Time-varying Treatments in Observational Studies: Marginal Structural Models of the Effects of Early Grade Retention on Math Achievement

Machteld Vandecandelaere[a], Stijn Vansteelandt[b], Bieke De Fraine[a], and Jan Van Damme[a]

[a] Centre for Educational Effectiveness and Evaluation

The Education and Training Research Group, University of Leuven, Leuven, Belgium

[b] Department of Applied Mathematics, Computer Science and Statistics

Ghent University, Ghent, Belgium

Correspondence concerning this article should be addressed to Machteld Vandecandelaere, The Education and Training Research Group, Centre for Educational Effectiveness and Evaluation, Dekenstraat 2 (pb 3773), 3000 Leuven, Belgium.

E-mail: Machteld.Vandecandelaere@ppw.kuleuven.be

To cite this article: 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.* Advance online publication.

Abstract

One of the main objectives of many empirical studies in the social and behavioral sciences is to assess the causal effect of a treatment or intervention on the occurrence of a certain event. The randomized controlled trial is generally considered as the gold standard to evaluate such causal effects. However, because of ethical or practical reasons, social scientists are often bound to the use of non-experimental, observational designs. When the treatment and control group are different with regard to variables that are related to the outcome, this may induce the problem of confounding. A variety of statistical techniques, such as regression, matching and subclassification, is now available and routinely used to adjust for confounding due to measured variables. However, these techniques are not appropriate for dealing with time-varying confounding, which arises in situations where the treatment or intervention can be received at multiple time points. In this article, we explain the use of marginal structural models and inverse probability weighting to control for time-varying confounding in observational studies. We illustrate the approach with an empirical example of grade retention effects on mathematics development throughout primary school.

Keywords: time-varying treatment, marginal structural models, grade retention

This article results from our research experience in dealing with time-varying confounding in the evaluation of a grade retention policy. Confounding refers to the phenomenon whereby common causes of the treatment and the outcome distort the association between both to the extent that it does not equal the corresponding causal effect. The example in this article investigates the effects of early grade retention on development in mathematics achievement throughout primary education. Grade retention is the practice of holding back struggling children for one school year.

Assessing the effects of grade retention yields the following issues. First, a clear choice of the comparison strategy is important. In general, two approaches can be used. In a same-grade comparison, the outcomes of grade repeaters are compared with those of their one-year-younger grade-mates. This strategy answers the question of how, at the cost of one extra year of education, grade repeaters develop compared to similar promoted grade-mates who are one year younger (e.g., Wu, West & Hughes, 2008a). In a same-age comparison, on the other hand, the outcomes of grade repeaters are compared with their one-grade-higher age-mates. A same-age comparison aims to address the counterfactual question of how the retainees would have developed had they been promoted instead (e.g., Hong & Raudenbush, 2005). The drawback of both strategies is that retention effects can be ascribed to at least one other difference between the retained and promoted group under study. In a same-grade comparison, the effects of grade retention can be attributed to age, for example, retainees score initially higher because they are one year older compared to similar grade-mates. In a same-age comparison, the effects can be ascribed to the grade, for example, promoted students score higher because they are confronted with higher level learning material in a higher grade. Both types of comparisons are informative, but they clearly answer different questions (Wu, West & Hughes, 2008a). Depending on the question one wishes to answer, one or the other strategy may be more satisfying, as has been contended by multiple researchers in the field of grade retention research. For example, while

Lorence (2006) argues that same-grade comparisons are a more valuable comparison because they compare retained students to their new peers, Hong and Raudenbush (2005) see this as a drawback because such comparison does not show how the retained students would fare had they been promoted instead.

A second issue in grade retention research is that retainees and promoted children differ with regard to a large number of variables. For example, compared to children who are promoted to first grade, children who are retained in kindergarten are on average younger, have lower scores in mathematics, language and psychosocial skills, speak more often a foreign language, and have a lower socio-economic status (Vandecandelaere, Schmitt, Vanlaar, De Fraine, & Van Damme, 2015a). This raises concerns that the observed effects of grade retention may be confounded. Third, children who are on the edge of being retained but who are promoted anyway, are likely to be retained in the next grade instead (Jacob & Lefgren, 2009; Vandecandelaere, Schmitt, Vanlaar, De Fraine, & Van Damme, 2014; Wu, West, & Hughes, 2008a). In other words, children can be retained at different points in time. Acknowledging the time-dependent nature of grade retention, the practice may influence some of the confounding factors (e.g., mathematics scores). This issue suggests that confounders may lie on the causal path from grade retention to later outcomes. Conventional regression methods are not generally appropriate to deal with such confounders (Robins, 1989; Robins, Hernán, & Brumback, 2000).

This article starts with a brief description of the setting in studies with a time-fixed treatment. We then show how marginal structural models can be used to deal with time-varying treatments. This is followed by our empirical example of marginal structural models with regard to grade retention effects on mathematics development throughout primary education.

Studies with a Time-fixed Treatment

We first define the notation. For N independent subjects, we observe a treatment Z, which might affect the outcome Y. This is illustrated in Figure 1, which demonstrates a time-

fixed causal model. Let *X* be a vector of time-fixed covariates which might confound the relationship between *Z* and *Y*. *U* represents a vector of unmeasured variables, affecting the outcome.

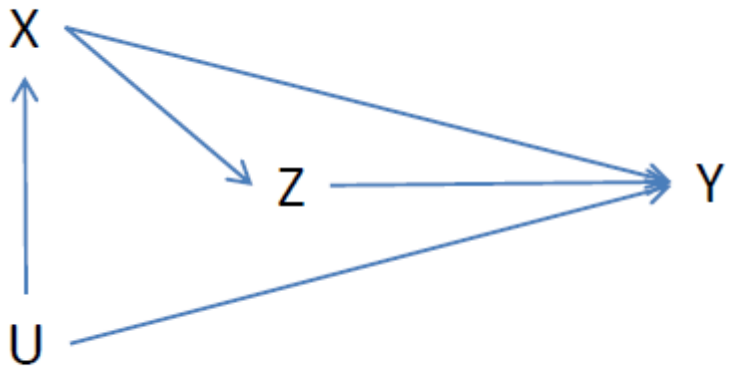


Figure 1. Causal diagram of treatment *Z* into outcome *Y*, representing a time-fixed causal model. *X* represents a vector of time-fixed covariates which might confound the relationship between *Z* and *Y*. *U* represents a vector of unmeasured variables, affecting the outcome. The missing arrow from *U* to *Z* encodes the assumption that there is no unmeasured confounding, an assumption that is routinely made in the analysis of most observational studies.

Potential Outcomes Framework

To formalize marginal structural models, we make use of the *potential outcomes framework* (Hernán, 2004; Hong, 2015; Imbens & Rubin, 2015; Rubin, 1974). Consider for example the causal effect of kindergarten retention on math achievement one year later. Imagine that we could observe mathematics achievement for every child in the population concurrently under two conditions: math achievement after kindergarten retention (treatment 1) and math achievement after promotion to first grade (treatment 0). In Table 1, the potential outcomes of six hypothetical children are displayed. The individual treatment effect is then the difference between these two potential outcomes: $Y(1)$-$Y(0)$. The average treatment effect (ATE) is the mean of these differences: $E[Y(1)$-$Y(0)]$, which encodes the average difference in outcome if the entire population was retained versus if the entire population was promoted. In Table 1, the

ATE of kindergarten retention is negative (-2.5). We may alternatively calculate the ATE as the difference between the marginal means of the potential outcomes under the two conditions: $E[Y(1)]$-$E[Y(0)]$, which is equal to $E[Y(1)$-$Y(0)]$. In Table 1, the marginal mean if all children were retained is 43, and the marginal mean if all children were promoted is 45.5, which gives an ATE of -2.5.

Table 1

Potential outcomes of six hypothetical students.

Student	***Y*(1)**	***Y*(0)**	***Y*(1)-*Y*(0)**
1	47	51	-4
2	50	54	-4
3	44	48	-4
4	39	40	-1
5	38	38	0
6	40	42	-2
Mean	43	45.5	-2.5

$Y(1)$ = outcome when treated; $Y(0)$ = outcome when untreated

In reality, of course, one of the two outcomes is an unobservable potential outcome or counterfactual. The fact that each individual has some potential outcomes missing, defines what is called the fundamental problem in causal inference (Holland, 1986).

The observed potential outcomes of our six hypothetical students are displayed in Table 2. When using only the observed outcomes, the difference in marginal means might be biased. This problem arises when there is a confounder X that simultaneously influences the treatment (and thus determines what potential outcome is observed) and the outcome. For example, consider a dummy variable X, indicating whether or not a child speaks the language of instruction at home. Suppose that non-native speakers are more likely to be retained and have generally lower math achievement scores. In Table 2, we observe more potential outcomes under the retention condition in the non-native group than in the native group. As a

consequence, the difference in marginal means of the observed potential outcomes gives a biased estimate of the ATE. Kindergarten retention appears more harmful (-5.6) than when we used the true marginal means (-2.5).

Table 2

Observed outcomes, propensity scores and weights of six hypothetical students.

Student	***Z***	***X***	***Y*(1)**	***Y*(0)**	***Y*(1)-*Y*(0)**	**PS**	**W**
1	1	1	47	?	?	1/3	3
2	0	1	?	54	?	2/3	1.5
3	0	1	?	48	?	2/3	1.5
Mean			47	51	-4		
4	0	0	?	40	?	1/3	3
5	1	0	38	?	?	2/3	1.5
6	1	0	40	?	?	2/3	1.5
Mean			39	40	-1		
Overall mean			41.7	47.3	-5.6		

Z = treatment; *X* = covariate; *Y*(1) = outcome when treated; *Y*(0) = outcome when untreated; *PS* = propensity score; *W* = weight

Marginal Structural Models in a Time-fixed Setting

A marginal structural mean model is a model for the population average of the potential outcomes at each treatment level *Z* (Robins et al., 2000); for instance,

(1)

$$E[Y(z)] = \beta_0 + \beta_1 z,$$

for z=0,1, in which $\beta_0 = E[Y(0)]$ is the marginal mean if all subjects in the population received treatment 0, and $\beta_1 = E[Y(1)]-E[Y(0)]$ is the ATE of *Z*.

The example in Table 2 illustrates how not adjusting for *X* can lead to a spurious relationship between treatment and outcome due to a violation of the so-called exchangeability assumption. *Exchangeability* between treatment groups implies that the treated, had they been untreated, would have experienced the same average outcome as the untreated did, and vice versa. Exchangeability holds when treatment assignment is independent of both potential

outcomes, as in a randomized controlled trial (RCT). Although it does not generally hold in observational studies, in such studies, exchangeability may sometimes still be met within more homogenous covariate strata, in which case we say that the treated and untreated populations are *conditionally exchangeable*. In that case, treatment assignment is independent of the potential outcomes, given measured covariates. This assumption is typically referred to as the *no unmeasured confounding assumption*, because it implies that no confounding remains after conditioning on those measured covariates. Such conditioning is most easily achieved by calculating separate treatment effects for the different covariate strata. However, it becomes unfeasible when there are several potential confounders, of which some can be continuous, for then, the number of subgroups becomes very large. Pretreatment differences that confound the treatment-outcome relationship can then be adjusted by means of regression. Using regression, the missing potential outcomes can be predicted for each possible covariate combination. In our simple example with only one covariate, we can postulate a model for the average math achievement score as a function of Z and X; for instance:

(2)

$$E[Y|Z,X] = \beta_0 + \beta_1 Z + \beta_2 X$$

Regression adjustment is appropriate in simple settings (see e.g., Rubin, 2001). However, when adjustment for a large number of covariates is needed, the degrees of freedom for the estimation of the treatment effect can become relatively small. Moreover, when treated and untreated subjects are very different in pretreatment characteristics, then regression adjustment is prone to extrapolation, a problem that may easily go undiagnosed (Vansteelandt & Daniel, 2014). In view of these concerns, alternative techniques to deal with confounding have been developed and have been increasingly used in social and behavioral sciences, of which propensity score methods are very popular. Propensity score methods enable adjustment

for confounding in a way that is not susceptible to extrapolation and allows for diagnosing lack of overlap in pretreatment characteristics between treatment groups (Ho, Imai, King, & Stuart, 2007; Schafer & Kang, 2008).

