha_0 + \alpha_1 H\left(\psi_0^\dagger,\psi_1^\dagger\right) + \alpha_2 x_0 + \alpha_3 y_0 \text{ (19)}$$

$$\mathrm{logit}\left(\mathrm{P}\left[Z_2 = 1\middle|H\left(\psi_1^\dagger\right),x_0,y_0,y_1,z_0\right]\right) = \alpha_4 + \alpha_5 H\left(\psi_1^\dagger\right) + \alpha_6 x_0 + \alpha_7 y_0 + \alpha_8 y_1 + \alpha_9 z_0 \text{ (20)}$$

In equation 19 $H(\psi_0^\dagger\ ,\psi_1^\dagger)$ was the observed outcome Y_2 minus candidate values for the blips ψ_0 and ψ_1. In equation 20 $H(\psi_1^\dagger)$ was the observed outcome Y_2 minus candidate values for the blip ψ_1.

g-estimation of the structural model was achieved by using the '*DTRreg*' package in R 3.4.3, which uses an optimization algorithm for finding the blips.

2.2.4 Bias and variance average effect estimates

Bias and variance were assessed for $\widehat{\Delta}$(1,0) and $\widehat{\Delta}$(0,1). Bias was estimated by subtracting the generated average effect of a treatment history (Δ) with the Monte Carlo mean ($\overline{\Delta}$) of the average effect estimates ($\widehat{\Delta}$). The Monte Carlo mean was calculated as:

$$\overline{\Delta} = \frac{\sum_{i=1}^{r} \widehat{\Delta}_i}{r} \text{ (21)}$$

where *r* was the number of simulations and $\widehat{\Delta}_i$ was the estimate of an average treatment effect of a single replication. The variance in the average effect estimates was calculated as:

$$Var_{\Delta} = \frac{\sum_{i=1}^{r} (\widehat{\Delta}_i - \overline{\Delta})^2}{r-1} \text{ (22)}$$

2.3 Results comparing regression models, MSMM and SNMM

Table 3 and Figures 4, 5, 6 and 7 show the bias and variance of $\widehat{\Delta}$(1,0) and $\widehat{\Delta}$(0,1) when using the different models. Whether and why these models were biased when applied to the four simulated datasets is described in the following paragraphs.

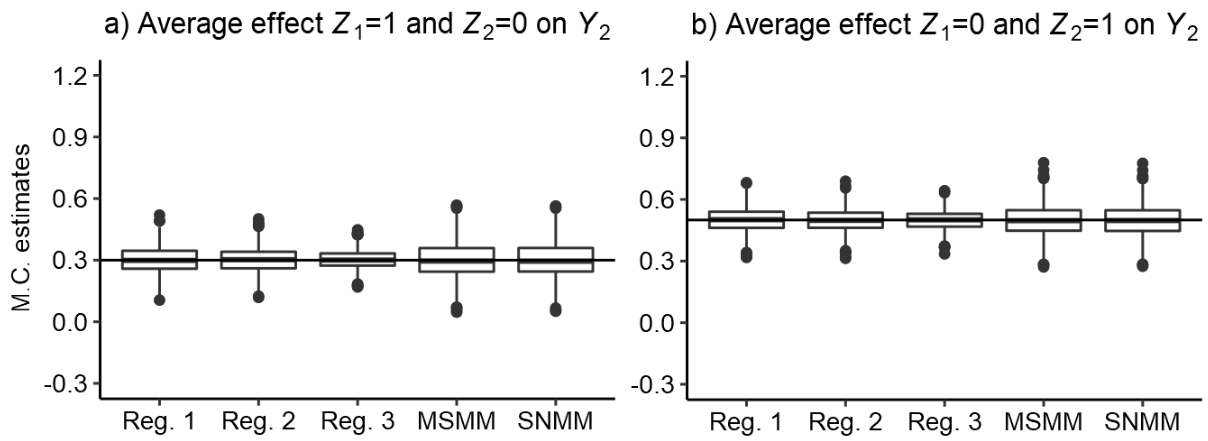

Figure 4. Boxplots Monte Carlo (M.C.) estimates average effects of treatment histories on outcome Y_2 in no confounders scenario. The horizontal line in each panel shows the true average effect.

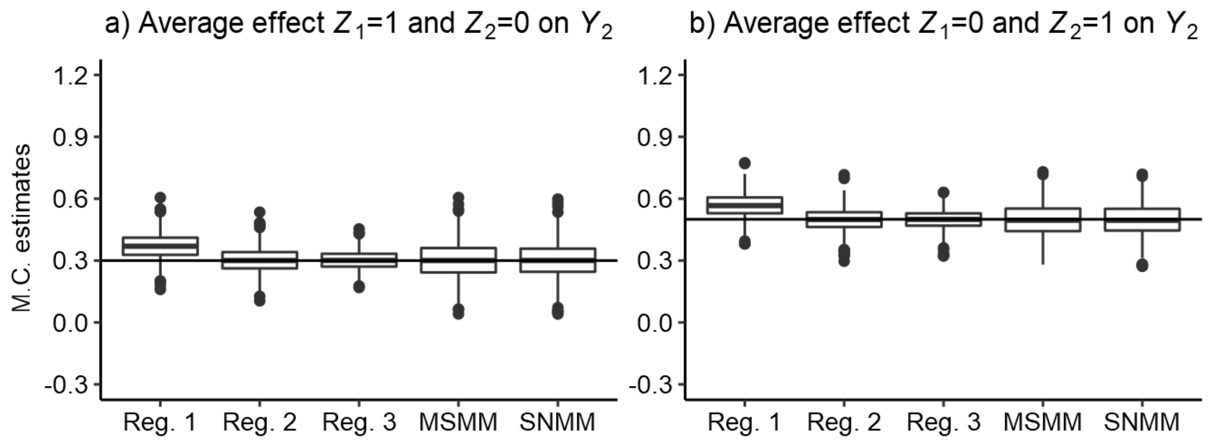

Figure 5. Boxplots Monte Carlo (M.C.) estimates of average effects on outcome Y_2 in time-fixed confounder scenario. The horizontal line in each panel shows the true average effect.

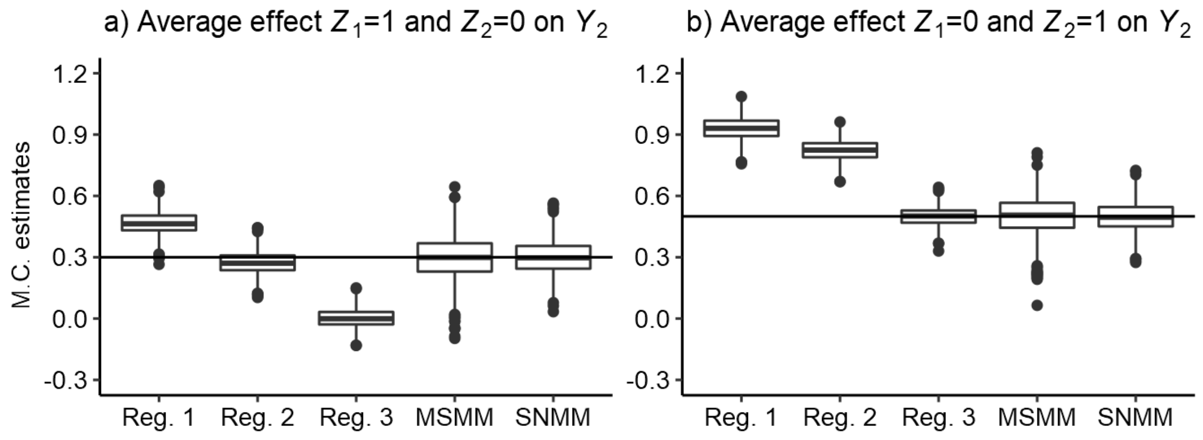

Figure 6. Boxplots Monte Carlo (M.C.) estimates of average effects on outcome *Y*2 in time-varying confounder scenario. The horizontal line in each panel shows the true average effect.

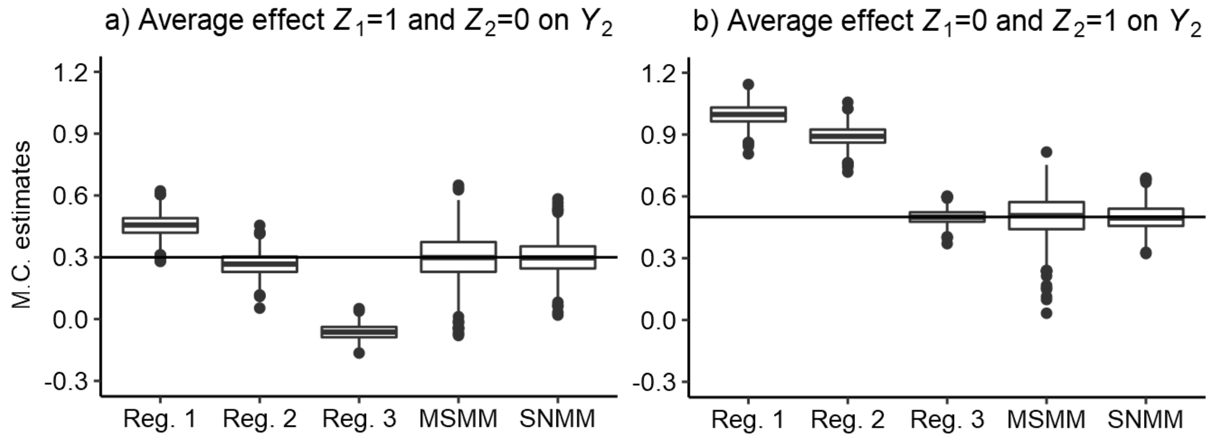

Figure 7. Boxplots Monte Carlo (M.C.) estimates of average effects on outcome *Y*2 in collider scenario. The horizontal line in each panel shows the true average effect.

Table 3

Comparison of bias, variance and MSE average effect estimates of treatment histories on outcome *Y*2 in 5 models.

	Reg. 1		Reg. 2		Reg. 3		MSMM		SNMM	
	$\widehat{\Delta}$(1,0)	$\widehat{\Delta}$(0,1)	$\widehat{\Delta}$(1,0)	$\widehat{\Delta}$(0,1)	$\widehat{\Delta}$(1,0)	$\widehat{\Delta}$(0,1)	$\widehat{\Delta}$(1,0)	$\widehat{\Delta}$(0,1)	$\widehat{\Delta}$(1,0)	$\widehat{\Delta}$(0,1)
No confounders dataset (1)										
Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SD	0.06	0.06	0.05	0.05	0.05	0.09	0.08	0.09	0.08	0.06
MSE	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.00
Time-fixed confounder dataset (2)										
Bias	0.07	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SD	0.06	0.06	0.06	0.05	0.05	0.04	0.09	0.08	0.08	0.08
MSE	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01
Time-varying confounder dataset (3)										
Bias	0.17	0.43	-0.03	0.32	-0.30	0.00	0.00	0.00	0.00	0.00
SD	0.05	0.05	0.06	0.05	0.04	0.04	0.11	0.10	0.08	0.07
MSE	0.03	0.19	0.00	0.11	0.09	0.00	0.01	0.01	0.01	0.00
Collider dataset (4)										
Bias	0.16	0.50	-0.03	0.39	-0.36	0.00	0.00	0.00	0.00	0.00
SD	0.05	0.05	0.05	0.05	0.04	0.03	0.12	0.10	0.08	0.06
MSE	0.03	0.25	0.00	0.16	0.13	0.00	0.01	0.01	0.01	0.00

Note. Reg. 1 = No confounders regression model; Reg. 2 = Time-fixed confounders regression model; Reg. 3 = Time-varying confounders regression model; MSMM = Marginal structural mean model; SNMM = Structural nested mean model; $\widehat{\Delta}$(1,0) = Average effect treatment history (1,0); $\widehat{\Delta}$(0,1) = Average effect treatment history (0,1); SD = Standard deviation; MSE = Mean squared error

The no covariates regression model gives unbiased estimates for the no confounders dataset (1). However, the model is biased for the time-fixed confounders dataset (2), time-varying confounders dataset (3) and collider dataset (4). This bias is caused by not accounting for the confounding effects of the time-fixed confounder X_0, time-varying confounder Y_t and collider U_0.

The time-fixed covariates regression model gives unbiased estimates for the no confounders dataset (1) and time-fixed confounders dataset (2). However, the model is biased for the time-varying confounders dataset (3) and collider dataset (4). This bias is caused by not accounting for the confounding effects of the time-varying confounder Y_t and collider U_0.

The time-varying covariates regression model gives unbiased estimates for the no confounders dataset (1) and time-fixed confounders dataset (2). It also gives unbiased estimates for Δ(0,1) for the time-varying confounders dataset (3) and collider dataset (4). However, the model is biased for Δ(1,0) for the time-varying confounders dataset (3) and collider dataset (4). This bias is caused by using Y_1 as a covariate, blocking the effect of Z_1 on Y_2, which biases $\hat{\Delta}(1,0)$. Furthermore, the model does not account for the confounding effect of collider U_0.

The MSMM and SNMM are unbiased when applied to the no confounders dataset (1), time-fixed confounders dataset (2), time-varying confounders dataset (3) and collider dataset (4).

In conclusion, we illustrated that when estimating average treatment effects of a time-varying treatment with a time-varying confounder and a collider, suitable methods are required. However, the variance of the estimates when using these suitable methods, the MSMM and the SNMM, are larger compared to the regression models. This requires a further inquiry on analysis strategies to decrease variance in the estimates when using MSMMs and SNMMs.

2.4 Results extreme propensity score removal, truncation and doubly robust estimation

Table 4 and Figure 8 show the bias and variance of $\hat{\Delta}(1,0)$ and $\hat{\Delta}(0,1)$ when using 99th or 95th percentile truncation, removing extreme propensity scores (PS<0.05, PS>0.95), and including the baseline confounders when using MSMMs. Table 4 and Figure 8 also show the bias and variance of $\hat{\Delta}(1,0)$ and $\hat{\Delta}(0,1)$ when using SNMMs, with and without double robust estimation. The results are discussed in the following paragraphs.

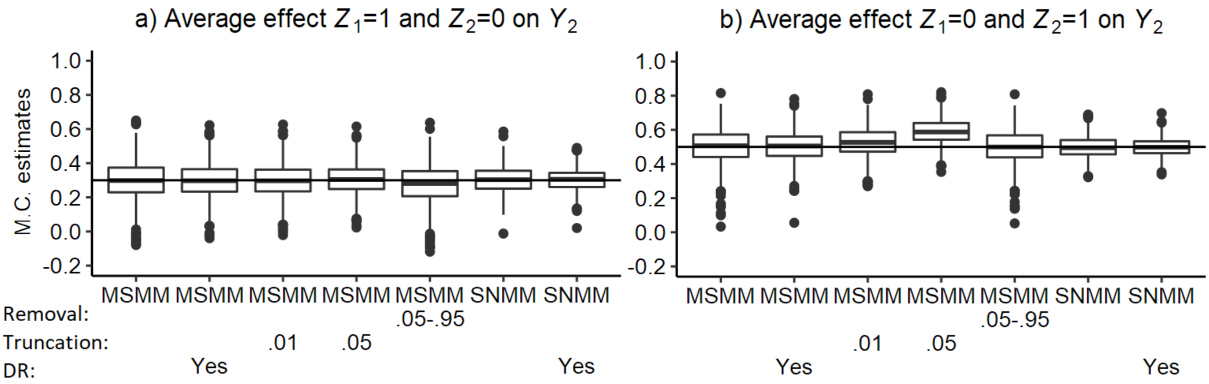

Figure 8. Boxplots Monte Carlo (M.C.) estimates of average effects on outcome Y_2 across MSMMs and SNMMs using different analysis strategies. The horizontal line in each panel shows the true average effect.