Inverse probability weighting. Propensity score methods enable adjustment for confounding by explicitly modelling, and thus acknowledging, that the probability to be assigned to the treatment condition may depend on pretreatment characteristics. In particular, the propensity score (PS) is the probability to be in the observed treatment condition, given the pretreatment covariates: $\Pr(Z=1|X)$ (Rosenbaum & Rubin, 1983). Propensity scores summarize all potential confounders into a scalar summary, and thereby allow for simple adjustment methods based on stratification, regression adjustment, matching or weighting. In this article, we focus on inverse probability weighting (IPW), since this is the default estimation method for marginal structural models in a time-varying setting. IPW changes the empirical distribution of the observed outcomes to make it representative of the complete dataset of potential outcomes. Each subject is weighted by $W=P[Z|X]^{-1}$, being the inverse of the probability to be in the treatment condition he or she is really in, given the covariates. The weights are thus defined as $P[Z=1|X]^{-1}$ for the treated subjects and as $P[Z=0|X]^{-1}$ for the control subjects.

The use of IPW is comparable to the use of survey sampling weights which are routinely employed to make samples representative of a particular population (Austin, 2011). IPW creates a pseudo-population. This is a population in which each subject appears under each treatment condition, thereby mimicking data as they would have looked if they originated from a RCT (provided the conditional exchangeability assumption holds). Each subject contributes W copies of itself to the pseudo-population. For instance, in Table 2, student 1 has a propensity score of 1/3 because 1 out of 3 students in the native language group is retained. His data are thus weighted 3 times, one time for himself and 2 additional times for students 2 and 3 for whom $Y(1)$ is missing. In contrast, the data for students 2 and 3 are weighted by the reciprocal of 2/3,

i.e. 1.5 times, 1 time for themselves, and 0.5 times to make up for student 1 for whom $Y(0)$ is missing. The weighting ensures that treatment status becomes independent of measured pretreatment characteristics in the weighted dataset, as in a RCT, although it does not eliminate possible associations with unmeasured pretreatment characteristics, unlike in a RCT. The weighting thus eliminates the arrow leaving from X into Z in Figure 1, so that X no longer induces confounding. The weighted data can therefore be analyzed as an RCT, ignoring data on measured confounding variables. It follows in particular that the MSM (1) can be fitted by regressing Y on Z on the weighted data (Hernán, Brumback, & Robins, 2000). In our artificial example, using the aforementioned weights, the parameters of the MSM can be estimated as: $E[Y(1)]$ = ((3 x 47) + (1.5 x (38 + 40)))/6 =43 , and $E[Y(0)]$ = ((1.5 x (54 + 48)) + (3 x 40))/6 = 45.5, which again returns the ATE using the complete dataset. Note that, using IPW, we have actually doubled our number of observations. On average, each student is replaced by two copies in the pseudo-population: one under the retention condition and one under the control condition.

Studies with a Time-varying Treatment

In observational studies, the treatment is often not fixed in time. Examples of time-varying treatments are drug use, behavioral therapy, instructional programs, or grade retention, which is the empirical example in this article. Figure 2 illustrates a time-varying confounded setting with two treatment occasions. In a time-varying setting, we have a vector of time-varying covariates L_t at each time t (e.g., prior math achievement), which may affect variables after time t. X_0 now indicates a vector of time-fixed covariates (e.g., gender, month of birth) which may affect all other variables. Important is that the variables in L_t are measured prior to Z_t. We use overbars to indicate the history of treatments $\bar{Z}$ and covariates $\bar{L}$, prior to t.

The potential outcomes framework in a time-varying setting becomes more complicated because now there are potential outcomes for every treatment pattern. In grade retention research, this means for example that we have a potential outcome for each school trajectory a student may follow. In principle, every possible school trajectory forms a different treatment pattern, associated with a potential outcome. The average treatment effect at a given time *t* may then be defined as the difference between the marginal means of two or more potential outcomes at time *t*. For example, $E[Y_t(\bar{z})]$-$E[Y_t(\bar{z}')]$ represents the difference at time *t* between the potential outcome if the population was treated according to regime $\bar{Z}$ (e.g., Year 6 math achievement after being retained once, in kindergarten) and the potential outcome if the population was treated according to treatment regime $\bar{Z}'$ (e.g., Year 6 math achievement after never being retained).

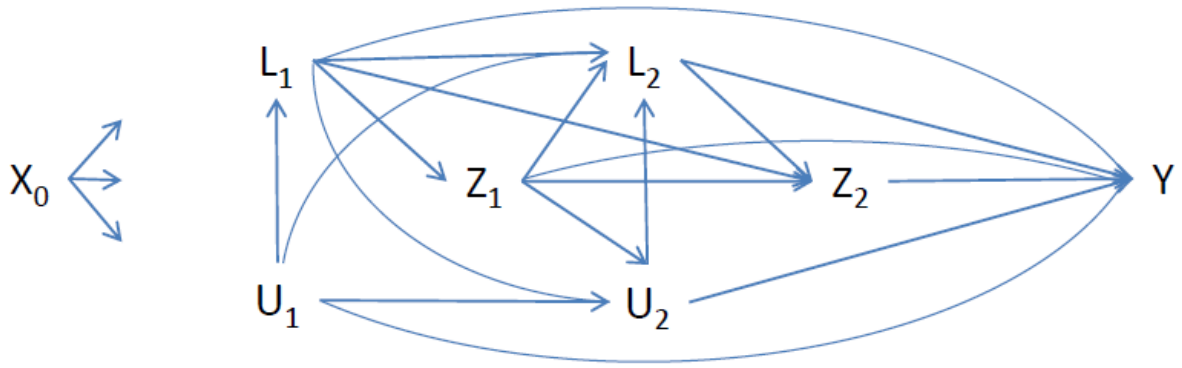


Figure 2. Time-varying causal model with two treatment occasions Z_1 and Z_2. L_1 and L_2 represent a vector of time-varying covariates at time t=1 and t=2 respectively, which may affect variables after time *t*. X_0 indicates a vector of time-fixed covariates which may affect all other variables. *Y* represents the end-of-study outcome.

The Problem with Standard Methods

In a time-varying setting, standard regression methods to adjust for confounding can become problematic for two reasons, both resulting from the inclusion of time-varying

covariates in the regression model. First, the treatment at time *t* might affect potential confounders at later times (we will refer to these as intermediate confounders as they might have been affected by earlier treatment). Simply controlling for earlier treatment or intermediate confounders through standard regression (i.e., by means of including these variables as covariates in the outcome model), blocks the effect of earlier treatment on the outcome via those intermediate variables, so that only part of the overall treatment effect is maintained (Rosenbaum, 1984). Second, controlling for an intermediate confounder can induce collider stratification bias. A collider is a common effect of two variables. In Figure 2, L_2 is a common effect of Z_1 and U_2. Also, U_2 is a common cause of L_2 and Y. Thus, conditioning on L_2 might induce collider stratification bias, a form of selection bias that alters the relationship between Z_1 and Y, and which may even lead one to systematically conclude that Z_1 and Y are associated in the absence of an effect (Hernán, Hernández-Diaz, & Robins, 2004; Morgan & Winship, 2015). For example, imagine that kindergarten retention has no effect on math achievement. Consider an unmeasured variable, illness, and a measured variable, low self-esteem. Now suppose kindergarten retention and illness are the only two causes of a low self-esteem, and that these two variables are not themselves related in the population. Then within the group of children with a low self-esteem, a negative correlation may be expected between kindergarten retention and illness: children who have low self-esteem but who were not retained must have been ill (as it must be their illness that explains their low self-esteem). Even if kindergarten retention were not to affect math achievement and even if retention were randomized, adjusting for self-esteem might thus induce an association via underlying illness. Specifically, when looking at the group of children with low self-esteem, kindergarten retention will show a positive relation with math achievement, because the children in that group who were retained in kindergarten were less likely ill, and therefore were more likely to have higher math scores.

Conditional exchangeability in a time-varying setting. In a time-varying setting, conditional exchangeability means that at each time *t*, there are no prognostic factors of the outcome that are differentially distributed between the treatment and the control group, given the treatment history $\bar{Z}_{t\text{-}1}$, the baseline covariates X_0, and the covariate history $\bar{L}_t$. This is also called the sequential randomization assumption. The assumption would hold if at each time, treatment were randomly assigned with randomization probabilities that are possibly depending on the treatment and confounder history (Robins & Hernán, 2008).

Marginal Structural Models to deal with Time-varying Confounding

In the previous section, we explained how both ignoring time-varying confounding and using standard regression to control for time-varying confounding can lead to biased results. Robins, Hernán and Brumback (2000) introduced inverse probability weighting under marginal structural models to deal with time-varying confounding. This has become a popular approach in epidemiology, however, relatively few applications are found in the behavioral sciences (e.g., Barber, Murphy, & Verbitsky, 2004; Hong & Raudenbush, 2008; VanderWeele, Hawkley, Thisted, & Cacioppo, 2011). The approach proposed by Robins et al. (2000) is to mimic data from a sequentially randomized experiment by reweighting the treatment groups at each point in time. Indeed, the weights are such that, after reweighting, as in a sequentially randomized experiment, treatment at each time *t* is no longer associated with the history of treatment and potential confounders prior to that time, but the effect of treatment on outcome is the same. The weights make use of time-varying propensity scores that model the relationship between treatment and the history of treatment and potential confounders prior to each time, but do not adjust for covariates measured at later times in order to avoid similar problems as with regression adjustment. Because the weighting creates balance (with regard to the history of treatment and potential confounders) between treatment groups at each time point, treatment effects can now be estimated by regressing the outcome on the treatment history using the

reweighted data, without further regression adjustment for potential confounders. By not conditioning on time-varying covariates in the model for the outcome, the two problems inherent to standard methods (i.e. blocking intermediate treatment effects via those covariates; collider stratification bias) are overcome (Robins et al., 2000; Hong & Raudenbush, 2008). The full set of assumptions underlying weighting is given in the Appendix.

We can formulate a MSM for the end-of-study outcome *Y* in a setting with two measurement occasions as

(3)

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

where z_1 and z_2 are binary (0/1) treatment indicators. Here, β_0 is the marginal mean of the outcome when the population was never treated (0,0) and β_1 corresponds to the average change in *Y* under treatment regime (1,0) (relative to treatment (0,0)), i.e., when the population was treated at time 1 but not at time 2. Accordingly, β_2 and β_3 refer to the average change in *Y* under treatment regimes (0,1) and (1,1) (relative to treatment (0,0)), i.e., when the population was only treated at time 2 or at both times 1 and time 2, respectively.

The required sample size for MSMs is dependent upon how strongly individuals within different treatment patterns are different from each other in terms of measured time-varying covariates. When individuals in different treatment patterns are comparable, as would be the case if treatment were randomly assigned at each time, then similar sample sizes to a randomized experiment with the same number of time points and the same outcome model would be needed. As usual, smaller sample sizes are then required when the treatment effect is large. The choice of model can also be very influential. For instance, when very few subjects switch treatment from time 1 to time 2, then the data will carry very little information about β_3 so that large sample sizes may be needed to fit model (3). This can be remedied by deleting the interaction term $z_1 z_2$, at the cost of needing to assume this term is 0. When individuals in

different treatment patterns are very different in terms of measured time-varying covariates, then large sample sizes may be needed (e.g., at least 500 subjects) as the inverse probability weights may then become extreme (close to 0 or 1) for some individuals. Unfortunately, no general sample size requirements can be given. The required sample size is heavily dependent on the choice of model, number of time points, magnitude of the treatment effect, outcome variation, and the propensity scores. The width of confidence intervals for the parameters of interest gives the necessary guidance as to how reliable the results are.