Table 4
Comparison of bias, variance and MSE average effect estimates of treatment histories on outcome Y_2 using MSMMs and SNMMs with either extreme propensity removal, truncation or doubly robust estimation.

Model	MSMM		MSMM		MSMM		MSMM		MSMM	
Cut-off									.05-.95	
Trunc.					.01		.05			
DR			Yes							
	$\widehat{\Delta}(1,0)$	$\widehat{\Delta}(0,1)$	$\widehat{\Delta}(1,0)$	$\widehat{\Delta}(0,1)$	$\widehat{\Delta}(1,0)$	$\widehat{\Delta}(0,1)$	$\widehat{\Delta}(1,0)$	$\widehat{\Delta}(0,1)$	$\widehat{\Delta}(1,0)$	$\widehat{\Delta}(0,1)$
Bias	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.09	-0.02	0.00
SD	0.12	0.10	0.10	0.09	0.10	0.09	0.09	0.08	0.12	0.10
MSE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Model	SNMM		SNMM							
Cut-off										
Trunc.										
DR			Yes							
	$\widehat{\Delta}(1,0)$	$\widehat{\Delta}(0,1)$	$\widehat{\Delta}(1,0)$	$\widehat{\Delta}(0,1)$						
Bias	0.00	0.00	0.00	0.00						
SD	0.07	0.06	0.06	0.05						
MSE	0.01	0.01	0.01	0.01						

Note. MSMM = Marginal structural mean model; SNMM = Structural nested mean model; Cut-off .05-.95 = Removal of sampling unit when probability less than 0.05 or 0.95 for first treatment exposure or second treatment exposure; Trunc. .01 = Truncation of time-specific weight to 0.01; Trunc. .05 = Truncation of time-specific weight to 0.05; DR = doubly robust estimation for SNMM and including baseline confounders MSMM; $\widehat{\Delta}(1,0)$ = Average effect treatment history (1,0); $\widehat{\Delta}(0,1)$ = Average effect treatment history (0,1); SD = Standard deviation; MSE = Mean squared error

First, we discuss the MSMMs. When we used 99th percentile truncation the variance of $\widehat{\Delta}(1,0)$ decreased 28.24%, whereas the variance of $\widehat{\Delta}(0,1)$ decreased 30.79%. $\widehat{\Delta}(1,0)$ was now biased with -1.20%, whereas $\widehat{\Delta}(0,1)$ was biased 5.57%. When we used 95th percentile truncation the variance of $\widehat{\Delta}(1,0)$ decreased 43.38%, whereas the variance of $\widehat{\Delta}(0,1)$ decreased 47.26%. $\widehat{\Delta}(1,0)$ was now biased 1.68%, whereas $\widehat{\Delta}(0,1)$ was biased 17.94%. When we included baseline confounders in the structural model the variance of $\widehat{\Delta}(1,0)$ decreased 27.69%, whereas the variance of $\widehat{\Delta}(0,1)$ decreased 31.08%. $\widehat{\Delta}(1,0)$ was now biased -0.67%, whereas $\widehat{\Delta}(0,1)$ was biased 0.94%. When we removed extreme propensity scores (PS<0.05 or PS >0.95) the variance of $\widehat{\Delta}(1,0)$ increased 2.40%, whereas the variance of $\widehat{\Delta}(0,1)$ increased 2.70%. $\widehat{\Delta}(1,0)$ was now biased -8.11%, whereas $\widehat{\Delta}(0,1)$ was biased -0.17%.

Lastly, we applied SNMMs with and without doubly robust estimation. Doubly robust estimation caused no tangible for early treatment $\Delta(1,0)$ or late treatment $\Delta(0,1)$. The variance decreased 25.23% for $\widehat{\Delta}(1,0)$ and 39.49% for late treatment $\widehat{\Delta}(0,1)$.

In conclusion, the results illustrate that both truncation and to inclusion of baseline confounders in the structural model substantially reduce variance in the estimates of MSMMs. However, truncation achieves this at the cost of increased bias, especially 95th percentile truncation. This trade-off does not occur for the inclusion of baseline confounders in the structural model, it only reduced variance. Surprisingly, removing extreme propensity scores does not decrease the variance, whereas bias does increase. That this removal does not decrease variance may be specific to our simulated dataset, because the overlap between treatment conditions is quite high. For SNMMs doubly robust estimation reduces variance with no apparent trade-off. Comparing the MSMMs with the SNMMs shows that the latter always has a lower variance. However, the MSMMs are always less biased than the regression models in the former section.

3 Empirical study

3.1 Introduction

The empirical study on track effects had three objectives. The first was to show how the conditional exchangeability assumption and positivity assumption are used in an applied study. The second objective was to introduce an often-encountered type of time-varying treatment, a monotonic time-varying treatment. The third objective was to show how MSMMs and SNMMs are applied, this includes software usage and analysis strategies.

The research question of this empirical study was how being in a higher track affects academic performance and academic self-concepts. Our first hypothesis was that higher track allocation benefits students' academic performance. Our second hypothesis was that higher track allocation negatively affects students' academic self-concept. For the hypotheses of track change, we based ourselves on the prevailing thought that track changes have negative effects. Accordingly, our third hypothesis was that downward track change negatively affects academic performance. Our fourth hypothesis was that downward track change negatively affects academic self-concept. In the following section the sample and methods are described in more detail.

3.2 Methodology

3.2.1 Sample

This study used a sample of 6328 students who were in the first year of secondary education in September 2013, using data from the longitudinal LiSO-project (LiSO-project, 2018). A subsample was taken where students from de-tracked schools, students in a sports or arts program and students who were redoing their first year in secondary education were removed. Hence, our final subsample consisted of 5417 students. At the start of secondary education 1419 students were in the classical track, 2229 were in the modern track, 1033 students were in the technical track and 736 students were in the vocational track. There were five measurement occasions: the start of secondary education September 2013 (T0), the end of the first year of secondary education May 2014 (T1), the end of the second year of secondary education May 2015 (T2), the end of the third year of secondary education May 2016 (T3), and the end of the fourth year of secondary education May 2017 (T4).

3.2.2 Variables

3.2.2.1 Treatment variable

The treatment variable was track allocation to the lower track. Lower track allocation was the active treatment condition (Z_t = 1), whereas higher track allocation was the control condition (Z_t = 0). It was not possible to compare nonconsecutive tracks due to the absence of comparable students (this will be further explained in the section 'Area of common support'). Three pairwise comparisons were made: the classical track with the modern track, the modern track with the technical track, and the technical track with the vocational track. For each comparison of two tracks, five track allocation histories were distinguished: staying in the higher track continuously (0,0,0,0), starting in the higher track but changing to the lower track after T3 (0,0,0,1), starting in the higher track but changing to the lower track after T2 (0,0,1,1), starting in the higher track but changing to the lower track after T1 (0,1,1,1) and staying in the lower track continuously (1,1,1,1). These track allocation histories are also shown in Table 5 for each comparison.

Table 5

Overview track allocation histories of classical and modern track comparison, modern and technical track comparison, and technical and vocational track comparison.

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

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

3.2.2.2 Outcomes

The first outcome of interest was student academic performance. Two measures for academic performance were used: academic performance in mathematics and Dutch reading comprehension. The second outcome of interest was student academic self-concept. Three measures for academic self-concept were used: general academic self-concept, self-concept in mathematics and self-concept in Dutch. So, five student outcomes were studied in total.

Mathematics performance was measured at T0, T1, T2, T3 and T4. The number of items ranged from 32 to 42 and encompassed following domains: algebra, geometry, geometric calculation, and data- and information processing. The tests were based on the educational goals set by the government and are considered valid measurements in the Flemish context. Each test had a mix of multiple-choice and open-ended questions. Item Response Theory was used during test development for vertical equating, to test for differential item functioning and to select items in a broad range of difficulty parameters with high discrimination parameters (Embretson & Reise, 2000). Ability scores were estimated using Warm's weighted likelihood estimation (Warm, 1989). The classical, modern and technical track at T0 were used as a reference group with a mean of 100 and standard deviation of 10. The Cronbach's Alphas of the tests ranged from 0.81 to 0.87.

Dutch reading comprehension was developed and measured in a similar way as mathematics. However, it was measured only at T0 and T4, with the number of items ranging from 32 to 43. For a correct answer, students were required to either perform a descriptive analysis, structure text elements or give a personal judgement on different text elements. The texts also varied in complexity of sentence formulation, complexity of text structuring, extensiveness of the texts, extent of visual support, familiarity of content and if the text was more practical or abstract. This was based on the educational goals set by the government and the tests are considered valid in this context. The classical, modern and technical track at T0 were used as a reference group with a mean of 100 and standard deviation of 10. The Cronbach's Alpha's of the tests were 0.80 and 0.83.

Academic self-concepts were measured at T0, T1, T2, T3 and T4. General academic self-concept was measured based on four items, which were translated to Dutch from the short form of general academic self-concept of the Self-Description Questionnaire II (Marsh, Ellis, Parada, Richards, & Heubeck, 2005). The items of the domain specific measures (self-concept for mathematics and self-concept for Dutch) are based on the Self-Description Questionnaire III (Marsh & O'Neill, 1984), and are reduced to four of the original six items. Multiple group factor analyses in Mplus 8 were used to investigate measurement invariance across measurement occasions (Baumgartner & Steenkamp, 2006; Cheung & Rensvold, 2002). The cutoff criteria from Hu and Bentler (1999) were used for fit indices CFI, TLI and RMSEA. Factor analyses showed that a one-factor structure with assumed measurement invariance fitted well for general academic self-concept (CFI= .99, TLI = .99, RMSEA = .04). Factor analyses showed that a one-factor structure with assumed measurement invariance fitted well for academic self-concept in mathematics. However, we were required to let two pairs of indicators freely correlate to achieve satisfactory model fit (CFI= .97, TLI = .97, RMSEA = .07). For academic self-concept in Dutch we did not find entirely satisfactory model fit for a one-factor structure with assumed measurement invariance (CFI= .94, TLI = .95, RMSEA = .09). However, there was no obvious way to improve model fit and we decided to use this as our final factor model. Composite reliabilities ranged from 0.77 to 0.84 for general academic self-concept, from 0.89 to 0.92 for self-concept in mathematics and from 0.85 to 0.88 for self-concept in Dutch. Maximum a posteriori (MAP) estimation was used for student factor scores with a zero mean and unit variance in the whole sample at T0.

3.2.2.3 Independent variables

To satisfy the sequential conditional exchangeability assumption (Robins & Hernán, 2009) we needed to control for a selection of covariates. This selection of covariates should make it tenable that any effect of the track allocation history on academic performance could not be ascribed to (time-varying) confounders. We based our selection on the literature for causal inference of time-fixed exposures (e.g., Stuart, 2010). Most authors agree that all variables that predict both the treatment (i.e., track allocation and track change in this study) and the outcome (i.e., academic performance and academic self-concept) should be included. If sample size allows it, all variables related to the outcome should also be included (e.g., Brookhart et al., 2006; Myers et al., 2011; Stuart, 2010). Hence, we included those variables that predict academic performance and academic self-concept. Table 6 gives a brief overview of these variables.

Table 6
Descriptions, information sources and properties of time-varying and time-fixed confounder measures.

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

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

3.2.3 Area of common support

Before estimating weights, we assessed the area of common support between the different track allocation histories of a track comparison. This should be considered a test of the positivity assumption. Accordingly, by assessing the area of common support it is tested whether comparable students exist across different track trajectories. This was achieved by first estimating the propensity score of being continuously in the higher track. The overlap in the resulting propensity scores was then used to assess the area of common support (Steiner & Cook, 2013). Note that it was immediately clear that there was only a substantial area of common support between pairs of tracks that are consecutive in the hierarchy of tracks. Hence, as described in the 'Treatment variable' section, we made three pairwise comparisons of tracks. For each comparison a limited area of common support was found. Hence, when applying MSMMs, truncation or extreme propensity score removal should be used. We used a minimum of 0.05 and a maximum of 0.95 as cutoff values for extreme propensity score removal (e.g., Crump, Hotz, Imbens, & Mitnik, 2009). We also used 99th percentile truncation (e.g., Lee, Lessler, & Stuart, 2011). This step will be shown first in the results section. Because SNMMs automatically account for which treatment probabilities the area of common support is strongest, no students were removed for having extreme propensity scores when using SNMMs.

3.2.4 Application marginal structural mean model

3.2.4.1 Structural model

The first step in applying the MSMM was linking the marginal mean of each treatment history to a structural model. For the marginal mean at measurement occasion T4 we have the following equation:

$$\mathrm{E}[Y_4(z_1,z_2,z_3,z_4)] = \beta_0 \ + \beta_1 z_1 z_2 z_3 z_4 \ + \beta_2(1-z_1)z_2 z_3 z_4 \ + \beta_3(1-z_1)(1-z_2)z_3 z_4 \ + \beta_4(1-z_1)(1-z_2)(1-z_3)z_4 + \boldsymbol{\beta}_5 \boldsymbol{x}_0 + \boldsymbol{\beta}_6 \boldsymbol{l}_0 \quad (23)$$

In this equation $\mathrm{E}[Y_4(z_1,z_2,z_3,z_4)]$ is the marginal mean, whereby parameters β_1, β_2, β_3 and β_4 respectively describe the ATEs of track allocation histories (1,1,1,1), (0,1,1,1), (0,0,1,1) and (0,0,0,1). β_0 is equal to the track allocation history of always being in the higher track (0,0,0,0). $\boldsymbol{\beta}_5$ and $\boldsymbol{\beta}_6$ are vectors of parameters describing the average change in marginal means for time fixed covariates in vector $\boldsymbol{x}_0$ and the baseline measurements for time-varying confounders in vector $\boldsymbol{l}_0$. Note that equivalent structural models were specified for the marginal means at measurement occasions T1, T2 and T3.

For estimation we used GEEs with the Newton-Raphson algorithm. We specified an independent correlation matrix (Liang & Zeger, 1986) and estimated sandwich standard errors (Joffe, Ten Have, Feldman, & Kimmel, 2004). Note that either specifying a non-independent correlation structure or using mixed methods would make outcome measurements dependent on future treatment exposures, which would cause bias (Robins, Hernan, & Brumback, 2000, p. 554). Inverse probability treatment weights were incorporated into the model estimation. To examine differences between track allocation histories at each time point, contrasts were tested using one degree of freedom Wald tests (Kuhn, Weston, Wing, & Forester, 2016). GEE-models were estimated using the geepack 1.2-1 package (Højsgaard, Halekoh, & Yan, 2006) in R 3.4.3. Cohen's *d* was used for effect size interpretation (Cohen, 1977).

3.2.4.2 Inverse probability treatment weighting

The second step was weight estimation with the following equations for T4:

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

$$\frac{\mathrm{P}[Z_1=1]}{\mathrm{P}[Z_1=1|y_0,\boldsymbol{x}_0,\boldsymbol{l}_0]} * \frac{\mathrm{P}[Z_2=1|Z_1=0]}{\mathrm{P}[Z_2=1|y_0,y_1,\boldsymbol{x}_0,\boldsymbol{l}_0,\boldsymbol{l}_1,Z_1=0]} * \frac{\mathrm{P}[Z_3=1|Z_2=0]}{\mathrm{P}[Z_3=1|y_0,y_2,\boldsymbol{x}_0,\boldsymbol{l}_0,\boldsymbol{l}_2,Z_2=0]} * \frac{\mathrm{P}[Z_4=1|Z_3=0]}{\mathrm{P}[Z_4=1|y_0,y_3,\boldsymbol{x}_0,\boldsymbol{l}_0,\boldsymbol{l}_3,Z_3=0]} \text{ (24)}$$

In this equation $\overline{SW_4}$ was the total weight at *t*=4, and SW_1, SW_2, SW_3 and SW_4 were the time-specific stabilized weights. When we compare equation 24 to the weight estimation of the simulation study, an important difference is that the estimated probabilities are conditional on either $Z_1 = 0$, $Z_2 = 0$ or $Z_3 = 0$. This change is caused by track allocation being a monotonic time-varying treatment. Therefore, reweighting is only necessary for students who can still change to a lower track at time *t* (i.e., students who were in the higher track at time *t*-1). When the time-specific weight was not estimated for a student (i.e., student was already in the lower track), it was replaced by value one (i.e., the total weight is unchanged). Note that, for the total weights at measurement occasions T1, T2 and T3, it was only necessary to multiply time-specific weights until that measurement occasion.

To attain stable weights estimates we chose not to use the entire history of time-varying covariates but only their values at T0 and their values at time *t*-1 for track allocation at time *t*. We found that this was enough to balance the entire confounder history and stabilize the weight estimation. As mentioned, extreme propensity score removal and 99th percentile truncation were used as well. We used the twang 1.5 package (Ridgeway, McCaffrey, Morral, Griffin, & Burgette, 2017) in R 3.4.3 for weight estimation. To estimate the propensity score we used generalized boosted regression models (GBMs, McCaffrey, Ridgeway, & Morral, 2004), a nonparametric regression technique whith an automated and data-adaptive algorithm to predict propensity scores. Given that it automatically optimizes the predictive power of a set of covariates for the propensity scores, it is considered best practice for propensity score estimation (e.g., Stuart, 2010). However, for the technical and vocational track comparison we used the covariates as linear predictors, for this led to a better balance.

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

3.2.5 Application structural nested mean model

3.2.5.1 Structural model

The first step in applying the SNMM was linking the blips of each treatment exposure to a structural model with the following equations:

$$\mathrm{E}[Y_4(z_1) - Y_4(0)|z_1] = \psi_0 z_1 \text{ (25)}$$

$$\mathrm{E}\big[Y_4(0,z_2) - Y_4(0,0)|z_2\,\big] = \psi_1 z_2 \text{ (26)}$$

$$\mathrm{E}\big[Y_4(0,0,z_3) - Y_4(0,0,0,)|z_3\,\big] = \psi_2 z_3 \text{ (27)}$$

$$\mathrm{E}\big[Y_4(0,0,0,z_4) - Y_4(0,0,0,0)|z_4\,\big] = \psi_3 z_4 \text{ (28)}$$

These equations were the ATTs of the different treatment exposures (i.e., initial track allocation to the lower and track change to the lower track), which are described by the blips ψ_0, ψ_1, ψ_2 and ψ_3. g-estimation of the structural model was achieved by using the '*DTRreg*' package in R 3.4.3 (Wallace, Moodie, & Stephens, 2017), using an optimization algorithm for finding the blips. Cohen's *d* was used for effect size interpretation (Cohen, 1977).

g-estimation

The second step in applying the SNMM was g-estimation with the following equations:

$$\text{logit}(\text{P}[Z_1 = 1 | H(\psi_0^\dagger, \psi_1^\dagger, \psi_2^\dagger, \psi_3^\dagger), x_0, l_0]) = \alpha_0 + \alpha_1 H(\psi_0^\dagger, \psi_1^\dagger, \psi_2^\dagger, \psi_3^\dagger) + \alpha_2 x_0 + \alpha_3 l_0 \quad (29)$$

$$\text{logit}(\text{P}[Z_2 = 1 | H(\psi_1^\dagger, \psi_2^\dagger, \psi_3^\dagger), l_0, l_1, x_0, z_1]) = \alpha_4 + \alpha_5 H(\psi_1^\dagger, \psi_2^\dagger, \psi_3^\dagger) + \alpha_6 x_0 + \alpha_7 l_0 + \alpha_8 l_1 + \alpha_9 z_1 \quad (30)$$

$$\text{logit}(\text{P}[Z_3 = 1 | H(\psi_2^\dagger, \psi_3^\dagger), l_0, l_2, x_0, z_2]) = \alpha_{10} + \alpha_{11} H(\psi_2^\dagger, \psi_3^\dagger) + \alpha_{12} x_0 + \alpha_{13} l_0 + \alpha_{14} l_2 + \alpha_{15} z_2 \quad (31)$$

$$\text{logit}(\text{P}[Z_4 = 1 | H(\psi_3^\dagger), l_0, l_3, x_0, z_3]) = \alpha_{16} + \alpha_{17} H(\psi_3^\dagger) + \alpha_{18} x_0 + \alpha_{19} l_0 + \alpha_{20} l_3 + \alpha_{21} z_3 \quad (32)$$

In these equations $H(\psi_0^\dagger, \psi_1^\dagger, \psi_2^\dagger, \psi_3^\dagger)$, $H(\psi_1^\dagger, \psi_2^\dagger, \psi_3^\dagger)$, $H(\psi_2^\dagger, \psi_3^\dagger)$ and $H(\psi_3^\dagger\ )$ each represent the observed outcome Y_4 minus candidate values for the blips ψ_0, ψ_1, ψ_2 and ψ_3.

3.2.6 Missing values

In our sample, 10.89% of the data was missing on average. We used multiple imputation by chained equations to attain unbiased and efficient estimates for missing values in the covariates and outcomes (Schafer & Graham, 2002) with the package mice 2.30 (van Buuren & Groothuis-Oudshoorn, 2011) in R 3.4.3. The incorporation of both confounders and outcomes as predictors should result in unbiased and efficient estimates under the missing at random assumption (MAR, e.g., Moodie, Delaney, Lefebvre, & Platt, 2008). We estimated ten imputed datasets, and combined the results as described by Rubin's (1987) rules. The average relative efficiency attained for the outcomes at T4 (when missingness was highest) was 98.28% on average for the MSMMs and 97.97% on average for the SNMMs.

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

3.3 Results

3.3.1 Balance after weighting

Figure 9 shows the overlap between the track allocation histories of each pairwise track comparison. Substantial overlap existed, but for each comparison very low propensity scores and very high propensity scores occurred. Accordingly, by applying the cutoff values of .05 and .95, students were removed from the dataset. The resulting sample sizes for all comparisons for each time point are shown in Table 5.

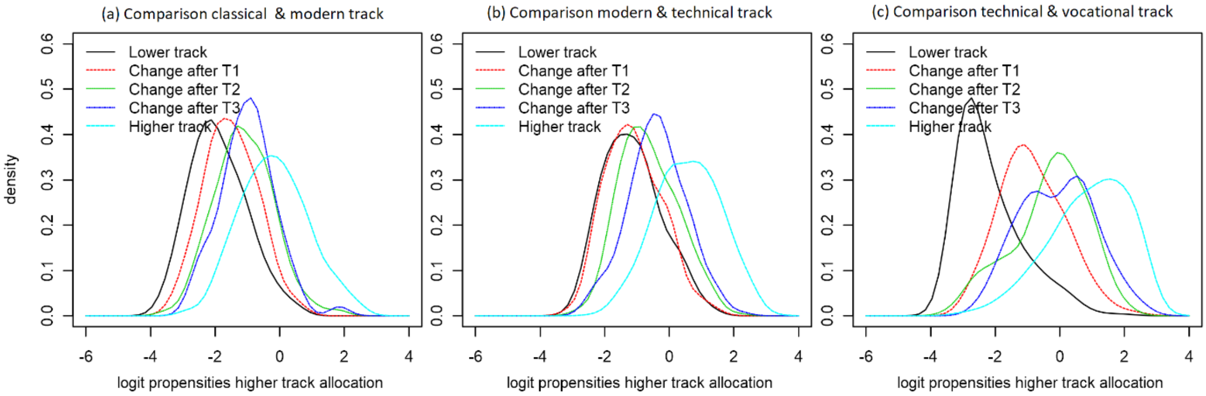

Figure 9. Density plots logit propensities of lower track allocation.

The minimum, maximum and mean SMDs after applying the MSMM weights are shown in Table 7. Figures 10, 11 and 12 show the SMDs for each covariate, before and after weighting. In general, satisfactory balance is achieved. However, we do note that for the time-varying measure of the general academic self-concept, while the bias is severely reduced, the cutoff is not always reached. At T4 there is some imbalance remaining for the modern and technical track comparison, and the technical and vocational track comparison. Generally, though, balance across the confounders is reached.

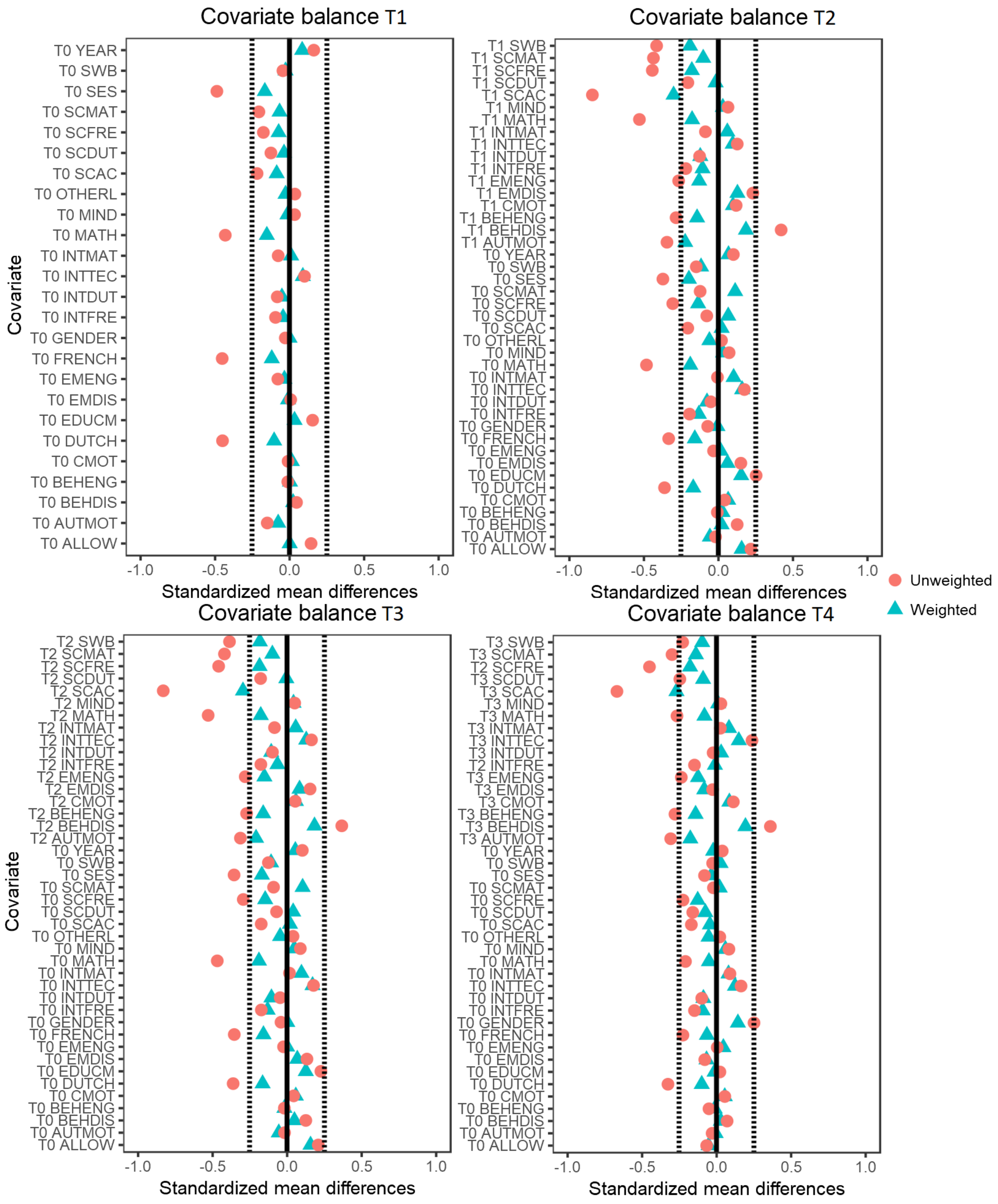

Figure 10. Classical and modern track comparison: SMDs before and after weighting for covariate(s) (histories) at T1, T2, T3 and T4.