The procedure for fitting the MSM (3) consists of three steps. First, at each time *t*, a weight is calculated for each subject. Weights can be calculated using the 'ipw' package in R by making use of time-varying propensity scores (van der Wal & Geskus, 2011). Second, at each time *t*, the weights and the resulting covariate balance are evaluated. The first two steps involve an iterative process of evaluation of the weight distribution and covariate balance and modification of the weight model. In the final step, the measurements for each subject at each time are reweighted by the time-specific weights estimated in the previous steps to estimate the parameters of the MSM.

Estimation of the weights. At each time *t*, a weight *W* is estimated for each subject:

(4)

$$W_{\mathrm{t}} = \prod_{\mathrm{s}=1}^{\mathrm{t}} \mathrm{P}[Z_{\mathrm{s}} | \, \overline{Z}_{\mathrm{s}-1}, \bar{L}_s, X_0]^{-1}$$

Equation (4) is a product of time-specific weights, from baseline up to time *t*, for each subject. Each term in the denominator represents the probability of the observed treatment Z_s at time $s<t$, given the treatment history up to the previous time point $\bar{Z}_{s\text{-}1}$, the covariate history $\bar{L}_s$, and the baseline covariates X_0. Table 3 illustrates the weights at each time *t* for three hypothetical subjects.

Table 3

Unstabilized and stabilized weights of three hypothetical subjects.

ID	t	Z_t	W	SW
1	1	1	$W_1 = P[Z_1=1\|X, \bar{L}_1]^{-1}$	$W_1 = (P[Z_1=1])/ (P[Z_1=1\|X, \bar{L}_1])$
1	2	1	$W_2 = W_1 / P[Z_2=1\|Z_1=1, X, \bar{L}_2]$	$W_2 = W_1 (P[Z_2=1\| Z_1=1])/ (P[Z_2=1\|Z_1=1, X, \bar{L}_2])$
1	3	1	$W_3 = W_2 / P[Z_2=1\|Z_2=1, Z_1=1, X, \bar{L}_3]$	$W_3 = W_2 (P[Z_3=1\| Z_2=1, Z_1=1])/ (P[Z_3=1\| Z_2=1, Z_1=1, X, \bar{L}_3])$
2	1	0	$W_1 = P[Z_1=0\|X, \bar{L}_1]^{-1}$	$W_1 = (P[Z_1=0])/ (P[Z_1=0\|X, \bar{L}_1])$
2	2	0	$W_2 =W_1 / P[Z_2=0\|Z_1=0, X, \bar{L}_2]$	$W_2 = W_1 (P[Z_2=0\| Z_1=0])/ (P[Z_2=0\|Z_1=0, X, \bar{L}_2])$
2	3	1	$W_3 =W_2 / P[Z_3=1\|Z_2=0, Z_1=0, X, \bar{L}_3]$	$W_3 = W_2 (P[Z_3=1\| Z_2=0, Z_1=0])/ (P[Z_3=1\|Z_2=0, Z_1=0, X, \bar{L}_3])$
3	1	0	$W_1 = P[Z_1=0\|X, \bar{L}_1]^{-1}$	$W_1 = (P[Z_1=0])/ (P[Z_1=0\|X, \bar{L}_1])$
3	2	0	$W_2 =W_1 / P[Z_2=0\|Z_1=0, X, \bar{L}_2]$	$W_2 = W_1 (P[Z_2=0\| Z_1=0])/ (P[Z_2=0\|Z_1=0, X, \bar{L}_2])$
3	3	0	$W_3 =W_2 / P[Z_3=0\| Z_2=0, Z_1=0, X, \bar{L}_3]$	$W_3 = W_2 (P[Z_3=0\| Z_2=0, Z_1=0])/ (P[Z_3=0\| Z_2=0, Z_1=0, X, \bar{L}_3])$

W = weight; *P* = probability; Z_t = treatment at time *t*; *X* = vector of baseline covariates; $\bar{L}$ = covariate history; *SW* = stabilized weight

With time-varying treatments, the product of the probabilities can vary greatly. For example, when some of the probabilities (which are included in the denominator of the weight) are close to zero, some of the inverse probability weights can become very large. In later steps of the procedure, this can lead to imprecision in the estimators of the parameters of the MSM. Robins et al. (2000) therefore strongly recommend to use stabilized inverse probability weights (SW_t), which are obtained by multiplying the time-specific weights with $P[Z_s \,|\overline{Z}_{s-1},\, X_0]$, being the probability to be in the observed treatment condition, given the treatment history. Additionally, one may include baseline covariates in the numerator of the weight model, which makes the numerator and the denominator of the weights agree better, and consequently leads to reduced variability in the weights. By including treatment history and covariates in the numerator, the weighting mimics a RCT in which the probabilities are allowed to vary according to the treatment trajectory and baseline covariates, but have no residual dependence on the time-varying covariate history. Table 3 illustrates the stabilized weights for three hypothetical subjects.

Equation (5) gives the formula for stabilized weights SW_t.

(5)

$$SW_{\mathrm{t}} = \prod_{\mathrm{s}=1}^{\mathrm{t}} \frac{\mathrm{P}[Z_s|\overline{Z}_{\mathrm{s}-1},\, X_0]}{\mathrm{P}[Z_s|\overline{\mathrm{Z}}_{\mathrm{s}-1}, \bar{L}_s, X_0]}$$

The denominator is the same as in equation (4), and the numerator is equal to the denominator but without adjusting for the time-varying covariate history. Robins et al. (2000) demonstrate that the use of stabilized weights is valid as long as the MSM correctly includes the baseline covariates X_0. This approach mimics a RCT in which the treatment trajectories do not necessarily have the same sample size. Stabilized weights maintain the original sample size in the weighted sample and alleviate the problem of extreme weights.

Evaluation of the weights and the resulting balance. The next step is to evaluate the weights and the balancing properties. The weights can be evaluated by examining their distribution at each time *t*. Because stabilized weights keep the size of the pseudo-population on average equal to the size of the study population, the mean of the weights is expected to be one. When the mean is far from one, this might indicate a violation of the so-called positivity assumption or might signal misspecification of the propensity score model (Cole & Hernán, 2008). The *positivity assumption* refers to the condition that all treatment options are possible at every level of the covariates. When there is a covariate combination at which it is impossible to be treated, a structural zero probability of receiving treatment will occur. One way to deal with nonpositivity is to restrict the sample to subjects who exceed a minimum probability of treatment or no treatment at each time point given the baseline covariate information (Cole & Hernán, 2008). In the empirical example presented in this article, the analytic sample was restricted to children who had a probability of at least 5% of being retained in kindergarten.

As was mentioned, stabilized weights are one way to reduce extreme weights. When there are extreme weights after stabilization, one may choose to further reduce the variability by means of resetting or truncating the weights below a chosen percentile to the weight at that

percentile. For example, weights below the first percentile can be set to the weight at percentile 1 and weights above percentile 99 can be set to the weight at percentile 99. Truncation leads to reduced variability in the weights but may leave residual confounding bias. If truncation is considered, it is important to carefully evaluate the bias-variance trade-off by means of exploring the behavior of different truncation values with regard to the distribution of the weights and the covariate balance (Cole & Hernán, 2008).

When the weight model is correctly specified and the sample is sufficiently large, the weighted sample is balanced in terms of measured covariates across the treatment groups at each time *t*. Balance refers to the similarity of the covariate distributions (Harder, Stuart, & Anthony, 2010). For example, in a subset of children with the same probability to be retained at time *t*, retained and promoted children have similar distributions of the covariate history at time *t*. Balance diagnostics are frequently examined in propensity score studies, however, in MSM studies this step is often ignored. This may be due to the complexity of the time-varying weights. Nevertheless, in the typical case in which the true probability to be treated is not known, it is important to verify that the balancing property holds, i.e., that the weighted treatment groups are similar with regard to their covariate histories. Covariate balance between treatment regimes can be examined by evaluating their standardized mean differences (*SMD*) in the weighted sample. The *SMD* is the difference in covariate means, divided by the pooled standard deviation (Rubin, 2001).

Specifically, suppose we want to evaluate balance at time 2 between treatment regimes (0,0) and (0,1). We then first calculate the product of time-specific weights up to time 2. Next, we use these weights to calculate the weighted *SMDs* between the two regimes with regard to baseline covariates (that are not incorporated into the numerator of the weights), time-1 covariates and time-2 covariates. A *SMD* of 0.25 is commonly considered as a maximum acceptable value of imbalance (Ho et al., 2007); however, this should be considered as a rule

of thumb rather than a strict cut-off (Harder et al., 2010). Another way of examining balance is to plot the covariate distributions of the treatment and control group before and after weighting at time *t*. The central tendencies of the covariates should be closer after weighting. If the balance diagnostics indicate serious imbalance for some covariates, then that result may signal that the propensity score model is not correctly modelling those covariates. It should then be modified by including interactions or higher order terms with regard to those covariates, by transforming covariates, or by exploring machine learning approaches to estimating propensity scores (see Austin, 2011; Caliendo & Kopeinig, 2008; Cham & West, in press; Lee, Lessler & Stuart, 2010). Importantly, note that balance of the treatment groups at each time should only be assessed with respect to the history of treatments and covariates at that time, and not with regard to future covariates.

Estimation of the MSM. The MSM can be fitted by regressing the outcome at each measurement time on the predictors in the MSM (treatment history, time, and possibly baseline covariates), thereby weighting the contribution of each subject at each time *t* using the weights (4) or (5) retrieved from the previous steps. Important here is that the standard errors must account for the fact that there are repeated outcome measures per individual over time, as well as for the uncertainty in fitting models for the weights. Robust variance estimation, ignoring the estimation of the weights, is most commonly used as it is relatively simple and known to deliver standard errors that are not too small, although generally somewhat conservative (Joffe, Ten Have, Feldman, & Kimmel, 2004; Robins et al., 2000). This can be done by fitting the MSM using software for generalized estimating equations (GEE) with an independence working correlation structure (Zeger & Liang, 1986). Weighted analyses with robust (conservative) standard errors could also be performed using typical software for survey analysis (Robins & Hernán, 2008). GEE with an independence working correlation structure estimates the parameters in a way that treats each subject at each time as separate observations.

However, in the estimation of the standard errors, one accounts for the correlation of observations within a subject (Robins, 2000). GEE is here preferred over the use of mixed models, which are more familiar to social and behavioral researchers, because the former does not invoke the implicit assumption that the outcome at a particular measurement occasion is independent of future treatment exposure, which is often not justified in the time-varying treatment setting (Vansteelandt, 2007). For the same reason, GEE analyses should not use other choices of working correlation structure than independence (Vansteelandt, 2007).

Empirical Example

We illustrate the use of MSMs with an empirical study about the effects of grade retention on the mathematics achievement scores throughout primary education. Grade retention is the practice of holding back children or students who have not mastered their grade's curriculum. By letting them repeat a grade, they have more time to master particular knowledge, skills or competencies, which could prevent failure and frustration in later years. In European countries like Belgium, Spain, France, the Netherlands and Luxembourg, this idea is supported by the teaching profession, the school community and parents (Eurydice network, 2011). Over the past decades, grade retention has been a major issue in the field of educational effectiveness research. The general premise in the research base is that children do not benefit from repeating a grade. This belief largely stems from widespread meta-analyses (i.e., Holmes (1989) and Jimerson (2001)), in which grade retention was characterized as an ineffective or, in some cases, even harmful practice. This conclusion gave rise to policy interventions aiming to reduce retention rates. However, the meta-analyses in turn have been criticized for being based on studies that show significant methodological shortcomings (e.g., Allen, Chen, Willson, & Hughes, 2009; Lorence, 2006), questioning the validity of these conclusions. A recent generation of studies has used techniques to adjust for pre-treatment differences. Recent studies using same-age comparisons have found that at-risk children may benefit more from

receiving intellectual challenges that are offered in a higher grade (Goos, Van Damme, Onghena, Petry, & De Bilde, 2013; Hong & Yu, 2007; Hong & Raudenbush, 2005; Hong & Raudenbush, 2006; Vandecandelaere et al., 2015a; Vandecandelaere, Vansteelandt, De Fraine & Van Damme, 2015b; Wu et al., 2008a; Wu, West & Hughes, 2008b). Recent research using same-grade comparisons has demonstrated that grade repeaters, compared to their younger grade-mates, score better in math during the retention year. In the long term, the advantage disappears or even reverses (Dong, 2010; Goos et al., 2013; Moser, West & Hughes, 2012; Wu et al., 2008a).

An important limitation in previous research is that grade retention usually is treated as a fixed, one-time intervention. Typically, the achievement of students who were retained in a specific grade is compared with that of equivalent peers who were not retained. In reality, however, low performing students who were promoted after the grade of interest are much more likely to repeat a later grade, which demonstrates the time-varying nature of grade retention. Simply ignoring students' post-treatment trajectories impedes a clear understanding of post-treatment outcome differences. When outcomes are compared at a given number of years after retention, it is possible that children who were promoted after the grade of interest are in the same grade as children who were retained in the grade of interest. In our empirical example, 31.89% of the children with at least 5% probability to be retained in kindergarten but who promoted anyway were retained in first grade instead. In this scenario, at least a part of the 'promoted' group should more accurately be considered as a 'delayed retention' group. This delayed treatment situation should be accounted for when children who were retained after the grade of interest are included in the sample. On the other hand, estimating treatment effects based on samples excluding students who were not continuously promoted after the treatment, only tells part of the story and may induce selection bias. Vandecandelaere et al. (2014) elaborated on this issue and, using a same-age comparison with the present sample, compared

math development after kindergarten retention with math development in two alternative treatment regimes: first-grade retention and continuous promotion. The results indicated that retainees would perform better if they would be promoted each year; kindergarten retention was found to be less harmful than first-grade retention. The study was limited in that the alternative trajectories were not handled together in one model, and that only two retention regimes were addressed. Moreover, retention was still treated as a time-fixed event in the sense that the treatment groups were assumed to be exchangeable within levels of baseline covariates, assessed before anyone was retained.

The research questions that we address in the present article are:

a) How do children who are at least 5% at risk to be retained in kindergarten develop throughout primary education with regard to their mathematics achievement, under the following treatment patterns: no retention, kindergarten retention, first-grade retention, and second-grade retention? The children compared are of the same age.
b) Does the impact of grade retention vary by the grade in which the student was retained?

Each year, a child can be retained or can be promoted to the next grade. In Flanders, there is no national standardized testing. In general, the retention decision heavily relies on the teachers' consideration of the child's cognitive and non-cognitive skill set. The retention decision lies with the parents who receive advice from the teacher-team and the pupil guidance center. The four treatment regimes $\bar{Z}$ considered in this study are presented in Table 4. Of course, more treatment regimes are possible. For example, children can be retained after Year 3 or might spend one or more years in a separate school for special education. Because of the small number of students in these trajectories, the data for children in these groups were censored. Our approach for dealing with censored students (along with the corresponding assumptions) is explained later in this article.

Table 4

Treatment regimes $\bar{Z}$ considered in this study.

	Year 1	**Year 2**	**Year 3**	**Year 4**	**Year 5**	**Year 6**
K retention (KR)	*K*	*G1*	*G2*	*G3*	*G4*	*G5*
G1 retention (G1R)	G1	*G1*	*G2*	*G3*	*G4*	*G5*
G2 retention (G2R)	G1	G2	*G2*	*G3*	*G4*	*G5*
No retention (noR)	G1	G2	G3	G4	G5	G6

Italic= one year delay compared to continuously promoted children. K= kindergarten; G1-G6 = Grade 1 through Grade 6

Students in different treatment regimes were compared at multiple time points. As we explained in the introduction of this article, two approaches can be used: a same-age comparison or a same-grade comparison. Because our research question involves the causal question of how at-risk children would perform under four treatment patterns, we conducted a same-age comparison. All math achievement scores (seven measurement occasions) were equated on the same scale. The test scores were vertically linked using Item Response Theory and estimated with a Bayesian model. Vertical equating of scores enabled us to assess the mathematics achievement scores of each child over time, and to compare the mathematics scores of children from different grade levels.

Data

The research questions were answered through analyses of the data from the large-scale longitudinal SiBO-project. A cohort of approximately 6,000 Flemish children was followed, from kindergarten (age 5-6; school year 2002-2003) until eighth grade (age 13-14; school year 2010-2011). For ease of presentation, in the remainder of this article, school years 2002-2003 to 2008-2009 are referred to as Year 0 to Year 6 respectively. In Year 0, all children were in kindergarten. In Year 6, continuously promoted children were in Grade 6, and grade repeaters were in Grade 5. This is illustrated in Table 4.

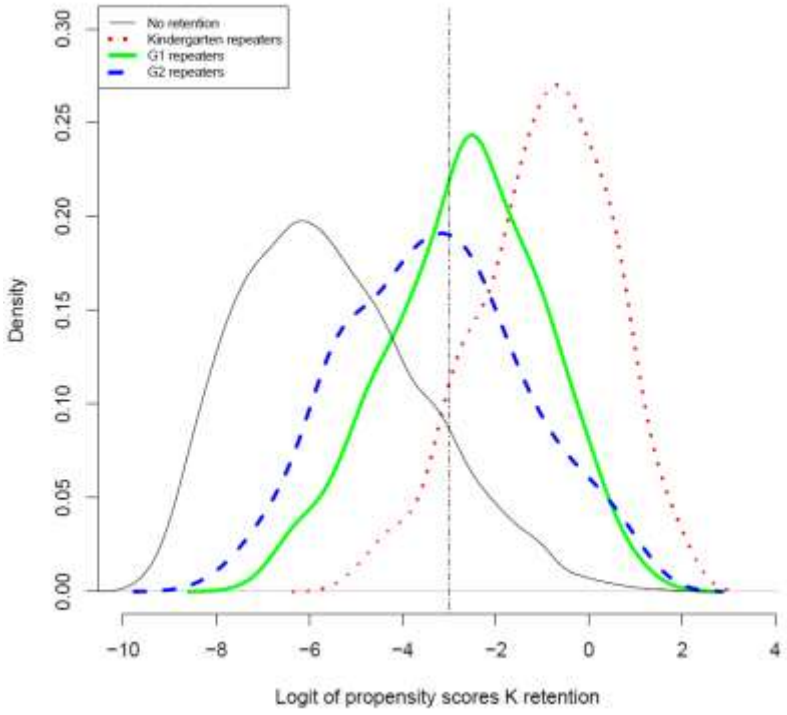
A wide variety of information was collected at the school, teacher, class and student level. Achievement tests for mathematics were administered at the end of each year in primary school. Each test consisted of 50 to 80 items and covered the following domains: number sense (e.g., "In the middle of 12 and 16 lies the number …… "), number procedures (e.g., "Fill in: 20 = 4 + 5 + 6 + ……"), measurement (e.g., "What time is missing in the following time table? All trains take the same time to travel the same distance."), geometry (e.g., "How many corners does a rectangle have?"), and applied math problem solving (e.g., "Polly is saving money for a new bike. The bike costs EUR 425. She already saved EUR 280. How much does she still need?"). Cronbach's α coefficients of the mathematics tests ranged between .88 and .93.

Retention policies vary across regions within a country and across countries. To conceptualize the present study, information about the Flemish educational context is provided in the supplementary online material.

Selection of the Analytic Sample

All children in the SiBO dataset who were for the first time in kindergarten in Year 0 were considered (N=5,616). Children in alternative schools and children who had missing treatment or outcome information from Year 0 or Year 1 onwards were excluded from the sample, before calculating treatment and censoring weights. This resulted in a sample of 4,196 children of whom 846 (20%) were retained at least once in elementary school. To guarantee sufficiently similar treatment groups in view of the positivity assumption, we chose to further restrict the analytic sample to those children who, on the basis of baseline covariates, were predicted to have at least a 5% probability of being retained in kindergarten. This was done prior to fitting the final weights and outcome models. This strategy prevented violation of the positivity assumption, and also implies that the effect of retention is evaluated for a group of children for whom retention may be meaningful. Using logistic regression, we estimated the probability of being retained in kindergarten as a function of pretreatment characteristics. As

we explain later in this article, the potential set of characteristics was initially identified based on a literature review. Stepwise logistic regression was used to select the characteristics in the final prediction model. The 5% cut-off was based on two criteria: 1) the overlap in the distributions of the probability to be retained in kindergarten of the four treatment groups and 2) the covariate balance after applying time-varying IPW. We started with a 1% cut-off and progressively increased this percentage until we achieved acceptable weight and balance diagnostics at a 5% cut-off. The density plots of the distributions of the propensities for kindergarten retention before and after the selection are shown in Figure 3. The 5% probability criterion gave a good overlap between the treatment regimes under consideration. It is clear from Figure 3 that the probability to be retained in kindergarten is most predictive for kindergarten retention. The selection resulted in an analytic sample of 857 children of whom 200 were retained in kindergarten, 199 were retained in first grade, and 61 were retained in second grade[i]. By selecting this sample, our empirical example reports on the effects of early grade retention for the population of children who had a probability of kindergarten retention of at least 5%. In the remainder of this article, the term 'at-risk children' refers to this group.



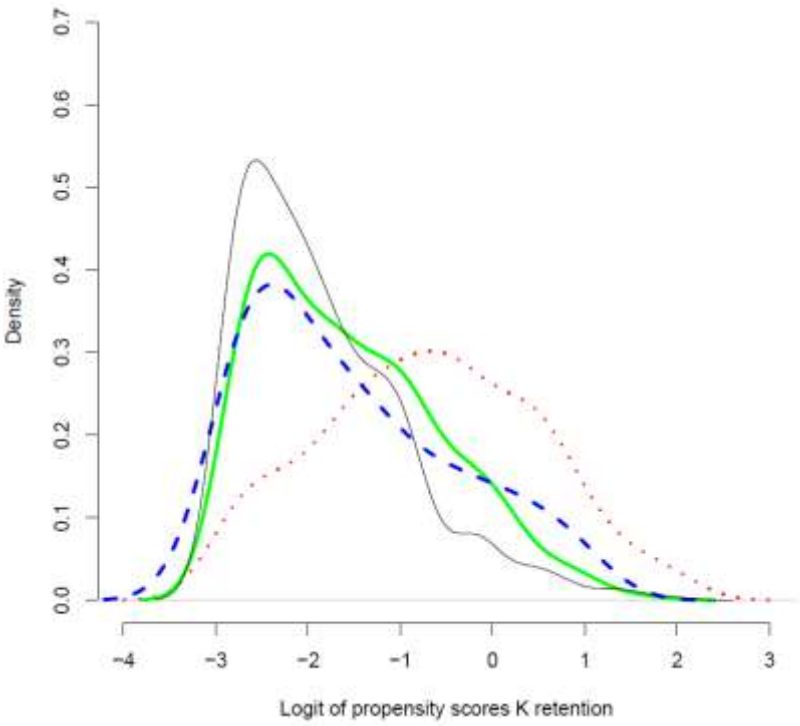


Figure 3. Density plots of the logit of propensity scores of kindergarten retention before (upper) and after (lower) selection. The vertical line indicates the 5% probability of retention cut-off. On average, the retained students have a higher propensity score compared to the promoted students. After the selection, the distributions demonstrate a clear overlap.