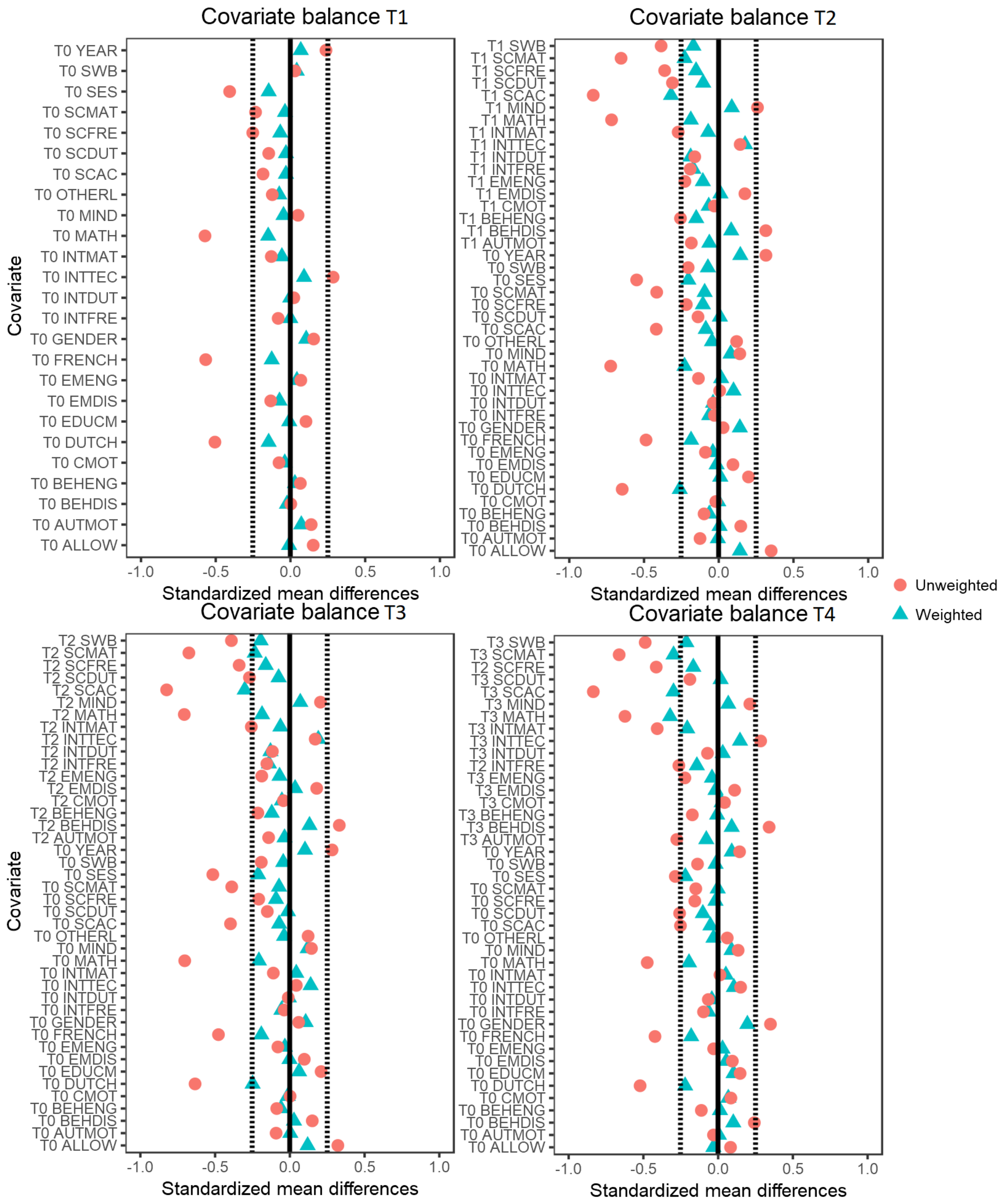

Figure 11. Modern and technical track comparison: SMDs before and after weighting for covariate(s) (histories) at T1, T2, T3 and T4.

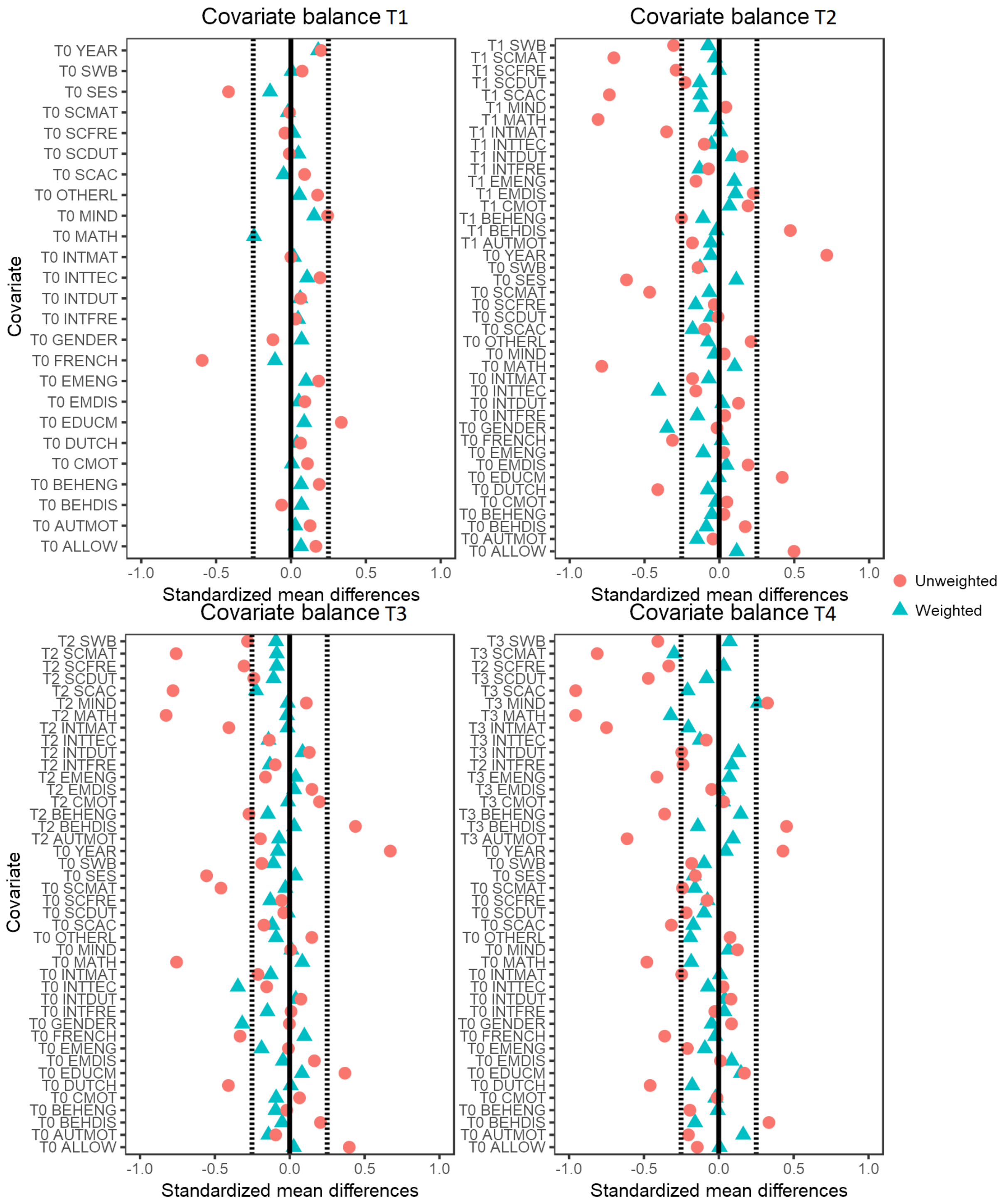

Figure 12. Technical and vocational track comparison: SMDs before and after weighting for covariate(s) (histories) at T1, T2, T3 and T4.

Table 7
SMDs after weighting

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

3.3.2 Analysis of track effects

In what follows, we will discuss the results for each of the five dependent variables. And for each dependent variable, we describe the results for the three track comparisons. Every track comparison is made with the two methods: the MSMM and the SNMM.

3.3.2.1 Academic performance in mathematics

For academic performance in mathematics, the growth estimated with the MSMM for each of the three track comparisons is shown in Figure 13. Table 8 shows the MSMM ATE estimates and SNMM ATT estimates. For brevity, we only discuss the results at T4.

For the classical and modern track comparison, students who are continuously in the higher track make significantly more learning gains compared to students who are continuously in the lower track, with a small effect size. Furthermore, students who are continuously in the higher track generally make significantly more learning gains compared to students who changed from the higher to lower track, with small to nonmeaningful effect sizes. However, at no point do the students who change from the higher to lower track make significantly less learning gains compared to students who are continuously in the lower track.

For the modern and technical track comparison, students who are continuously in the higher track make significantly more learning gains compared to students who are continuously in the lower track, with a small effect size. Furthermore, students who are continuously in the higher track make significantly more learning gains compared to students who changed from the higher to lower track, with small to medium effect sizes. However, at no point do the students who change from the higher to lower track make significantly less learning gains compared to students who are continuously in the lower track.

For the technical and vocational track comparison, students who are continuously in the higher track make significantly more learning gains compared to students who are continuously in the lower track, with a medium effect size. Furthermore, students who are continuously in the higher track generally make significantly more learning gains compared to students who changed from the higher to lower track, with small to medium effect sizes. Students who change from the higher to lower track make significantly more learning gains compared to students who are continuously in the lower track.

Table 8

Contrast estimates academic performance in mathematics across track comparisons using the MSMMs and SNMMs.

	Classical & modern track comparison		Modern & technical track comparison		Technical & vocational track comparison	
	MSMM	SNMM	MSMM	SNMM	MSMM	SNMM
	d (SE)	*d* (SE)	*d* (SE)	*d* (SE)	*d* (SE)	*d* (SE)
T1						
High – low	1.67*	1.74*	1.47*	1.59*	4.57*	4.60*
	(0.31)	(0.31)	(0.36)	(0.38)	(0.77)	(0.63)
T2						
High – low	1.49*	0.65	4.01*	3.61*	1.11	2.11*
	(0.41)	(0.41)	(0.39)	(0.44)	(0.70)	(0.69)
High – T1 Change	0.37	0.03	3.25*	3.66*	0.88	-2.76*
	(0.64)	(0.59)	(0.75)	(0.74)	(1.36)	(0.86)
T3						
High – low	2.01*	1.60*	1.81*	1.45*	6.14*	6.63*
	(0.38)	(0.36)	(0.39)	(0.41)	(0.70)	(0.62)
High – T1 Change	2.89*	1.93*	4.00*	3.01*	2.78*	1.65*
	(0.55)	(0.57)	(0.66)	(0.66)	(0.81)	(0.82)
High – T2 Change	0.92	0.82	1.98*	1.03*	1.60	1.30
	(0.53)	(0.42)	(0.55)	(0.51)	(1.59)	(0.99)
T4						
High – low	3.24*	2.96*	3.54*	2.92*	6.60*	7.44*
	(0.46)	(0.40)	(0.50)	(0.51)	(1.14)	(0.95)
High – T1 Change	2.38*	1.58*	5.29*	4.47*	3.24*	1.83
	(0.60)	(0.59)	(0.76)	(0.72)	(1.29)	(1.02)
High – T2 Change	2.02*	1.54*	3.62*	2.04*	5.13*	2.53*
	(0.58)	(0.48)	(0.64)	(0.64)	(1.31)	(1.18)
High – T3 Change	0.85	2.13*	3.44*	3.14*	5.42*	3.53*
	(1.16)	(0.72)	(0.87)	(0.73)	(2.18)	(1.43)

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

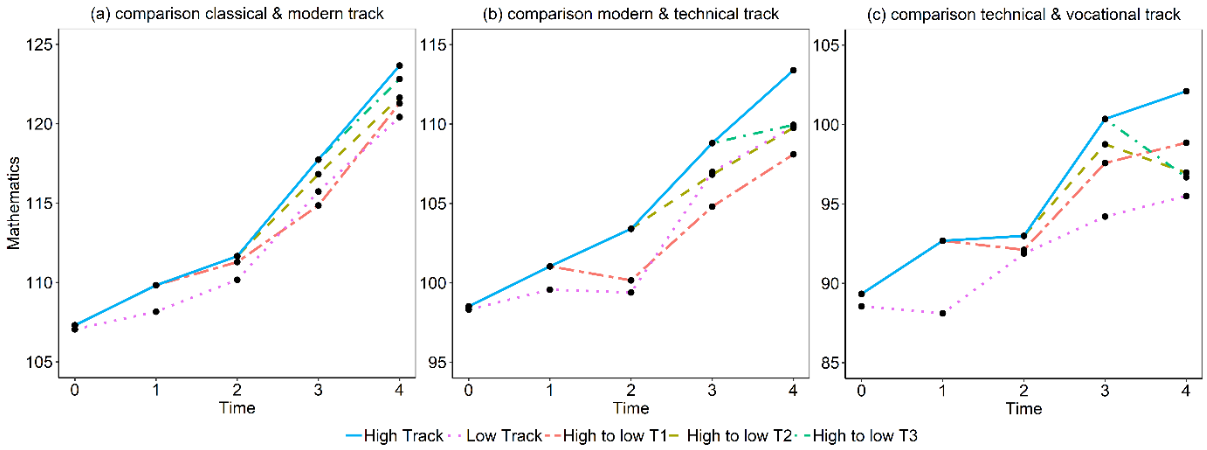

Figure 13. Mathematics performance estimated with MSMMs for each track allocation history of each track comparison.

3.3.2.2 Academic performance in Dutch reading comprehension

Table 9 shows the MSMM ATE estimates and SNMM ATT estimates for Dutch reading comprehension for each of the three track comparisons. Note that, because Dutch reading comprehension was only measured at T0 and T4, no visual representation of growth across school years is given.

For the classical and modern track comparison, students who are continuously in the higher track make significantly more learning gains compared to students who are continuously in the lower track, with a small effect size. Furthermore, students who are continuously in the higher track generally make significantly more learning gains compared to students who changed from the higher to lower track, with small to medium effect sizes. However, at no point do the students who change from the higher to lower track make significantly less learning gains compared to students who are continuously in the lower track.

For the modern and technical track comparison, students who are continuously in the higher track make significantly more learning gains compared to students who are continuously in the lower track, with a small effect size. Furthermore, students who are continuously in the higher track generally make significantly more learning gains compared to students who changed from the higher to lower track, with small to nonmeaningful effect sizes. However, at no point do the students who change from the higher to lower track make significantly less learning gains compared to students who are continuously in the lower track.

For the technical and vocational track comparison, there is no difference in the learning gains between students who are continuously in the higher track and students who are continuously in the lower track. Furthermore, the students who changed from the higher to lower track have no differences in learning gains as well.

Table 9
Contrast estimates academic performance in Dutch reading comprehension across track comparisons using the MSMMs and SNMMs.

	Classical & modern track comparison		Modern & technical track comparison		Technical & vocational track comparison	
	MSMM *d* (SE)	SNMM *d* (SE)	MSMM *d* (SE)	SNMM *d* (SE)	MSMM *d* (SE)	SNMM *d* (SE)
T4						
High – low	2.87*	3.13*	2.42*	2.60*	1.51	1.10
	(0.56)	(0.51)	(0.57)	(0.60)	(1.66)	(1.05)
High – T1 Change	4.12*	3.70*	2.09*	2.97*	-1.16	-0.38
	(0.77)	(0.80)	(0.87)	(1.01)	(1.54)	(1.33)
High – T2 Change	3.02*	2.99*	2.55*	2.41*	0.62	0.92
	(0.72)	(0.68)	(0.75)	(0.81)	(1.74)	(1.41)
High – T3 Change	2.52	2.04	1.61	1.39	0.59	0.67
	(1.43)	(1.12)	(1.08)	(0.91)	(2.45)	(1.88)

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

3.3.2.3 General academic self-concept

For general academic self-concept, the development estimated with the MSMM for each of the three track comparisons is shown in Figure 14. Table 10 shows the MSMM ATE estimates and SNMM ATT estimates. For brevity, we only discuss the results at T4.