Application of MSMs

Estimation of the weights. The weights were estimated as a function of both time-fixed information and time-varying information from Year 1 through Year 3. Variables were gathered from official records, achievement tests, teacher questionnaires about the child, and parent questionnaires. Based on a literature review on predictors of early grade retention and academic growth, we identified the potential set of covariates that might be confounders of the retention-achievement relationship. The covariates in the final prediction model were selected by means of stepwise logistic regression. The final model yielded 22 pretreatment covariates. Table 5 provides a list of the covariates. The covariates were measured every year[ii], except for Year-2 anxious behavior and Year-3 cooperative behavior, prosocial behavior and attitude to work. To allow these covariates to be time-varying, we used the most recent available score (last score carried forward). For example, unmeasured Year-2 anxious behavior was replaced with

measured Year-1 anxious behavior. This strategy is used in the fitting of MSMs because it is not obvious how to impute incomplete time-varying covariate data in a way that is compatible with the postulated propensity score models and the MSM. Furthermore, we assumed these psychosocial and attitudinal variables were relatively stable within the same person. If this assumption were violated, then the results from our MSM analysis might be biased, but only to the extent that these specific covariates both changed over time and confounded the association between treatment and outcome. In the initial dataset (N=5,616), other missing values in the covariates were multiply imputed using chained equations, under a missing at random (MAR) assumption (Azur, Stuart, Frangakis, & Leaf, 2011). The mean proportion of missing covariate information in this dataset increased from 9.9% (sd=5.9) in Year 1 up to 27.3% in Year 3 (sd=5.6). The imputation was performed using the MICE package in R (Buuren & Groothuis-Oudshoorn, 2011). Given the computational burden, the data were imputed ten times. This was deemed sufficient in the sense that it yielded a relative efficiency of at least 95% for all parameters (Rubin, 1987). Subsequent analysis steps were performed on each imputed dataset and combined using Rubin's (1987) rules.

Of importance, note that our analysis allows for the math score in Year t in the outcome model to be a potential confounder in Year t+1 in the weight model. For example, mathematics achievement measured in Year 1, which is the outcome in Year 1, was used as a Year-2 covariate in the weight model (a potential confounder between the relationship of grade retention in Year 2 and later math achievement). At the same time, math achievement in Year 1 was the outcome in Year 1. The adjustment for previous math scores was important because our analysis aims to contrast retained and promoted children with similar histories, in particular similar pre-treatment mathematics scores.

The weights were created using the 'ipw' package in R (van der Wal & Geskus, 2011) (the R code is provided in the online supplementary material). For continuously promoted

children, the treatment trajectory of continuous promotion from Year 1 to Year 6 was coded as (0,0,0,0,0,0). For kindergarten, first-grade and second-grade repeaters, the treatment trajectories were coded (1,1,1,1,1,1), (0,1,1,1,1,1) and (0,0,1,1,1,1) respectively (see Table 4): 0 means no delay, whereas 1 means delay. In other words, we considered only treatment regimes in which children, once retained, stayed delayed for one year compared to promoted children. For each year, within the subgroup of children who were not (yet) retained, weights were estimated by means of a logistic regression predicting the risk to be retained the next year. From the retention year onwards, the treatment weights of retained children remained constant. This is because the probability of being delayed at time *t* for children who were retained at time *t* -1 was equal to 1, hence, their time-specific weight at time *t* is 1 (i.e., regardless their pre-treatment history, all children who were retained at time *t* stay delayed from time *t* onwards). From Year 3 onwards, the weights remained constant for all children. The probability of being in the particular treatment regime, given the treatment history, is equal to 1 from then onwards (i.e., as can be seen in Table 6, there is no longer variation in the treatments from Year 3 onwards).

As mentioned above, the estimation and evaluation of the weights was a recursive process. We first estimated the weights as a function of the covariates, without any interactions. Next, we repeatedly extended the weight model by including meaningful interactions, followed by an evaluation of the weights and balance diagnostics. In particular, previous research has shown an interaction between month of birth and achievement scores with regard to the probability to be retained (Vandecandelaere et al., 2015b). At-risk children who are older have on average a lower math score compared to at-risk children who are younger. Other interactions that were considered were the interaction between month of birth and socio-economic status, and between socio-economic status and math achievement, since these covariates are known to be important predictors of early grade retention. The model including all covariates and an interaction of time-varying math achievement and month of birth led to the best weight and

balance diagnostics. We additionally included covariate-time interactions in the propensity score model to allow the coefficients for every time-varying covariate in that model to vary by time. Because some extreme weights occurred in two of the ten imputed datasets, weights below percentile 0.1 and above percentile 99.9 were truncated.

Table 5

Description and information source of time-fixed and time-varying covariates.

Name	Description	Information source
Time fixed covariates		
Month of birth	Month in which the child was born	PQ
Gender	Gender of the child	PQ
SES	Socioeconomic status: score computed by means of confirmatory factor analysis based on 7 indicators: (1) Highest diploma father, (2) Highest diploma mother, (3) Employment status father, (4) Employment status mother, (5) Occupational level father, (6) Occupational level mother and (7) Income.	PQ
Home Language	Language spoken at home (1: only other language; 2:Dutch and other language; 3: only Dutch)	PQ
Ethnicity	Ethnic background (1: the Maghreb and Turkey; 2: other non West-European; 3: West-European; 4: Belgium)	PQ
High risk student	Identified by the Flemish government as high risk student: A student meeting at least one out of five equal opportunity indicators: 1. The parent is a barge skipper, fairground worker, circus artist or a caravan dweller. 2. The mother has no qualification of secondary education. 3. The child is temporarily or permanently living outside the family. 4. The family lives on a replacement income. 5. The language the child speaks with his family at home is not Dutch.	OR
Time-varying covariates		
Math	Mathematics achievement	AT
Independent	Scale score of 4 items measuring ability to take initiative and to act autonomously with regard to school tasks	TQC
Cooperative	Scale score of 4 items measuring socially responsible behavior and ability to deal with authority in the classroom	TQC
Hyperactive	Scale score of 4 items measuring display of restless classroom behavior and lack of attention	TQC
Asocial	Scale score of 4 items measuring preference for solitary play and tendency to isolate from peers	TQC
Aggressive	Scale score of 4 items measuring the frequency of negative and dominating behavior	TQC

Attitude work	Scale score of 3 items measuring accuracy and ability to stay on task for a period of time	TQC
Self-confidence	Item measuring child's self-confidence	TQC
Peer relations	Scale score of 4 items measuring peer-acceptance and ability to get along with peers	TQC
Well-being	Scale score of 4 items measuring wellbeing of the child at school	TQC
Anxious	Scale score of 4 items measuring anxious-fearful behavior	TQC
Prosocial	Scale score of 7 items measuring the extent to which the child is prosocial with peers	TQC
TR language	Teacher-rated language ability	TQC
TR math	Teacher-rated math ability	TQC
Prognosis SE	Teacher's prognosis of success in secondary education	TQC
% High risk students	School-level variable: Percentage of high risk students at school in Year 0. Since children can change schools, this variable was used as a time-varying covariate.	OR

K = kindergarten; PQ = parent questionnaire; TQC = teacher questionnaire about the child; AT: achievement test; OR: official records; SES=socio-economic status; TR = teacher-rated

Evaluation of the weights distribution and balance properties. The means of the stabilized weights were close to one. More specifically, in Year 1, 2 and 3, the average stabilized weight was 1.00, 0.94 and 0.90 respectively. The stabilized weights (after truncation) ranged between 0.28-4.36 in Year 1, 0.19-18.59 in Year 2, and 0.19-18.93 in Year 3[iii].

Balance diagnostics consisted of the *SMDs* between the retained and promoted group, at each time sequence. The results are shown in Table 6. Figures 4, 5, and 6 present dot plots of *SMDs* in Year 1, 2 and 3 respectively, sorted according to the unweighted *SMD*. The weights were estimated ten times, for each of the ten imputed datasets. The diagnostics presented are based on the pooled estimates.

From the figures it is clear that the weighting substantially improved balance. In Year 1, all *SMDs* were below the threshold of 0.25. It should be noted that restricting the research sample to children who had at least 5% probability to be retained in kindergarten already reduced the unweighted covariate imbalance in Year 1 to a large extent. In Year 2, six of the 45

covariates had weighted *SMDs* larger than 0.25. In Year 3, 12 of the 61 variables had weighted *SMDs* larger than 0.25. The imbalance in the pre-treatment measurement of math achievement requires scrutiny when interpreting the results, since this is the strongest potential confounder of retention status and later math achievement. Prior to being retained, second-grade repeaters scored lower compared to promoted children.

Table 6

Standardized mean differences before and after weighting.

Treatment regime	**000000 vs. 111111**		**000000 vs. 011111**		**000000 vs. 001111**	
	Year 1		**Year 2**		**Year 3**	
	Unweighted	**Weighted**	**Unweighte**	**Weighte**	**Unweighte**	**Weighted**
Time-fixed						
Month of	0.20	0.00	**-0.29**	-0.07	-0.18	-0.07
Boys	-0.05	-0.04	0.01	-0.04	0.00	-0.01
Girls	0.06	0.04	-0.01	0.04	0.00	0.01
SES	-0.08	0.00	**-0.27**	-0.17	**-0.73**	-0.13
Home	0.09	-0.01	0.03	-0.05	-0.12	-0.13
Home	-0.02	-0.01	0.00	0.00	0.19	-0.03
Home	-0.06	0.01	-0.02	0.03	-0.08	0.11
Etnicity1	0.04	0.00	0.06	-0.05	**0.28**	0.07
Etnicity2	0.05	0.01	-0.06	-0.01	**-0.27**	-0.19
Etnicity3	0.03	0.00	-0.02	0.00	-0.21	-0.18
Etnicity4	-0.07	-0.01	0.00	0.04	-0.08	0.04
High risk0	-0.08	-0.05	-0.18	-0.09	**-0.39**	-0.09
High risk1	0.08	0.05	0.18	0.08	**0.41**	0.09
Time-varying covariates						
Math Y1	**-0.61**	-0.05	**-0.40**	-0.14	**-0.30**	0.07
TR language	**-0.37**	-0.01	**-0.27**	-0.17	**-0.31**	-0.25
TR math Y1	**-0.77**	0.06	-0.22	0.06	**-0.39**	**-0.37**
Independent	**-0.38**	0.04	**-0.41**	-0.19	-0.02	0.21
Cooperative	-0.09	0.01	-0.11	0.00	0.08	0.24
Hyperactive	0.14	-0.02	**0.29**	0.12	0.00	0.11
Asocial Y1	0.14	0.02	0.16	**0.29**	0.02	**-0.29**
Aggression	0.01	-0.10	0.01	-0.01	0.00	-0.02
Attitude Y1	-0.24	-0.04	**-0.37**	-0.24	-0.05	**0.26**
Self	-0.23	-0.09	-0.20	-0.05	-0.06	**0.28**
Peer relations	-0.20	-0.08	-0.18	-0.15	-0.25	-0.18
Welbeing Y1	-0.22	-0.02	-0.06	-0.06	-0.04	**0.32**
Anxious Y1	0.10	0.05	0.08	0.00	-0.11	-0.18
Prosocial Y1	-0.16	-0.01	-0.12	-0.08	0.13	**0.39**