For the classical and modern track comparison, students who are continuously in the higher track develop a significantly higher self-concept compared to students who are continuously in the lower track. The MSMM ATE has a small effect size, whereas the SNMM ATT has a nonmeaningful effect size. Furthermore, the MSMM shows that students who are continuously in the higher track generally develop a significantly higher self-concept compared to students who changed from the higher to lower track, with generally small effect sizes. However, the SNMM ATTs have nonmeaningful effect sizes. The students who change from the higher to lower track do not differ in development from the students who are continuously in the lower track.

For the modern and technical track comparison, students who are continuously in the higher track develop a significantly lower self-concept compared to students who are continuously in the lower track, with a small effect size. Furthermore, students who are continuously in the higher track generally develop a significantly lower self-concept compared to students who changed from the higher to lower track, with small to nonmeaningful effect sizes. The students who change from the higher to lower track do not differ in development from the students who are continuously in the lower track.

For the technical and vocational track comparison, students who are continuously in the higher track develop a significantly lower self-concept compared to students who are continuously in the lower track, with a small to medium effect size. Furthermore, students who are continuously in the higher track generally develop a significantly lower self-concept compared to students who changed from the higher to lower track, with medium to large effect sizes. Students who change from the higher to lower track generally develop a higher self-concept compared to students who are continuously in the lower track.Table 10

Contrast estimates general academic self-concept across track comparisons using the MSMMs and SNMMs.

	Classical & modern track comparison		Modern & technical track comparison		Technical & vocational track comparison	
	MSMM	SNMM	MSMM	SNMM	MSMM	SNMM
	d (SE)	*d* (SE)	*d* (SE)	*d* (SE)	*d* (SE)	*d* (SE)
T1						
High – low	-0.12*	-0.12*	-0.19*	-0.21*	-0.80*	-0.79*
	(0.04)	(0.04)	(0.04)	(0.05)	(0.08)	(0.09)
T2						
High – low	0.06	-0.03	-0.13*	-0.23*	-0.55*	-0.61*
	(0.04)	(0.04)	(0.05)	(0.05)	(0.09)	(0.09)
High – T1 Change	0.06	-0.11	-0.01	-0.20*	-0.60*	-1.06*
	(0.07)	(0.06)	(0.07)	(0.08)	(0.21)	(0.12)
T3						
High – low	0.22*	0.03	-0.33*	-0.40*	-0.50*	-0.55*
	(0.06)	(0.05)	(0.06)	(0.06)	(0.10)	(0.10)
High – T1 Change	0.37*	-0.03	-0.22*	-0.35*	-0.67*	-0.87*
	(0.09)	(0.08)	(0.09)	(0.08)	(0.12)	(0.13)
High – T2 Change	0.08	-0.04	-0.19*	-0.37*	-1.00*	-1.13*
	(0.08)	(0.06)	(0.08)	(0.08)	(0.19)	(0.15)
T4						
High – low	0.26*	0.15*	-0.29*	-0.32*	-0.48*	-0.55*
	(0.05)	(0.05)	(0.06)	(0.06)	(0.11)	(0.10)
High – T1 Change	0.31*	0.15*	-0.23*	-0.22*	-0.83*	-0.95*
	(0.08)	(0.08)	(0.09)	(0.09)	(0.13)	(0.10)
High – T2 Change	0.25*	0.06	-0.03	-0.19*	-0.69*	-0.89*
	(0.07)	(0.06)	(0.07)	(0.08)	(0.15)	(0.14)
High – T3 Change	0.18	0.05	-0.19*	-0.37*	-1.05*	-1.08*
	(0.14)	(0.11)	(0.09)	(0.09)	(0.22)	(0.15)

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

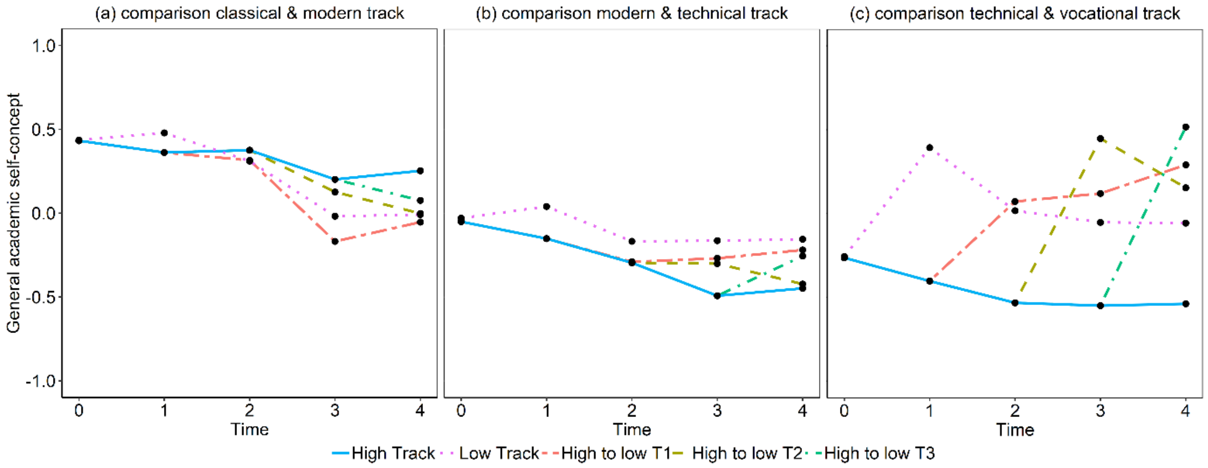

Figure 14. Development general academic self-concept estimated with MSMMs for each track allocation history of each track comparison.

3.3.2.4 Self-concept in mathematics

For self-concept in mathematics, the development estimated with the MSMM for each of the three track comparisons is shown in Figure 15. Table 11 shows the MSMM ATE estimates and SNMM ATT estimates. For brevity, we only discuss the results at T4.

For the classical and modern track comparison, students who are continuously in the higher track and students who are continuously in the lower track do not develop differently for self-concept in mathematics. Furthermore, the MSMMs show that students who are continuously in the higher track develop a significantly higher self-concept in mathematics compared to students who changed from the higher to lower track. However, the effects sizes are nonmeaningful. The SNMM shows no significant difference between students who are continuously in the higher track and students who changed from the higher to lower track. The effect sizes are also nonmeaningful.

For the modern and technical track comparison, students who are continuously in the higher track develop a significantly lower self-concept in mathematics compared to students who are continuously in the lower track, with a small effect size. For students who changed from the higher to lower track, the MSMM and SNMM mostly agree. Students who changed from the higher to lower track after T1 do not differ significantly in self-concept in mathematics from the students who are continuously in the higher track. Students who changed from the higher to lower track after T3 develop a significantly higher self-concept in mathematics compared to students who are continuously in the higher track, with a medium to large effect size. However, for students who changed from the higher to lower track after T2 the SNMM predict a significantly higher academic self-concept in mathematics with a small effect size, whereas the MSMM predicts a nonmeaningful effect size.

For the technical and vocational track comparison, students who are continuously in the higher track develop a significantly lower self-concept compared to students who are continuously in the lower track, with a small effect size. Students who are continuously in the higher track generally develop a significantly lower self-concept compared to students who change from the higher to lower track.

Generally, students who change from the higher to lower track develop a significantly lower self-concept compared students who are continuously in the lower track.

Table 11
Contrast estimates self-concept in mathematics across track comparisons using the MSMMs and SNMMs.

	Classical & modern track comparison		Modern & technical track comparison		Technical & vocational track comparison	
	MSMM	SNMM	MSMM	SNMM	MSMM	SNMM
	d (SE)	*d* (SE)	*d* (SE)	*d* (SE)	*d* (SE)	*d* (SE)
T1						
High – low	-0.03	-0.05	-0.06	-0.09*	-0.83*	-0.86*
	(0.04)	(0.04)	(0.04)	(0.04)	(0.07)	(0.08)
T2						
High – low	0.04	-0.10*	-0.12*	-0.23*	-0.64*	-0.80*
	(0.04)	(0.04)	(0.05)	(0.05)	(0.08)	(0.09)
High – T1 Change	0.06	-0.08	-0.03	-0.18*	-0.79*	-1.22*
	(0.07)	(0.06)	(0.07)	(0.07)	(0.15)	(0.11)
T3						
High – low	0.11*	-0.05	-0.28*	-0.38*	-0.43*	-0.53*
	(0.05)	(0.05)	(0.06)	(0.05)	(0.10)	(0.10)
High – T1 Change	0.35*	0.03	-0.18	-0.28*	-0.92*	-1.00*
	(0.08)	(0.07)	(0.09)	(0.09)	(0.10)	(0.11)
High – T2 Change	0.07	-0.03	-0.23*	-0.42*	-0.73*	-0.76*
	(0.07)	(0.05)	(0.07)	(0.07)	(0.17)	(0.12)
T4						
High – low	0.08	-0.01	-0.32*	-0.39*	-0.20*	-0.32*
	(0.06)	(0.05)	(0.06)	(0.05)	(0.10)	(0.10)
High – T1 Change	0.19*	0.02	-0.11	-0.15	-0.73*	-0.84*
	(0.08)	(0.08)	(0.08)	(0.08)	(0.10)	(0.11)
High – T2 Change	0.17*	-0.02	-0.12	-0.33*	-0.17	-0.46*
	(0.07)	(0.07)	(0.07)	(0.07)	(0.12)	(0.10)
High – T3 Change	0.07	-0.06	-0.63*	-0.85*	-0.21	-0.31*
	(0.14)	(0.10)	(0.08)	(0.08)	(0.19)	(0.13)

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

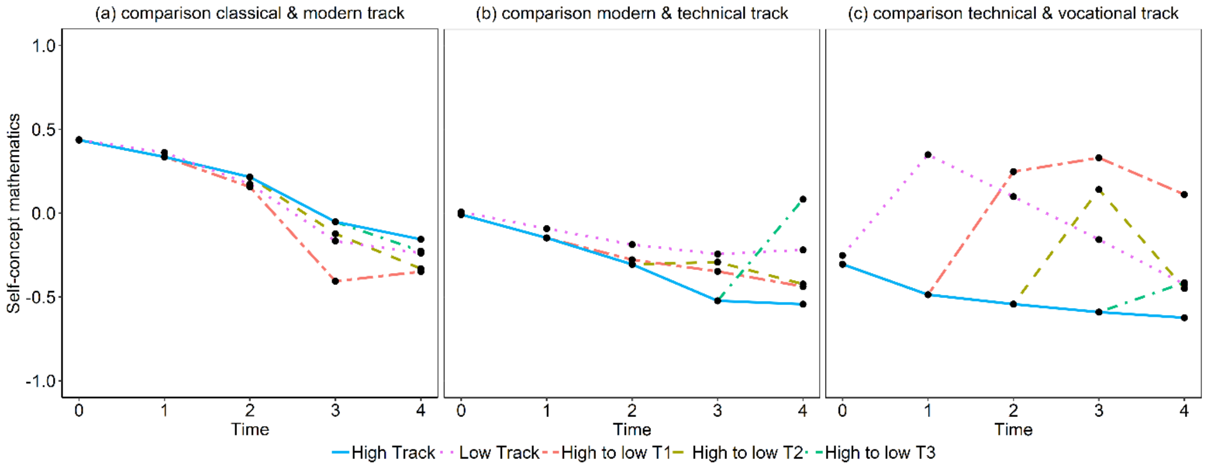

Figure 15. Development self-concept in mathematics estimated with MSMMs for each track allocation history of each track comparison.

3.3.2.5 Self-concept in Dutch

For self-concept in Dutch, the development estimated with the MSMM for each of the three track comparisons is shown in Figure 16. Table 12 shows the MSMM ATE estimates and SNMM ATT estimates. For brevity, we only discuss the results at T4.

For the classical and modern track comparison, students who are continuously in the higher track develop a significantly higher self-concept in Dutch compared to students who are continuously in the lower track, with a small effect size. Furthermore, students who changed from the higher to lower track generally do not develop significantly different from the other groups of students. The effects sizes are also nonmeaningful.

For the modern and technical track comparison, students who are continuously in the higher track develop a significantly lower self-concept in Dutch compared to students who are continuously in the lower track. However, the effect size is nonmeaningful. Generally, students who changed from the higher to lower track develop a significantly higher self-concept than students who are continuously in the higher track. The effect sizes are small to nonmeaningful.

For the technical and vocational track comparison, students who are continuously in the higher track do not differ from students who are continuously in the lower track. Students who change from the higher to lower track after T2 and T3 do not develop differently compared to students who are continuously in the higher or lower track. However, students who change from the higher to lower track after T1 do develop a significantly higher self-concept in Dutch than students who are continuously in the higher track or lower track.

Table 12
Contrast estimates self-concept in Dutch across track comparisons using the MSMMs and SNMMs.

	Classical & modern track comparison		Modern & technical track comparison		Technical & vocational track comparison	
	MSMM *d* (SE)	SNMM *d* (SE)	MSMM *d* (SE)	SNMM *d* (SE)	MSMM *d* (SE)	SNMM *d* (SE)
T1						
High – low	0.09*	0.08*	-0.03	-0.06	-0.39*	-0.35*
	(0.04)	(0.04)	(0.04)	(0.05)	(0.08)	(0.07)
T2						
High – low	0.16*	0.20*	0.15*	0.13*	-0.37*	-0.37*
	(0.04)	(0.04)	(0.05)	(0.05)	(0.08)	(0.08)
High – T1 Change	0.05	0.06	0.23*	0.21*	-0.37*	-0.54*
	(0.06)	(0.06)	(0.08)	(0.07)	(0.11)	(0.10)
T3						
High – low	0.31*	0.27*	-0.15*	-0.13*	-0.18*	-0.14
	(0.05)	(0.05)	(0.05)	(0.05)	(0.09)	(0.08)
High – T1 Change	0.22*	0.13*	-0.23*	-0.26*	-0.37*	-0.36*
	(0.07)	(0.07)	(0.08)	(0.08)	(0.11)	(0.10)
High – T2 Change	0.07	0.04	-0.21*	-0.22*	-0.28	-0.27*
	(0.06)	(0.05)	(0.06)	(0.06)	(0.19)	(0.12)
T4						
High – low	0.31*	0.29*	-0.19*	-0.14*	-0.01	0.04
	(0.05)	(0.04)	(0.06)	(0.05)	(0.09)	(0.11)
High – T1 Change	0.10	0.11	-0.26*	-0.21*	-0.30*	-0.28*
	(0.07)	(0.06)	(0.08)	(0.08)	(0.11)	(0.11)
High – T2 Change	0.16*	0.11*	-0.23*	-0.23*	0.05	0.04
	(0.06)	(0.05)	(0.06)	(0.06)	(0.13)	(0.12)
High – T3 Change	-0.06	0.01	-0.18	-0.14	-0.11	-0.13
	(0.13)	(0.10)	(0.09)	(0.07)	(0.18)	(0.14)

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

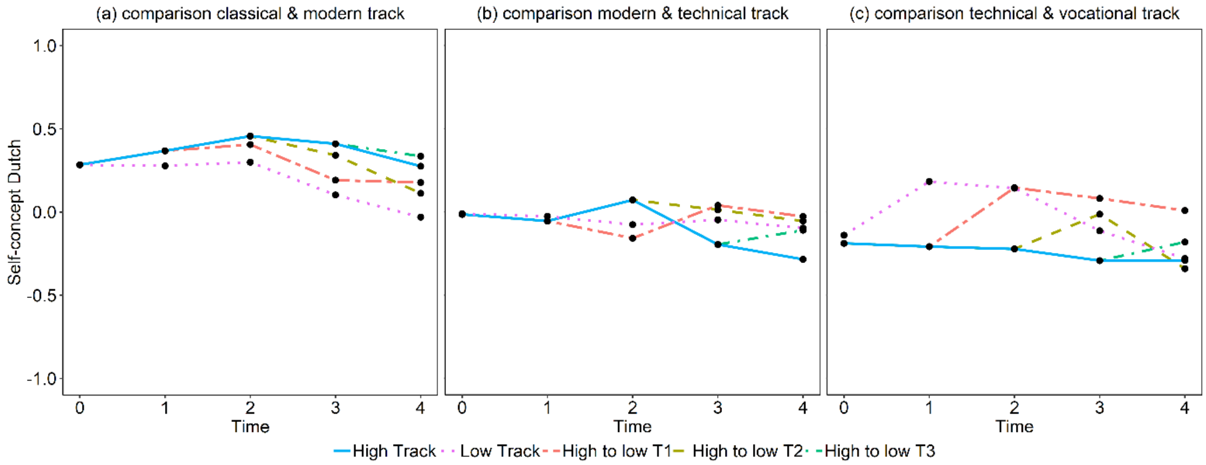

Figure 16. Development self-concept in Dutch estimated with MSMMs for each track allocation history of each track comparison.