Prognosis SE	**-0.32**	0.02	**-0.43**	-0.20	**-0.33**	-0.14
% High risk	0.10	0.03	-0.01	0.05	0.18	-0.08
Math Y2			**-0.76**	-0.21	**-0.39**	-0.16
TR language			**-1.32**	**-0.30**	-0.03	0.20
TR math Y2			**-1.66**	**-0.50**	**-0.45**	-0.25
Independent			**-1.15**	**-0.46**	**-0.66**	**-0.33**
Cooperative			**-0.37**	-0.06	0.06	0.01
Hyperactive			**0.56**	0.18	0.14	0.17
Asocial Y2			0.17	0.03	0.02	0.18
Aggression			0.19	0.05	-0.08	-0.23
Attitude Y2			**-0.84**	**-0.32**	**-0.31**	**-0.31**
Self			**-0.47**	-0.08	**-0.59**	-0.23
Peer relations			**-0.32**	-0.11	0.03	0.11
Welbeing Y2			**-0.52**	-0.12	0.12	0.04
Anxious Y2			-0.14	-0.01	0.10	-0.09
Prosocial Y2			-0.18	-0.01	0.02	0.01
Prognosis SE			**-1.21**	**-0.44**	**-0.49**	0.04
% High risk			-0.04	0.03	0.16	-0.11
Math Y3					**-0.67**	**-0.33**
TR language					**-0.74**	-0.16
TR math Y3					**-1.55**	**-0.58**
Independent					**-1.11**	-0.17
Cooperative					0.08	0.23
Hyperactive					**0.37**	-0.03
Asocial Y3					**0.34**	-0.16
Aggression					-0.16	**-0.38**
Attitude Y3					**-0.29**	-0.09
Self					**-0.67**	-0.06
Peer relations					**-0.32**	0.16
Welbeing Y3					**-0.32**	0.08
Anxious Y3					**0.29**	-0.03
Prosocial Y3					0.04	0.25
Prognosis SE					**-1.26**	**-0.67**
% High risk					0.17	-0.14

For continuously promoted children, the treatment trajectory was coded as (0,0,0,0,0,0). For kindergarten, first-grade and second-grade repeaters, the treatment trajectories were coded (1,1,1,1,1,1), (0,1,1,1,1,1) and (0,0,1,1,1,1) respectively (see Table 4): 0 means no delay, whereas 1 means delay. Y1 = Year 1, Y2 = Year 2, Y3 = Year 3; SES = socio-economic status; TR = teacher-rated; bold = absolute standardize mean difference (*SMD*) > 0.25; *SMD* with negative sign indicates lower score for retainees

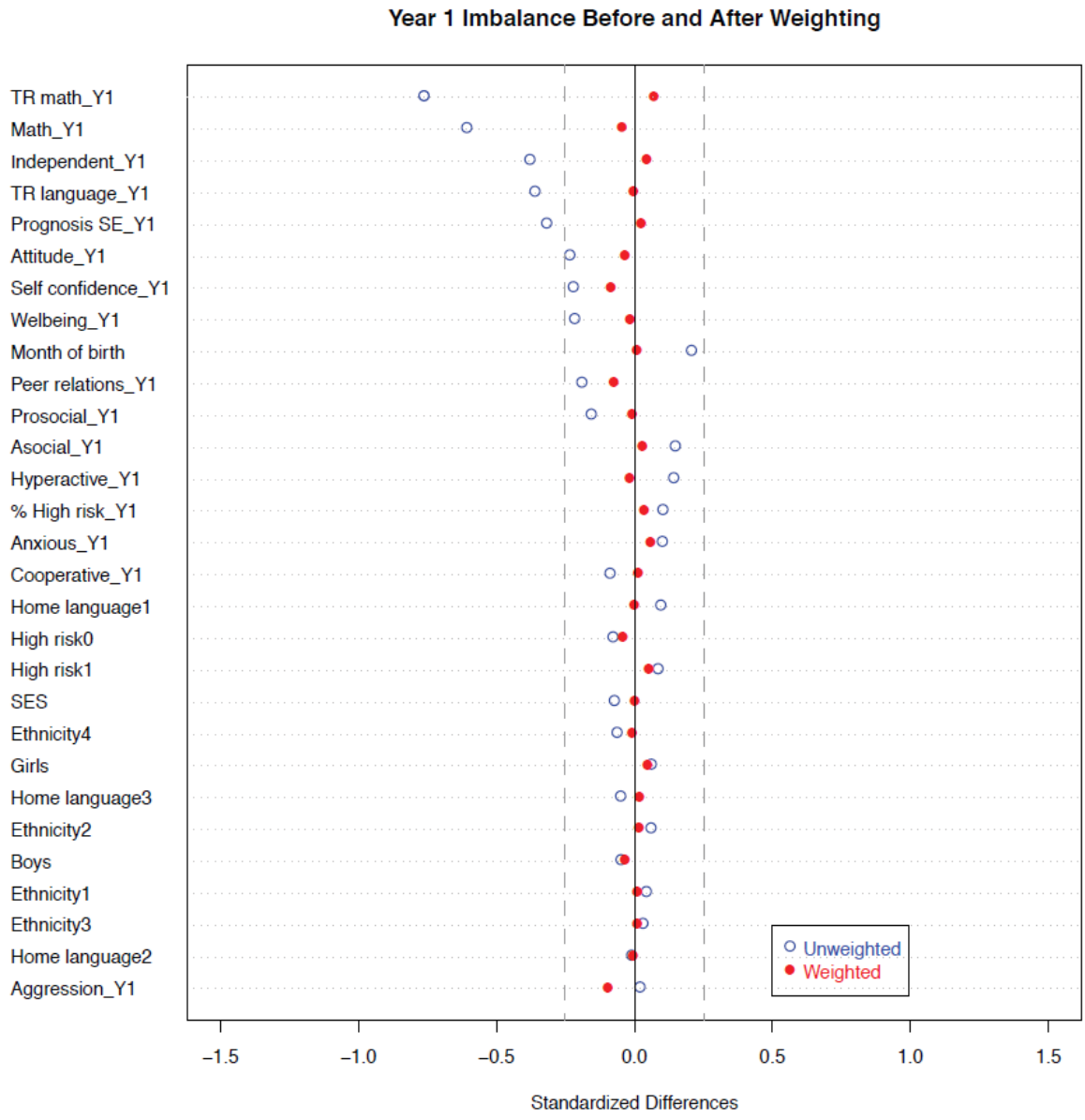


Figure 4. Standardized mean differences (*SMDs*) before and after weighting in Year 1, sorted according to the unweighted *SMD*. The *SMDs* indicate the difference in covariate means, divided by the pooled standard deviation (Rubin, 2001). Y1=Year-1 covariate. TR=teacher-rated. SES=socio-economic status. SE=secondary education.

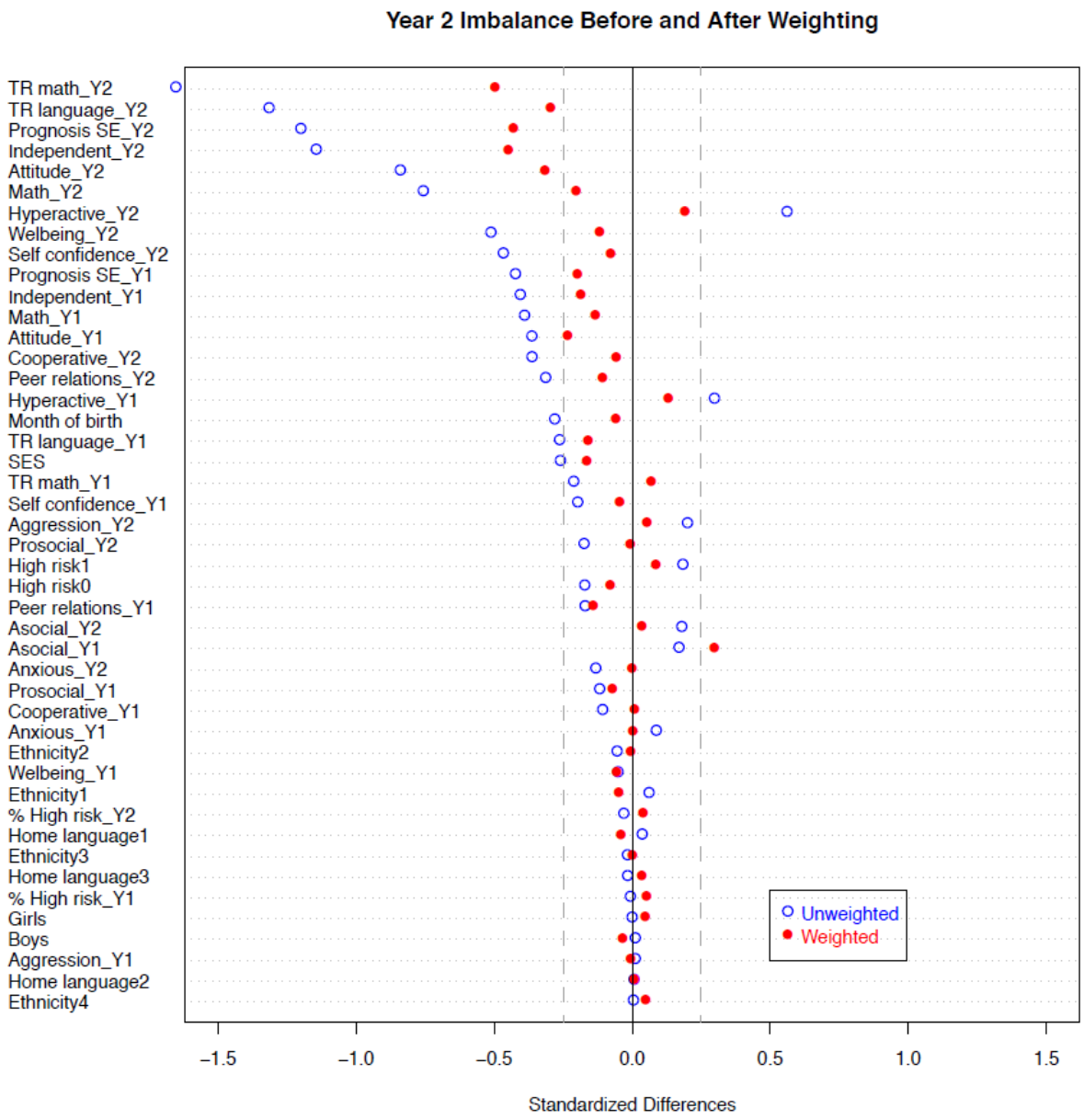


Figure 5. Standardized mean differences (*SMDs*) before and after weighting in Year 2, sorted according to the unweighted *SMD*. The *SMDs* indicate the difference in covariate means, divided by the pooled standard deviation (Rubin, 2001). Y1=Year-1 covariate, Y2=Year-2 covariate. TR=teacher-rated. SES=socio-economic status. SE=secondary education.

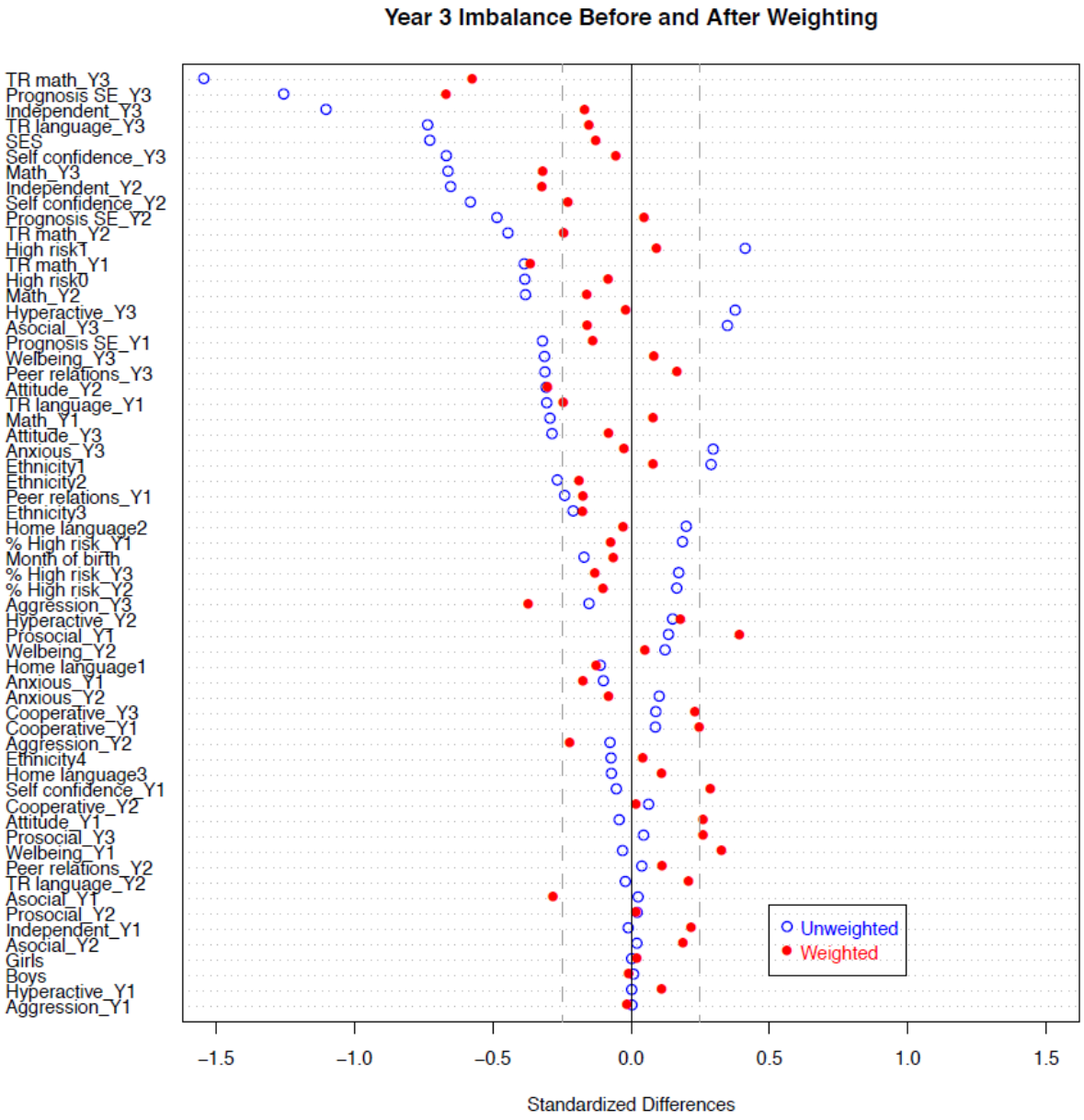


Figure 6. Standardize mean differences (*SMDs*) before and after weighting in Year 3, sorted according to the unweighted *SMD*. The *SMDs* indicate the difference in covariate means, divided by the pooled standard deviation (Rubin, 2001). Y1=Year-1 covariate, Y2=Year-2 covariate, Y3=Year-3 covariate. TR=teacher-rated. SES=socio-economic status. SE=secondary education.

For some covariates, the results indicate an increased *SMD* after weighting. For example, after weighting in Year 3, Year-1 prosocial behavior became imbalanced. Such an increase should be interpreted by the researcher considering its relative influence as a potential confounder (Harder et al., 2010). We were minimally concerned about this variable (prosocial

behavior) since it involves a distal Year-1 covariate and the more proximal measures of Year-2 and Year-3 prosocial behavior were balanced. Since the two most recent measurements of prosocial behavior were balanced, we assumed that later differences in mathematics could not be ascribed to differences in prosocial behavior three years earlier.

Missing data. Common in longitudinal research is the problem of missing data. In the current study, two types of missingness occurred. First, some children had temporarily missing outcome information at one or more measurement occasions (between 10% and 16% each year). This form of missingness was mainly due to absenteeism at the time of measurement. We assumed these missing data to be missing completely at random (MCAR). No further steps were taken for this type of missingness. Second, several children dropped out of the study (attrition) from one measurement occasion onwards (from 11% in Year 2 up to 51% in Year 6). These children were censored in the data analysis. Attrition occurred when children were retained after Year 3, retained for the second time, changed schools, or were transferred to special education. A comparison of this group with children who did not drop out indicated that these children had less favorable profiles; they scored lower on achievement tests and had a lower SES. Thus, we concluded that it was necessary to adjust these data for possible selection bias, which we did under the assumption that the missing data are MAR. To adjust for censoring, we selected the uncensored records and adjusted for possible selection bias by reweighting the (remaining) sample by the inverse of the probability of censoring, given the covariate history, treatment history, and other relevant potential confounders. The use of inverse probability weighting to address attrition under the MAR assumption has been used in previous research (Cole & Hernán, 2008; Hernán, Brumback & Robins, 2002; Robins et al., 2000). From Year 1 through Year 6, indicators for censoring were regressed on baseline covariates and pre-censoring treatment, outcome history and covariate history (amongst children who were previously uncensored). In addition to the covariates listed in Table 5, two dummies were

included as time-varying covariates, one indicating whether a child was delayed for at least one year and the second indicating that a child was in special education. The weights were stabilized (i.e., the weights were multiplied by the probability of being censored, given the censoring history). It is important to note that we could only adjust for MAR missingness to the extent that our measured time-varying covariates, e.g., history of retention and achievement scores, capture all predictors of censoring that are also associated with math achievement. If the previously stated conditional exchangeability assumption concerning retention is satisfied, then our adjustment for censoring due to retention after Year 3 – which is the main cause of censoring – is justified. The validity of our adjustment for nonrandom missingness is thus largely dependent upon the plausibility of the exchangeability assumption on which our entire analysis relies.

Estimating the parameters of the MSM using inverse-probability-of-treatment weighting. We fitted a multivariate response model, in which the dependent variables were the seven measurements of mathematics achievement. This model is different from that in most applications using MSMs, in which the outcome is only measured at the final time point. Our approach amounts to a weighted regression of the repeated measures of the outcome on the history of treatment, time and baseline covariates. The model was fitted using GEE with an independence working correlation matrix. This was done by means of the 'geepack'-package in R (Halekoh, Hojsgaard, & Yan, 2006). From Year 1 through Year 6, the available observations were weighed by a product of their treatment weight and their censoring weight up to the considered Year. Although the individual treatment weights remained constant from Year 3 onwards, their product with the individual censoring weights was time-varying up to Year 6. Time was treated as a categorical variable. Interactions between time and treatment regime were included as predictors in the model. The model for the outcome under the four treatment regimes is illustrated below:

(6)

No retention: $E(Y_{t_{000}}) = \beta_t$

Kindergarten retention: $E(Y_{t_{111}}) = \beta_t + \alpha_t$

First-grade retention: $E(Y_{t_{011}}) = \beta_t + \gamma_t$

Second-grade retention: $E(Y_{t_{001}}) = \beta_t + \delta_t$

where β_t indicates the average math score in Year *t* under the no retention treatment regime. Parameters α_t, γ_t, and δ_t refer to the average change in math score in Year *t* under the treatment regimes kindergarten retention, first-grade retention, and second-grade retention, respectively, relative to the no retention treatment regime. Contrasts were tested to examine at each time point the difference between the (up to) four treatment regimes. The outcome analysis was repeated ten times, once for each imputed dataset. The contrasts and outcome estimates for each of the ten datasets were combined according to Rubin's rules (Rubin, 1987).

Table 7

Contrast estimates of the comparisons in mathematics achievement.

Time	Contrast	Estimate		*SE*	*z*	ES
Year 1	noR-KR	11.91	***	1.44	8.26	1.33
Year 2	noR-KR	11.30	***	1.51	7.48	1.20
	noR-G1R	7.83	***	1.33	5.89	0.83
	KR-G1R	-3.47	**	1.22	-2.85	-0.37
Year 3	noR-KR	6.80	***	1.48	4.60	0.77
	noR-G1R	7.11	***	1.24	5.72	0.80
	noR-G2R	2.21		2.32	0.95	0.25
	KR-G1R	0.31		1.47	0.21	0.04
	KR-G2R	-4.59		2.48	-1.85	-0.52
	G1R-G2R	-4.90	*	2.36	-2.07	-0.55
Year 4	noR-KR	7.96	***	2.13	3.74	0.84
	noR-G1R	8.99	***	1.63	5.51	0.95
	noR-G2R	4.97	**	1.88	2.65	0.53
	KR-G1R	1.03		2.05	0.50	0.11
	KR-G2R	-2.99		2.35	-1.27	-0.32
	G1R-G2R	-4.02	*	1.94	-2.07	-0.43

Year 5	noR-KR	5.47	**	2.09	2.62	0.60
	noR-G1R	7.89	***	1.61	4.91	0.87
	noR-G2R	6.26	***	1.72	3.65	0.69
	KR-G1R	2.42		2.16	1.12	0.27
	KR-G2R	0.79		2.26	0.35	0.09
	G1R-G2R	-1.64		1.82	-0.90	-0.18
Year 6	noR-KR	4.95	*	2.47	2.00	0.54
	noR-G1R	8.91	***	1.81	4.93	0.97
	noR-G2R	6.02	*	2.79	2.16	0.66
	KR-G1R	3.96		2.58	1.54	0.43
	KR-G2R	1.07		3.37	0.32	0.12
	G1R-G2R	-2.89		2.81	-1.03	-0.32

Notes. noR = no retention; KR = kindergarten retention; G1R = first grade retention ; G2R = second grade retention. ES represents the effect size measure Cohen *d*, estimated in terms of the respective standard deviations of the outcome that year. Cohen (1988) labeled *d* values of 0.20, 0.50 and 0.80 as small, medium and large effect sizes respectively. *** $p<0.001$; ** $p<0.01$; * $p<0.05$

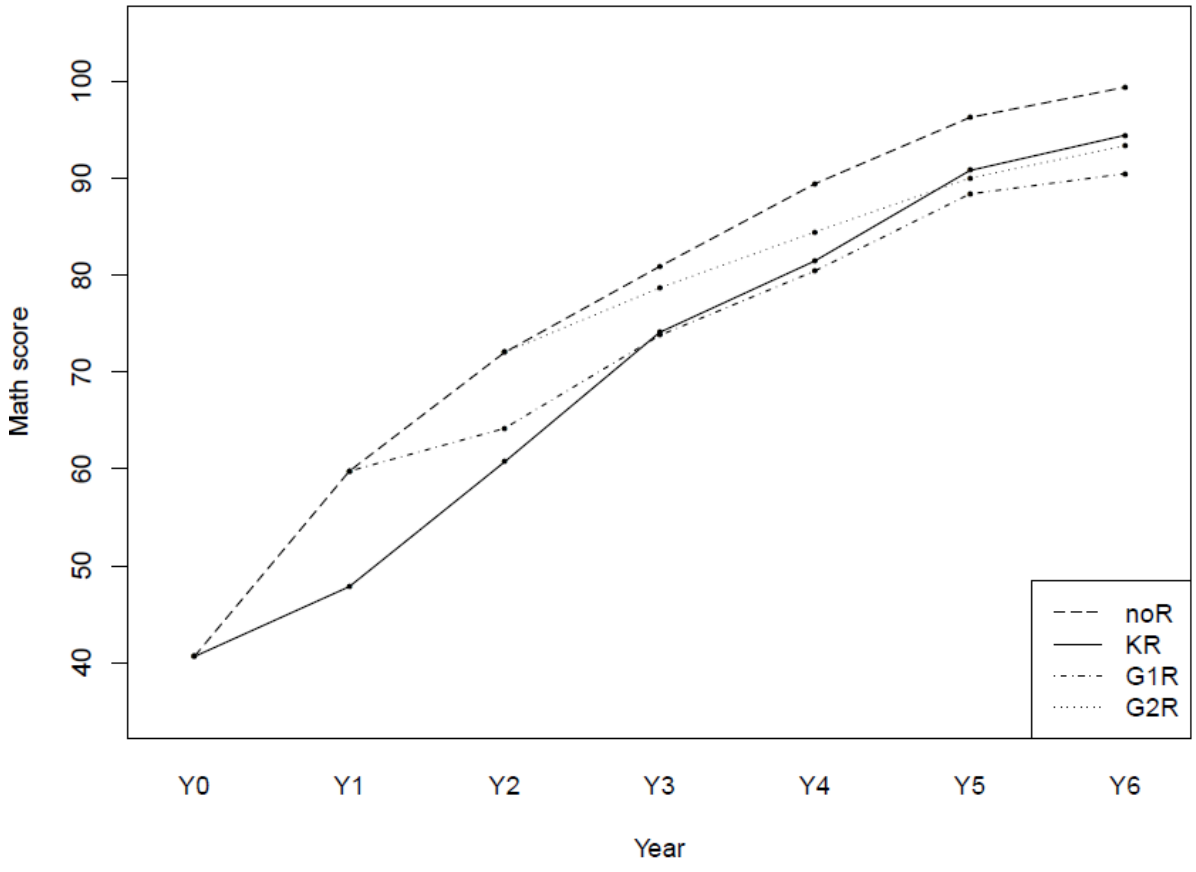


Figure 7. Same-age comparison of repeated measures of mathematics achievement for four treatment regimes. noR=no retention. KR=kindergarten retention. G1R=Grade 1 retention. G2R=Grade 2 retention.