3.4 Discussion empirical study

Our results supported the first hypothesis that being in a higher track is beneficial for academic performance relative to being continuously in a lower track. The hypothesis held true for each comparison and for both academic performance in mathematics and academic performance in Dutch reading comprehension. The one exception was that students continuously in the technical track and comparable students continuously in the vocational track did not differ for learning progress in Dutch reading comprehension.

Our results generally supported the second hypothesis that being in a higher track negatively affects academic self-concept, relative to being in a lower track. However, this only held true for the modern track and technical track comparison, and the technical track and vocational track comparison. For the classical and modern track comparison, the reverse was found, for being in the classical track was beneficial for academic self-concept. Also note that sometimes no significant differences were found. Hence, it remains arguable how strongly our results support the second hypothesis.

Our results somewhat supported the third hypothesis that downward track change negatively affects academic performance. Generally, we found that changing from the higher to lower track makes those students perform equal to the students who were continuously in the lower track. Hence, the relative gains made in the higher track disappeared. However, we found no support for the notion that students are better of starting in lower track instead of changing from the higher to lower track over time.

Our results did not support the fourth hypothesis that downward track change negatively affects academic self-concept. Our findings rather showed that the opposite is true. For both the modern and technical track comparison, and the technical and vocational track comparison we found that changing from the higher to lower track benefits academic self-concept. However, for the classical and modern track comparison, there were some indications that downward track change negatively affected academic self-concepts, though the effects were unstable across the MSMM and SNMM.

In this empirical study the MSMM and SNMM yielded results which led to equal conclusions on the effects of tracks. There were some differences, though this was to be expected given that the MSMM estimates ATEs and the SNMM estimates ATTs. Furthermore, both methods also dealt differently with students who went to an alternative track allocation history. The only notable difference between the MSMM estimates and SNMM estimates was the comparison between the classical and modern track on general academic self-concept. The MSMM estimates indicated that being in the classical track had a small positive effect, whereas the SNMM indicated that that no such effect existed. When interpreting the results of that comparison, we based ourselves on the MSMM estimates, for the MSMMs handle the students who have an alternative track allocation history more correctly than the SNMMs. Except that comparison, the estimates of the MSMM and SNMM were comparable in this empirical study.

4 Discussion

This study started from the observation that appropriate methods are required to estimate average effects of time-varying treatments that are subject to time-varying confounders. The potential outcomes framework describes two main assumptions that, if true, allow for the unbiased estimation of these average effects. The first, the sequential conditional exchangeability assumption, requires that the average effect cannot be ascribed to pretreatment differences between treatment conditions. The second, the positivity assumption, requires that overlapping strata of confounder and treatment histories exist between treatment conditions. The g-formula is an estimator that satisfies the sequential conditional exchangeability assumption by weighting, which removes all pretreatment differences between treatment conditions. However, the weights are estimated by stratifying across all observed combinations of confounder and treatment histories. This stratification is unrealistic in most empirical datasets and consequently the positivity assumption will not be true. The MSMM and SNMM, while both based on the g-formula, satisfy the assumptions in a manner better suited to empirical datasets. Hence, in this study, we compared the theories of both models and what they offer for practical application.

To prevent the problem of too many strata when using the g-formula, the MSMM and SNMM instead model treatment probabilities. For if the probability estimation is based on confounder and treatment histories, then the pretreatment differences in confounder and treatment histories will be reflected in unequal treatment probabilities. Consequently, an estimator that accounts for these unequal treatment probabilities will also account for pretreatment differences between treatment conditions. This will satisfy the sequential conditional exchangeability assumption. Furthermore, if an overlap exists in treatment probabilities across treatment conditions, the positivity assumption will also be satisfied. Therefore, the MSMM uses the inverse of predicted probabilities as weights to achieve balance in confounder and treatment histories between treatment conditions (i.e., IPTW). The SNMM directly uses the prediction of treatment probabilities to estimate average treatment effects (i.e., g-estimation). However, there are challenges in using treatment probabilities for estimating average effects of treatment histories, and MSMMs and SNMMs handle these differently.

The main challenge when using MSMMs is that confounder balance is achieved on average across infinite samples, but single sample imbalance will still exist by chance (e.g., Imai, King, & Stuart, 2008). This is because IPTW makes a dataset approximate a simple random sample (e.g., Rubin, 2007, pp. 25-27), which is unbiased but relatively inefficient compared to other sampling methods (e.g., Kish, 1965). If the effective sample size after IPTW is relatively large, then this inefficiency is of less concern. However, when the effective sample size after IPTW is relatively small, often because of large variability in weights, inefficiency will be a problem (e.g., Golinelli, Ridgeway, Rhoades, Tucker, & Wenzel, 2012). Truncation and removing respondents with extreme propensities will improve efficiency by preventing large weights, but they will also cause bias (e.g., Crump et al., 2009; Lee et al., 2011). Including the baseline measures in the structural model will also improve efficiency, but it will not directly reduce imbalance in later measures of time-varying confounders (e.g., Robins et al., 2000). Overall, strategies to reduce random imbalance after IPTW often induce bias, and researchers will have to try several strategies to prevent either from being too large.

While the SNMM also uses treatment probabilities to satisfy the sequential conditional exchangeability assumption, it suffers less from limited efficiency as the MSMM. The SNMM achieves this efficiency by basing the average effect estimate mainly on where the overlap in treatment probabilities is greatest (Vansteelandt et al., 2014, pp. 716-718). This prevents respondents with extreme treatment probabilities to cause uncertainty in the average effect estimate. Consequently, a higher efficiency is achieved. The resulting average effect estimate is of course under the assumption that the average effect is constant, even across respondents with extreme probability values who were barely involved in the estimation. Researchers will need to decide whether the constant treatment effect assumption is tenable for their study.

Note though that when the area of common support is limited, causal inference by applying a MSMM is only possible after removing respondents with extreme propensity scores. For the SNMM the causal inference will automatically only apply to the limited area of common support. The average effect estimates are then limited to only a small part of the population. However, how to interpret this estimate is debatable. Accordingly, Rosenbaum (2010, p. 86) argues that in this case it is preferable to redefine the population of interest based on observed covariates to simply prevent extreme probabilities from occurring.

Another challenge is specifying the prediction models for estimating the treatment probabilities. Typically, logistic regression models with covariates as linear and purely additive predictors are used. However, it is implausible that covariates always have noninteracting and nonlinear relationships with treatment probabilities (McCaffrey et al., 2004). Generalized boosted regression model (GBMs) use a data-adaptive algorithm that iteratively processes multiple regression trees for capturing all interacting and nonlinear relationships between the covariates and treatment probabilities. Using GBMs is considered best practice for MSMs (e.g., Stuart, 2010). Nevertheless, while GBMs give unbiased probability estimates, they will not minimize random sample imbalance. Different methods can minimize random sample imbalance, such as covariate balancing propensity scores (CBPS, Imai & Ratkovic, 2015). In a recent overview Griffin, McCaffrey, Almirall, Burgette, and Setodji (2017) compared different procedures for estimating treatment probabilities, showing that a method either minimizes random sample imbalance or is unbiased across samples. Researchers will again have to choose between reducing variance in sample estimates (i.e., efficiency) and reducing bias when specifying the prediction models for treatment probabilities. Furthermore, these methods are only available for MSMMs, not for SNMMs.

Related to specifying the prediction models is the question of which covariates to include in the prediction models. The consensus is that variables that predict both the treatment and the outcome (i.e., confounders) should be included. If sample size allows it, variables related to the outcome should also be included. However, variables that only predict the treatment should not be included, for they only increase the variance of the estimates (e.g., Brookhart et al., 2006; Myers et al., 2011; Stuart, 2010). GBMs will automate the process of variable inclusion in the model for estimating treatment probabilities (McCaffrey et al., 2004). Note that all sources of confounding should have been gathered during data collection before variable selection. This requires background knowledge about the phenomena under investigation. Of interest is that Steiner et al. (2010) showed that measuring all plausible confounders is not required, whereas McCaffrey, Lockwood and Setodji (2013) showed that unreliability in the covariates is not necessarily problematic. This is due to other variables correcting for the unmeasured and unreliably measured confounders. This facilitates bias reduction, but also makes the process of bias reduction somewhat opaque. Overall, researchers will still have to decide which variables to collect and what they consider predictive.

Lastly, choosing between the MSMM and the SNMM also means choosing between the ATE and the ATT of a treatment history. This choice depends on the estimand of interest; the average effect of a treatment history as if the whole population is in that treatment history, or the average effect of the treatment history for those who are in that treatment history. However, if only the weak sequential conditional exchangeability is tenable (Greenland & Robins, 2009, p. 4), then only the ATT is unbiased and the SNMM should be used.

5 Limitations

We introduced the g-formula as a theoretical computation formula for estimating average effects of time-varying treatments, but also described it as difficult to use in practice (e.g., Vansteelandt et al., 2014, p. 729). However, some authors have used the g-formula in their research. In these studies, the conditional exchangeability was considered tenable by either stratifying on relatively few variables and making parametric assumptions on specific confounders (e.g., Austin & Urbach, 2013; Snowden et al., 2011). Nonetheless, because time-varying confounding in psychological research is often characterized by many covariates, we considered its application less relevant. Hence, we limited this study to the application MSMMs and SNMMs.

We also refrained from discussing effect modification, which is the interaction effect between a confounder and treatment (Robins et al., 2000, pp. 556-557). In practice, we note that most studies are limited to average effect estimates. This is technically unproblematic for the MSMM, because the average effect estimate is averaged across all confounder levels. Accordingly, an interaction between confounder and treatment would not bias the average effect estimate. However, for the SNMM, such an interaction can cause bias, for the average effect estimate is based on where the area of common support is greatest (Vansteelandt et al., 2014, pp. 717-718). This area may be situated around a specific confounder value, and if an interaction exists with this confounder, the constant treatment effect assumption is untrue. Accordingly, an effect modification needs to be included in the SNMM to satisfy this assumption and give unbiased average effect estimates. Of note is that the MSMM can only include effect modification for baseline confounders, whereas the SNMM can also include effect modification for time-varying confounders (Robins et al., 2000, pp. 556-557; Vansteelandt et al., 2014, pp. 717-718).

6 Conclusion

The goal of this study was to compare the marginal structural mean model and the structural nested mean model for estimating average effects of time-varying treatments. At first glance, both models are similar, for both are estimators of average treatment effects and are situated in the same theoretical framework. However, the models estimate average treatment effects for different populations and differ in their assumptions. It was also shown how these differences have repercussions for the efficiency and the bias of the estimates. Accordingly, we used both models in a simulation study and an empirical study, showing how they can be applied in practice. Moreover, both models were situated in the potential outcomes framework, a theoretical framework that can help to understand the challenges of causal inference. We hope that, when combined with the information in the appendices, other researchers can also apply the models introduced in this article.

7 Appendix A: Collider stratification bias

Collider stratification bias will occur in any configuration of three variables where two independent variables, called the colliding variables, predict a third variable, called the collider (Cole et al., 2009; Whitcomb, Schisterman, Perkins, & Platt, 2009). In such a configuration, the two colliding variables will have a relation after conditioning on the collider. It may be counterintuitive why conditioning on a collider will result in a conditional relationship between colliding variables that is different from their marginal relationship. To clarify the source of this bias, we give a conceptual and statistical illustration. Afterwards, we provide a brief framework wherein this bias can be interpreted.

For the conceptual illustration (inspiration taken from Cole et al., 2009; Vandecandelaere et al., 2016), we use the hypothetical situation where students are assigned to an arts course based on their mathematical ability and painting ability. We imagine that both abilities are marginally independent and that both abilities predict the arts course assignment. Accordingly, the abilities are colliding variables whereas the assignment to the arts course is a collider. Subsequently, we assess the relation between mathematical and painting ability of students who were assigned to the arts course. This result will show that, within the group of students who were assigned to the arts course, both abilities are no longer independent. For a student with poor drawing skills in the arts course must be mathematically proficient and vice versa. Hence, while no relation existed between the mathematical and painting abilities for all students, there is a relation between mathematical and painting abilities within the group of students allocated to the arts course.

For the statistical illustration we generated two independent variables *A* and *B* (R syntax is shown at the end of this appendix). The scatterplot in Figure 1a accordingly shows that they do not have a tangible relationship. If these two variables predict a third variable *C*, then variables *A* and *B* will have tangible relationships within strata of *C*, as the scatterplot in Figure 1b shows. However, the strata averages of *A* and *B* will have a reverse relationship, as the scatterplot in Figure 1c shows. Partitioning the relationship between *A* and *B* according to strata of *C* is a simple way of conditioning the relation between *A* and *B* on *C*. Hence, a collider partitions the near zero marginal relationship of its colliding variables into two tangible conditional relationships, a relationship within collider strata and an opposing relationship between collider strata. The within and between relationships cancel each other out when summed.

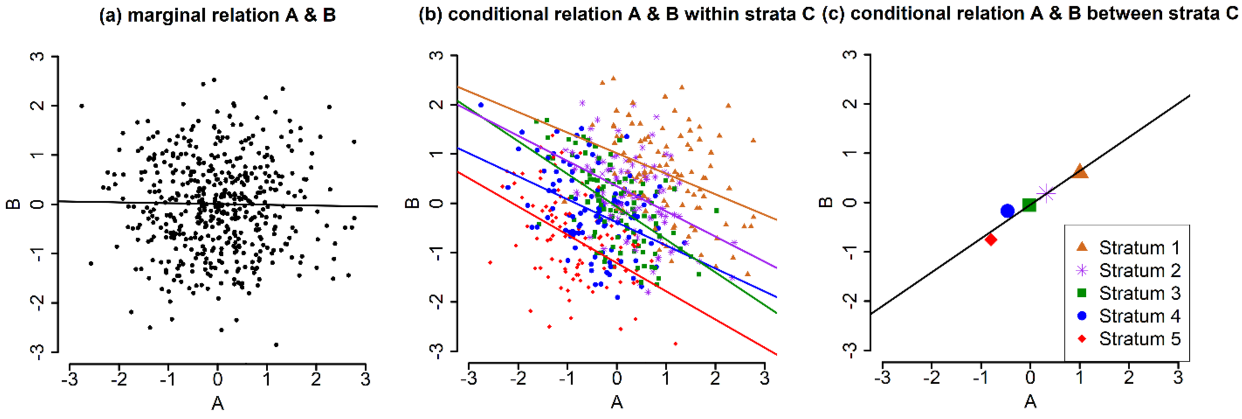

Figure 1. Comparing the marginal relation between two colliding variables A and B with their conditional relations within and between strata of collider C.

There are many examples in statistical literature of bias caused by confusing the conditional relationship between variables with their marginal relationship. The most well-known example is Simpsons Paradox, where a relationship exists between variables within groups, but reverses or disappears when combining the respondents of different groups into one dataset (Blyth, 1972). Another well-known example in medical research is Berkson's Paradox, where a relationship between two variables within a specific subpopulation (i.e., hospital patients) does not exist in the general population (Snoep, Morabia, Hernández-Diaz, Hernán, & Vandenbroucke, 2014).

Recently, different sources of bias have been connected to more general statistical concepts. For example, Pearl (2009) and Hernán et al. (2004) note that both confounding and collider stratification bias result in non-exchangeability between respondents across treatment groups. For both confounders and colliders, bias is removed by accounting for all paths originating from the colliding variables or confounders. This is achievable by either controlling for the colliding variables or confounder themselves or any variable that mediates the effects on the treatment exposure (Hernán et al., 2004; Pearl, 2009, pp. 16-18). This idea of having to block 'paths' has led to the statistical concept of 'd-separation' as a necessity for causal inference (Pearl 2009, pp. 16-18). However, despite recent advances in literature on causal literature to uncover different sources of bias (e.g., Porta, Vineis, & Bolúmar, 2015), discussion remains whether they are truly only nonmeaningful artifacts (e.g., Greenland, 2017, pp. 8-9; Krieger & Davey Smith, 2016)

```
#R code Appendix A

### Package preparation
install.packages("dplyr")
library(dplyr) #For ntile function, easily creates strata with equal
respondent sizes

### Data generation
A<-rnorm(500,mean=0,sd=1)
B<-rnorm(500,mean=0,sd=1)
C<-0.7*A+0.5*B+rnorm(500,mean=0,sd=0.509901951)

### Correlations variables
cor(A,B)
cor(A,C)
cor(B,C)

### Scatter plot A and B marginal
plot(A,B)
abline(lm(A~B),lwd=3)

### Strata creation

Cstrata<-ntile(C,5)
mean(c(
cor(A[Cstrata==1],B[Cstrata==1]),
cor(A[Cstrata==2],B[Cstrata==2]),
cor(A[Cstrata==3],B[Cstrata==3]),
cor(A[Cstrata==4],B[Cstrata==4]),
cor(A[Cstrata==5],B[Cstrata==5])
))

mean(c(
cov(A[Cstrata==1],B[Cstrata==1]),
cov(A[Cstrata==2],B[Cstrata==2]),
cov(A[Cstrata==3],B[Cstrata==3]),
cov(A[Cstrata==4],B[Cstrata==4]),
cov(A[Cstrata==5],B[Cstrata==5])
))

Amean<-aggregate(A,list(Cstrata),FUN=mean)[,2]
Bmean<-aggregate(B,list(Cstrata),FUN=mean)[,2]

### Scatter plots A and B conditional on C

plot(A,B)

abline(lm(B[Cstrata==1]~A[Cstrata==1]),lwd=3,col="red")
abline(lm(B[Cstrata==2]~A[Cstrata==2]),lwd=3,col="blue")
abline(lm(B[Cstrata==3]~A[Cstrata==3]),lwd=3,col="green4")
abline(lm(B[Cstrata==4]~A[Cstrata==4]),lwd=3,col="purple")
abline(lm(B[Cstrata==5]~A[Cstrata==5]),lwd=3,col="chocolate")

plot(Amean,Bmean)
abline(lm(Amean~Bmean),lwd=3)
```

8 Appendix B: g-estimation

8.1 An example of g-estimation for a time-fixed treatment

To understand g-estimation it seems helpful to remind ourselves of the central ideas of causal inference. The central idea of causal inference starts from the fundamental problem of causal inference (Holland, 1986), which is that the individual treatment effect for the potential outcomes, $\Delta_i = Y_i(1) - Y_i(0)$, cannot be observed. For a person *i* we can only observe either the potential outcome of being in the active treatment condition $Y_i(1)$ or the potential outcome of being in the control condition $Y_i(0)$. We cannot observe both $Y_i(1)$ and $Y_i(0)$ for one respondent *i*. However, if the conditional exchangeability assumption is tenable, an average treatment effect can still be estimated (Rosenbaum and Rubin, 1983). This assumption means that the average potential outcome of either the active treatment condition or the control condition is equal across the different treatment conditions when conditioning on the confounders *L*. Put formally, the potential outcome of being in the control condition is equal for those in the treated condition and those in the control condition when conditioning on *L*, with $E[Y(0)|Z = 1, L = l]= E[Y(0)|Z = 0, L = l]$. The same holds true for $Y(1)$. Hence, the fundamental problem of causal inference can be overcome if the conditional exchangeability assumption is tenable.

The conditional exchangeability assumption can also be interpreted in a different manner, which relates more clearly to the idea of the g-estimation. For the conditional exchangeability assumption also means that the potential outcomes of both conditions are independent of treatment exposure, with $Y(1),Y(0) \perp\!\!\!\perp Z|L$. Colloquially, treatment exposure *Z* should not predict the potential outcomes $Y(1)$ and $Y(0)$ after conditioning on confounders *L*. Accordingly, the reverse is also true, and the potential outcomes $Y(1)$ and $Y(0)$ should not predict whether a respondent is allocated to the treatment or the control condition. Note that the observed outcome *Y* does not need to be independent of the treatment exposure *Z*, for the research question is whether there is an average effect of the treatment exposure on the observed outcome. It is this understanding of the conditional exchangeability assumption that potential outcomes *Y*(1) and *Y*(0) should not predict *Z*, that is the starting point of g-estimation.

We should note that in the context SNMMs it suffices that the potential outcomes of being in the control condition *Y*(0) are independent from treatment exposure *Z* for different confounder levels of *L*. This is the weak conditional exchangeability assumption, formally written as $Y(0) \perp\!\!\!\perp Z | L$.

For illustrating g-estimation, we use the example data of Table 1 which has three respondents in the control condition (Z_i=0) and three respondents in the active treatment condition (Z_i=1). Accordingly, for the former respondents we know $Y_i|Z_i$=0, whereas for the latter respondents we know $Y_i|Z_i$=1. This is the same dataset as in Table 2 of the main text. However, the values of Y_i are now mean centered for the potential outcome in the control condition. Naïvely, we may think that subtracting the averages of both $((\overline{Y_\iota|1}) - \overline{Y_\iota|0}) = (3) - (-1))$ yields the average treatment effect, 4. This is incorrect, for we note that confounder *L* is unequally distributed across the active treatment condition and control condition. The last three columns show the potential outcomes $Y_i(0)$, potential outcomes $Y_i(1)$ and individual treatment effects ψ_i. The latter are always 2 and the average treatment effect is therefore also 2. This treatment effect is the ‘blip’. In practice, these potential outcomes and blip will never be observed, but we reveal these now for illustrative purposes. Note that $Y_i|Z_i$=0 is equal to $Y_i(0)$ and $Y_i|Z_i$=1 is equal to $Y_i(1)$ as they should be according to the consistency assumption.

Table 1
Example dataset g-estimation

Respondent *i*	Z_i	$Y_i \mid Z_i$=0	$Y_i \mid Z_i$=1	L_i	$Y_i(0)$	$Y_i(1)$	ψ_i
1	0	-3	?	0	-3	-1	2
2	0	-3	?	0	-3	-1	2
3	0	3	?	1	3	5	2
4	1	?	-1	0	-3	-1	2
5	1	?	5	1	3	5	2
6	1	?	5	1	3	5	2

In Figure 1, a DAG shows the relationship between the variables of Table 1. Confounder L_i predicts treatment exposure Z_i and the outcome Y_i, whereas Z_i also has a direct effect on Y_i (i.e., the average treatment effect of interest). Accordingly, the confounding effect of L_i is why we need the control for the values L_i if we want to estimate the treatment effect of Z_i on the outcome Y_i.

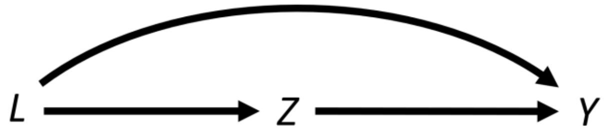

Figure 1. DAG example confounder *L*, treatment *Z* and outcome *Y*.

8.2 Estimating the probability of treatment exposure

In this section we show how the potential outcomes $Y_i(Z_i=0)$ in the example of Table 1 and Figure 1 do not contribute to the prediction of treatment exposures Z_i when we condition on the confounder values of L_i. Equation 1 shows the predictive model of Z_i when we have $Y_i(0)$ and L_i as predictors:

$$P(Z_i|y_i(0), l_i) = \alpha_0 + \alpha_1 l_i + \alpha_2 y_i(0)_i \ (1)$$

Equation 1 relates the probability of Z_i to a parametric function of both $Y_i(0)$ and L_i. In this case the link function is a logit function. If the conditional exchangeability assumption is true, $Y_i(0)$ should not predict Z_i when we condition on L_i. Therefore, α_2 should be zero, but only if using the correct parameter values for α_0 and α_1, otherwise the conditional exchangeability assumption would be untrue.

Luckily, we know both the true values $Y_i(0)$ and parameters $\alpha_0$and α_1. We generated the probability of treatment assignment to be 1/3 when L_i=0, accordingly α_0=-0.693, for this is the logit of $P(Z_i|L=0)=1/3$. Furthermore, we generated the probability of treatment assignment to be 2/3 when L_i=1, accordingly α_1=1.386, for α_0+α_1= 0.693 is the logit of $P(Z_i|L=1)=2/3$. In short, the true parameter values of α_0 and α_1 are -0.693 and 1.386 respectively, leaving us only with an unknown value for parameter α_2.

First, we try α_2=0 in combination with the true parameter values α_0=-0.693 and α_1=1.386. If the conditional exchangeability assumption is true, α_2=0 should yield unbiased predictions. Accordingly, Equation 1 now predicts that a respondent with L_i=1 has a probability of 2/3 of being assigned to the active treatment condition, Z_i=1. Furthermore, we predict that a respondent with L_i=0 has a probability of 1/3 of being assigned to the active treatment condition, Z_i=1. This prediction matches up exactly with the data in Table 1.

We also try some alternative values for α_2 in Equation 1, to find out whether potential outcomes $Y_i(0)$ cannot be made to help predict treatment exposure Z_i. First, we keep the true parameter values α_0=-0.693 and α_1=1.386 but make α_2=0.5. Now the model predicts probability of treatment assignment to be 9/10 when L_i=1 and 1/10 when L_i=0. Obviously, this does not correspond with our data, and α_2=0.5 is incorrect. As a second attempt, we kept the true parameter values α_0=-0.693 and α_1=1.386 but make α_2=-0.5. Now the model predicts probability of treatment assignment to be 31/100 when L_i=1 and 69/100 when L_i=0. Again, this does not correspond with our data, and α_2= -0.5 seems incorrect. In conclusion, the potential outcomes of being in the control condition $Y_i(0)$ seem unable to improve our prediction after conditioning on confounder L_i.

However, it should be noted that if we do not condition on confounder L_i, which makes the conditional exchangeability assumption untenable, then $Y_i(0)$ will contribute to the prediction of treatment exposure Z_i. For example, if we set α_0 and α_1 to zero, effectively removing L_i as predictor, the predictions of Equation 1 will still be correct if setting α_2 equal to 0.231. Accordingly, whether the potential outcomes of being in the control condition *$Y_i(0)$* do not contribute to the prediction of treatment exposures depends on the tenability of the conditional exchangeability assumption.

In conclusion, this section shows how the potential outcomes $Y_i(0)$ do not contribute to the prediction of treatment exposures Z_i when conditioning on the confounder values L_i. This is due the potential outcomes $Y_i(0)$ being conditionally exchangeable on Z_i after conditioning on L_i. However, in practice, the potential outcomes of the control condition are unknown for respondents in the active treatment condition and need to be estimated. We have learned though that under the conditional exchangeability assumption these unknown potential outcomes should not predict treatment exposure, after conditioning on the confounder. Hence, we can use that information to procure the unknown potential outcomes.

8.3 Estimating the blip

So far, we have assumed that we know the potential outcome $Y_i(0)$ for each respondent in our example of Table 1 and Figure 1. However, in practice we only know $Y_i(0)$ for those respondents with $Y_i|Z_i$=0; but not for respondents with $Y_i|Z_i$=1. For respondents with $Y_i|Z_i$=1, their potential outcome of being in the control condition $Y_i(0)$ is equal to their observed outcome Y_i minus the blip ψ. Hence Equation 1 can be rewritten as Equation 2.

$$P(Z_i|y_i - z_i\psi, l_i) = \alpha_0 + \alpha_1 l_i + \alpha_2(y_i - z_i\psi)\ (2)$$

Equation 2 shows that the challenge is in jointly finding a value for the blip ψ and correctly modeling the treatment exposure probabilities based on confounder L_i. We know from the former section that for unbiased estimates of $(y_i - z_i\psi)$ should make α_2=0. This does require that we condition on L_i to satisfy the conditional exchangeability assumption.

So far, we have written the conditional exchangeability assumption as $Y_i(0)\perp\perp Z_i|L_i$. This equation means that the potential outcomes $Y_i(0)$ are independent from the treatment exposures Z_i after conditioning on L_i. Another way to describe the independence between $Y_i(0)$ and $Z_i|L_i$ is that their covariance is zero. Hence, cov($Y_i(0)$, $Z_i|L_i$)=0. We can use this as basis for deriving an estimating equation for ψ, in which we also replace $Y_i(0)\ by$ Y_i - $Z_i\psi$ in Equation 3:

$$0 = Cov(Y_i - Z_i\psi, Z_i|L_i)\ (3)$$

The covariance from Equation 3 is equal to the expected value of the product of both Y_i - $Z_i\psi$ and $Z_i|L_i$, minus the product of both the expected values of Y_i - $Z_i\psi$ and the expected value of $Z_i|L_i$. This is expressed in Equation 4:

$$0 = E[(Y_i - Z_i\psi)Z_i|L_i] - E[Y_i - Z_i\psi]E[Z_i|L_i]\ (4)$$

Next, we rearrange the equation according to two properties of expected values. First, the expected value of the sum (this includes subtraction) is equal to the sum of the expected values. Second, because the expected value of a constant (in our case the blip ψ) is the constant itself, it can be brought outside the expected value operator. This allows us to rewrite Equation 4 into Equation 5:

$$0 = E[Y_iZ_i|L_i] - \psi E[Z_iZ_i|L_i] - E[Y_i]E[Z_i|L_i] + \psi E[Z_i]E[Z_i|L_i]\ (5)$$

Next, we rearrange Equation 5 so that the terms that contain blip ψ are on the left side of the equation, and the other terms are on the right side of the equation. This gives Equation 6:

$$\psi(E[Z_iZ_i|L_i] - E[Z_i]E[Z_i|L_i]) = E[Y_iZ_i|L_i] - E[Y_i]E[Z_i|L_i]\ (6)$$

Then, we arrange Equation 6 so that the blip ψ becomes a function of the remaining expected values, as shown in Equation 7:

$$\psi = \frac{E[Y_i Z_i|L_i] - E[Y_i]E[Z_i|L_i]}{E[Z_i Z_i|L_i] - E[Z_i]E[Z_i|L_i]} \text{ (7)}$$

The numerator and denominator of Equation 7 now have the interesting property to be an expression of the covariance formula for cov(Y_i,Z_i|L_i) and cov(Z_i,Z_i|L_i) respectively. The covariance of two random variables can also be expressed as the expected value of the product of the deviations of these random variables. This is how we reformulate both the numerator and denominator, as shown in Equation 8:

$$\psi = \frac{E[(Y_i - E[Y_i])(Z_i|L_i - E[Z_i|L_i])]}{E[(Z_i - E[Z_i])(Z_i|L_i - E[Z_i|L_i])]} \text{ (8)}$$

Using the distributive law (i.e., the product of a sum is equal to the sum of the products), we reformulate both the numerator and denominator, as shown in Equation 9:

$$\psi = \frac{E[Y_i(Z_i|L_i - E[Z_i|L_i]) - E[Y_i](Z_i|L_i - E[Z_i|L_i])]}{E[Z_i(Z_i|L_i - E[Z_i|L_i]) - E[Z_i](Z_i|L_i - E[Z_i|L_i])]} \text{ (9)}$$

Next, because the expected value of the sum (this includes subtraction) is equal to the sum of the expected values, we can rewrite Equation 9 into Equation 10:

$$\psi = \frac{E[Y_i(Z_i|L_i - E[Z_i|L_i])] - E[E[Y_i](Z_i|L_i - E[Z_i|L_i])]}{E[Z_i(Z_i|L_i - E[Z_i|L_i])] - E[E[Z_i](Z_i|L_i - E[Z_i|L_i])]} \text{ (10)}$$

It is important to realize that the expected value which is part of another expected value is a constant and can be treated as such. Accordingly, past the middle minus sign in both the denominator and numerator we bring E[Y_i] and E[Z_i] outside the expected value operator, for they are constants. To the right of the middle minus sign in both the denominator and numerator the expected value of the sum is also made the sum of the expected values. Hence, we procure Equation 11:

$$\psi = \frac{E[Y_i(Z_i|L_i - E[Z_i|L_i])] - E[Y_i](E[Z_i|L_i] - E[Z_i|L_i])}{E[Z_i(Z_i|L_i - E[Z_i|L_i])] - E[Z_i](E[Z_i|L_i] - E[Z_i|L_i])} \text{ (11)}$$

It is immediately clear that the latter part of the equation in both the numerator and denominator is zero. This gives us equation 12:

$$\psi = \frac{E\big[Y_i\big(Z_i|L_i - E(Z_i|L_i)\big)\big]}{E\big[Z_i\big(Z_i|L_i - E(Z_i|L_i)\big)\big]} \text{ (12)}$$

In this equation the blip is a function of the observed outcomes Y_i, observed treatment exposures Z_i, and expected treatment exposures E(Z_i). Now, we only need to define estimators. For both the numerator and denominator we simply need to calculate the mean value. The estimator of the expected value of Z_i should be based on the confounders to satisfy the conditional exchangeability assumption. The estimator of the expected value of Z_i is therefore Ê(Z_i|L_i). This estimator can be a logistic regression function that predicts the treatment probability Z_i as a function of the confounder L_i. Accordingly, the final equation is given in Equation 13:

$$\hat{\psi} = \frac{\sum_{i=1}^{n}\left[Y_i\left(Z_i - \hat{E}(Z_i|L_i)\right)\right]}{\sum_{i=1}^{n}\left[Z_i\left(Z_i - \hat{E}(Z_i|L_i)\right)\right]} \ (13)$$

We use the closed form solution in Equation 13 to estimate the blip in our example of Table 1. Accordingly, for Ê(Z_i|L_i) we predict that a respondent with L_i=0 has a probability of 1/3 and that a respondent with L_i=1 has a probability of 2/3. This information and the information in Table 1 concerning the observed outcomes and treatment allocations are then plugged into equation 13, leading to the following results in Equation 14:

$$\hat{\psi} = \frac{-3 * \left(0 - \frac{1}{3}\right) - 3 * \left(0 - \frac{1}{3}\right) + 3 * \left(0 - \frac{2}{3}\right) - 1 * \left(1 - \frac{1}{3}\right) + 5 * \left(1 - \frac{2}{3}\right) + 5 * \left(1 - \frac{2}{3}\right)}{0 * \left(0 - \frac{1}{3}\right) + 0 * \left(0 - \frac{1}{3}\right) + 0 * \left(0 - \frac{2}{3}\right) + 1 * \left(1 - \frac{1}{3}\right) + 1 * \left(1 - \frac{2}{3}\right) + 1 * \left(1 - \frac{2}{3}\right)}$$

$$= \frac{\left(\frac{3}{3}\right) + \left(\frac{3}{3}\right) - \left(\frac{6}{3}\right) - \left(\frac{2}{3}\right) + \left(\frac{5}{3}\right) + \left(\frac{5}{3}\right)}{-0 - 0 - 0 + \frac{2}{3} + \frac{1}{3} + \frac{1}{3}} = \frac{\frac{8}{3}}{\frac{4}{3}} = 2 \ (14)$$

Accordingly, the estimate resulting from this g-estimation has resulted in the blip estimate $\hat{\psi}$ = 2, based solely on the observed outcomes Y_i, observed confounder values L_i and observed treatment assignments Z_i. This estimate is equal to the true value of ψ.

9 Appendix C: Rank preservation

In the section 'SNMMs' it was stated that the blip value was assumed to be constant on average across all respondents. In the literature this is also referred to as an assumption of respondent rank preservation (Robins & Hernán, 2008, p. 577). This entails that respondents retain the same rank in potential outcome across all treatment conditions. For our simple example, comparing an active treatment condition (*Z* = 1) with a control condition (*Z* = 0), this means that respondents retain the same rank for potential outcomes *Y*(*Z*=1) and *Y*(*Z*=0). For example, if someone has the highest score for *Y*(*Z*=1) in a population, that person will also have the highest score for *Y*(*Z*=0) in the same population. This assumption of respondent rank preservation means that for each respondent the potential outcome $Y_i^*(0)$ can be estimated if knowing blip ψ_0. For example, we could estimate the $Y_i^*(0)$ for a respondent who was in the active treatment condition *Z* = 1:

$$Y_i^*(0) = Y_i - \psi\alpha_0$$

This illustrates that a respondents' potential outcome of being in the control condition is simply the result of a respondent's observed outcome Y_i minus a blip ψ_0 if respondent *i* was in the active treatment condition. However, respondent rank preservation seems unrealistic and it may be more plausible to expect heterogeneity in treatment effects. It has been shown though that the rank preservation assumption does not need to hold at the respondent level, but only on average (Robins & Hernán, 2008, p. 579). Consequently, in the section 'SNMMs' it was stated that the blip value was assumed to be constant on average across all respondents.

Bibliografie

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, *27*(12), 2037–2049. Retrieved from http://doi.wiley.com/10.1002/sim.3150

Austin, P. C., & Urbach, D. R. (2013). Using G-computation to estimate the effect of regionalization of surgical services on the absolute reduction in the occurrence of adverse patient outcomes. *Medical Care*, *51*(9), 797–805.

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.

Blyth, C. R. (1972). On Simpson's Paradox and the sure-thing principle. *Journal of the American Statistical Association*, *67*(338), 364–366. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/01621459.1972.10482387

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. Retrieved from https://www.tandfonline.com/doi/full/10.1080/01425692.2012.704720

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*(12), 1149–1156.

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

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

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

Cole, S. R., & Frangakis, C. E. (2009). The consistency statement in causal inference: A definition or an assumption. *Epidemiology*, *20*(1), 3–5. Retrieved from https://insights.ovid.com/crossref?an=00001648-200901000-00003

Cole, S. R., & Hernan, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, *168*(6), 656–664. Retrieved from https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwn164

Cole, S. R., Platt, R. W., Schisterman, E. F., Chu, H., Westreich, D., Richardson, D., & Poole, C. (2009). Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology*, *39*(2), 417–420.

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

Daniel, R. M., Cousens, S. N., De Stavola, B. L., Kenward, M. G., & Sterne, J. A. C. (2013). Methods for dealing with time-dependent confounding. *Statistics in Medicine*, *32*(9), 1584–1618.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

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.

Greenland, S. (2017). For and against methodologies: some perspectives on recent causal and statistical inference debates. *European Journal of Epidemiology*, *23*(1), 3–20.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, *10*(1), 37–48.

Greenland, S., & Robins, J. M. (2009). Identifiability, exchangeability and confounding revisited. *Epidemiologic Perspectives & Innovations*, *6*(4), 1–9.

Griffin, B. A., McCaffrey, D. F., Almirall, D., Burgette, L. F., & Setodji, C. M. (2017). Chasing balance and other recommendations for improving nonparametric propensity score models. *Journal of Causal Inference*, *5*(2), 1–37.

Hernán, M. A., Brumback, B., & Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, *11*(5), 561–570.

Hernán, M. A., Hernández-Díaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, *15*(5), 615–625.

Højsgaard, S., Halekoh, U., & Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software*, *15*(2), 1–11.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

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.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(2), 481–502.

Imai, K., & Ratkovic, M. (2015). Robust estimation of inverse probability weights for marginal structural models. *Journal of the American Statistical Association*, *110*(511), 1013–1023.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706–710.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, *86*(1), 4–29.

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

Cambridge, United Kingdom: University Press.

Joffe, M. M. (2012). Commentary: Structural nested models, G-estimation, and the healthy worker effect: The promise (mostly unrealized) and the pitfalls. *Epidemiology*, *23*(2), 220–222.

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

Kish, L. (1965). *Survey sampling*. New York City, NY: Wiley.

Krieger, N., & Davey Smith, G. (2016). The tale wagged by the DAG: Broadening the scope of causal inference and explanation for epidemiology. *International Journal of Epidemiology*, *45*(6), 1787–1808.

Kuhn, M., Weston, S., Wing, J., & Forester, J. (2016). The contrast package. *CRAN Package Repository*, 1–14.

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS ONE*, *6*(3), 1–6.

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

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

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

Marsh, H. W., & O'Neill, R. (1984). Self Description Questionnaire III: The construct validity of multidimensional self-concept ratings by late adolescents. *Journal of Educational Measurement*, *21*(2), 153–174.

McCaffrey, D. F., Lockwood, J., & Setodji, C. (2013). Inverse probability weighting with error-prone covariates. *Biometrika*, *100*(3), 671–680.

McCaffrey, D. F., Ridgeway, G., & Morral, A. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*(4), 403–425.

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

Moodie, E. E. M., Richardson, T. S., & Stephens, D. A. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics*, *63*(2), 447–455.

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.

Pearl, J. (2009). *Causality: models, reasoning and inference*. Cambridge, United Kingdom: University Press.

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

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, *61*(2), 317–337.

Porta, M., Vineis, P., & Bolúmar, F. (2015). The current deconstruction of paradoxes: one sign of the ongoing methodological “revolution.” *European Journal of Epidemiology*, *30*(10), 1079–1087.

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

Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, *7*(9–12), 1393–1512.

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, NY: Springer.

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

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

Robins, J. M., Mark, S. D., & Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 479–495.

Rosenbaum, P. R. (1984). The consquences of adjustment for a concomitant variable that has been effected by the treatment. *Journal of the Royal Statistical Society. Series A (General)*, *147*(5), 656–666.

Rosenbaum, P. R. (2002). *Observational studies*. New York City, NY: Springer.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. New York City, NY: Springer New York.

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.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

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

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, *5*(4), 472–480.

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

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, *26*(1), 20–36.

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.

Snoep, J. D., Morabia, A., Hernández-Diaz, S., Hernán, M. A., & Vandenbroucke, J. P. (2014). Commentary: A structural approach to Berkson's fallacy and a guide to a history of opinions about it. *International Journal of Epidemiology*, *43*(2), 515–521.

Snowden, J. M., Rose, S., & Mortimer, K. M. (2011). Implementation of G-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology*, *173*(7), 731–738.

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

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15*(3), 250–267.

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.

Talbot, D., Atherton, J., Rossi, A. M., Bacon, S. L., & Lefebvre, G. (2015). A cautionary note concerning the use of stabilized weights in marginal structural models. *Statistics in Medicine*, *34*(5), 812–823.

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

van der Wal, W. M., Noordzij, M., Dekker, F. W., Boeschoten, E. W., Krediet, R. T., Korevaar, J. C., & Geskus, R. B. (2010). Comparing mortality in renal patients on hemodialysis versus peritoneal dialysis using a marginal structural model. *The International Journal of Biostatistics*, *6*(1), 41–57.

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

VanderWeele, T. J., Hawkley, L. C., Thisted, R. A., & Cacioppo, J. T. (2011). A marginal structural model analysis for loneliness: Implications for intervention trials and clinical practice. *Journal of Consulting and Clinical Psychology*, *79*(2), 225–235.

VanderWeele, T. J., & Shpitser, I. (2013). On the definition of a confounder. *Annals of Statistics*, *41*(1), 196–220.

Vansteelandt, S., Joffe, M., & others. (2014). Structural nested models and g-estimation: The partially realized promise. *Statistical Science*, *29*(4), 707–731.

Vansteelandt, S., & Keiding, N. (2011). Invited commentary: G-computation--lost in translation? *American Journal of Epidemiology*, *173*(7), 739–742.

Wallace, M. P., Moodie, E. E. M., & Stephens, D. A. (2017). An R package for G-estimation of structural nested mean models. *Epidemiology*, *28*(2), e18–e20.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.

Whitcomb, B. W., Schisterman, E. F., Perkins, N. J., & Platt, R. W. (2009). Quantification of collider-stratification bias and the birthweight paradox. *Paediatric and Perinatal Epidemiology*, *23*(5), 394–402.