The mathematics growth in each of the four treatment regimes is depicted in Figure 7. The contrast estimates are shown in Table 7. For example, a value in the row "KR - G1R" represents the estimated average score if all at-risk children were to repeat kindergarten (KR) minus the estimated average score if all at-risk children were to repeat first grade (G1R). In Year 1, it appears that significantly higher scores would be obtained if all children were promoted to first grade rather than being retained in kindergarten (estimate = 11.91, *SE* = 1.44, *d* = 1.33). In Year 2, three conditions were compared: promotion, kindergarten retention and first-grade retention. The contrasts indicated that by the end of Year 2, higher scores would be obtained if all children were promoted rather than being retained in kindergarten (estimate = 11.30, *SE* = 1.51, *d* = 1.20) or first grade (estimate = 7.83, *SE* = 1.33, *d* = 0.83). The advantage of being continuously promoted compared to kindergarten and first-grade retention was observed until the end of primary school. Further, in Year 2, lower scores would be obtained if all children were retained in kindergarten rather than in first grade (estimate = -3.47, *SE* = 1.22, *d* = -0.37). In Year 3, there was no longer a significant difference between kindergarten retention and first-grade retention. Nevertheless, in the same year, significantly lower scores were observed after first-grade retention in comparison with second-grade retention (estimate = -4.90, *SE* = 2.36, *d* = -0.55). This advantage of second-grade retention compared to first-grade retention disappeared from Year 5 onwards. By the end of Year 6, the final measurement, it appeared that significantly higher scores would be obtained if all children were continuously promoted rather than being retained in kindergarten (estimate = 4.95, *SE* = 2.47, *d* = 0.54), first grade (estimate = 8.91, *SE* = 1.81, *d* = 0.97) or second grade (estimate = 6.02, *SE* = 2.79, *d* = 0.66). Any differences in math achievement among the three retention conditions failed to reach statistical significance ($\alpha = .05$).

In sum, the results indicate that children at risk for grade retention would score higher on mathematics throughout their entire primary school career if they were promoted each year.

The scores for kindergarten and first-grade repeaters, compared to continuously promoted children, were significantly lower from the end of their retention year onwards. Children who repeated second grade fell less behind during their retention year than children who repeated kindergarten or first grade. The advantage for promoted children in contrast with second-grade repeaters was significant from Year 4 onwards. By the end of primary school, kindergarten and second-grade repeaters appeared to score slightly higher in mathematics achievement than first-grade repeaters, but the effect sizes were small and the results were not statistically significant.

Discussion

In the social and behavioral sciences, one of the main objectives is to disentangle the causal connection between an event or intervention and the outcome. In the absence of randomization, researchers must adjust for variables that confound the relationship between the treatment and the outcome. When the treatment is initiated at different times for different subjects, confounding may also become time-varying. In the present study, we explained the use of marginal structural models, fitted by means of inverse probability weighting, to deal with time-varying treatments in the presence of time-varying confounding. Unlike conventional regression techniques, MSMs allow valid adjustment for time-varying variables that potentially confound the relationship between the (time-varying) treatment and the outcome. By means of reweighting the research sample at each treatment occasion, a pseudo-population is created in which confounding is removed (provided all relevant confounding factors were available and correctly modelled in the propensity score models). Although this approach has been widely used in epidemiologic research, applications in social and behavioral sciences to date have been limited. Given the increase of accessible software packages with which to apply MSMs, we believe that this approach will be of great value in future social and behavioral research. In this study, we illustrated the use of MSMs with an empirical example of the effects of grade retention on math achievement scores during primary school. Using a population of children

who had a probability of at least 5% of being retained in kindergarten, we compared the marginal distribution of mathematics achievement in four treatment regimes: no retention, kindergarten retention, first-grade retention and second-grade retention. In other words, children could be retained at three different points in time. To adjust for time-varying confounding, time-specific treatment weights were estimated at each of the three treatment times. The sample was reweighted each year using the product of time-specific weights up to each time t .

When the treatment can be received at multiple time points, the number of potential treatment patterns strongly increases. An increase in the number of treatment patterns and covariates may hamper achieving a good balance across all those combinations. In our empirical example, the reweighted standardized mean differences in the covariate history between retained and promoted children gave an indication of the amount of remaining imbalance at each time point. The weighting led to a good covariate balance in the first year. In the second and third year, the weighting failed to balance some covariates to an acceptable level. Although the balance improved dramatically relative to an unadjusted analysis, we thus cannot exclude some residual confounding bias.

The results indicate that at-risk children benefit most from being continuously promoted. Compared to same-age children who were one grade higher and who were therefore confronted with new subject matter, children who were delayed for one year scored lower on a measure of mathematics achievement. It seems that at-risk children may benefit more from receiving intellectual challenges that are offered in a higher grade. This is in line with previous research in which academic outcomes of early grade repeaters are compared with those of same-age peers who are one grade higher (Goos et al., 2013; Hong & Yu, 2007; Hong & Raudenbush, 2005; Hong & Raudenbush, 2006; Vandecandelaere et al., 2015a; Vandecandelaere et al., 2015b; Wu et al., 2008a; Wu et al., 2008b). Some authors argue that a same-age comparison is

unfair since we cannot expect grade repeaters know the material to which their same-age peers who are one grade higher have been exposed (Lorence, 2006; Moser et al., 2012). To control for this difference in instruction, the grade repeaters and promoted children can be compared when they are in the same grade. Previous research on early grade retention using a same-grade comparison has demonstrated that grade repeaters, compared to their younger grade-mates, score better in math during the retention year. However, in the long term, this initial advantage disappears or even reverses (Dong, 2010; Goos et al., 2013; Moser et al., 2012; Wu et al., 2008a). In this study, we did not compare the relative position of the retained students to that of their younger classmates. A same-grade comparison would require shifting back the promoted children one year in such a manner that all of the children are being compared in the same grade. This comparison would involve a different operationalization of the treatment patterns in estimating the treatment and censoring weights. To illustrate this, in Figure 8, we graphically shifted back the growth of the promoted children. This figure gives a rough indication of what the results of a same-grade comparison would look like. It is an approximate depiction because the corresponding shifted covariates were not used. The figure seems to indicate that, in the long term (in Grade 5), grade repeaters scored lower compared to their younger grade-mates. This lower mathematics score is of course at the cost of an extra year of primary school for the retained children. Future research is needed to investigate this result in more detail.

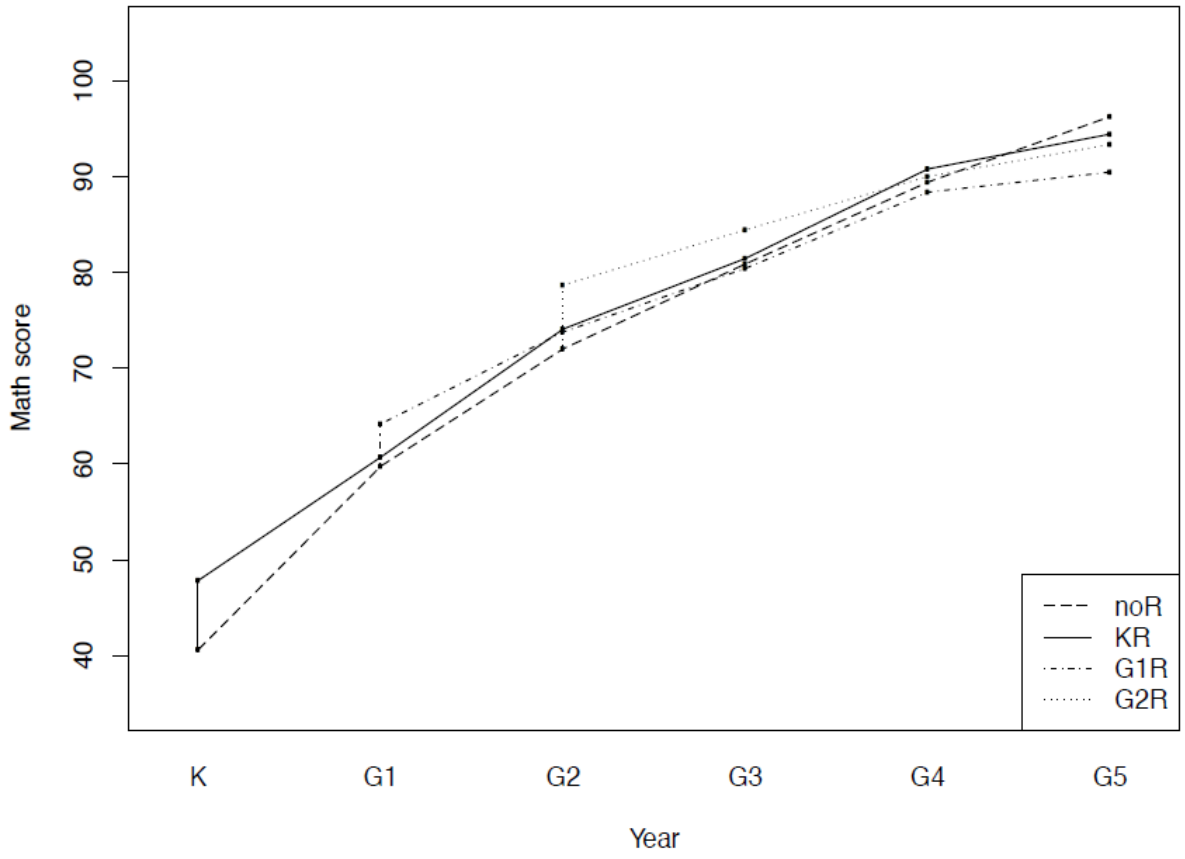


Figure 8. Same-grade comparison of repeated measures of mathematics achievement for four treatment regimes. The achievement scores of the promoted children in Figure 7 were graphically shifted back. It is an approximate depiction since the corresponding shifted covariates were not used. noR=no retention. KR=kindergarten retention. G1R=Grade 1 retention. G2R= Grade 2 retention.

Our results differ from a preceding study using the same data, in which we found kindergarten retention to be significantly less harmful compared to first-grade retention for mathematics achievement (Vandecandelaere et al., 2015a). In the current study, the difference was not statistically significant. These conflicting results might be explained by the different approaches of the two analyses. In the previous study, we matched treatment groups only with respect to baseline covariates, measured before anyone was ever retained (in Year 0). In the present analysis, we also balanced first-grade repeaters and promoted children with respect to

covariates measured right before they were retained (in Year 1, and thus after they could have been retained in kindergarten). In other words, we now also adjusted for time-varying confounding. We speculate that this adjustment for more recent differences between first-grade repeaters and promoted children improved the balancing properties and led to more accurate estimates. On the other hand, compared to a matching approach, weighted analyses introduce more variation and come with conservative standard errors, which might lead to increased uncertainty about the estimates.

Our results rely on the exchangeability assumption, meaning that we assume that all prognostic factors of the outcome that are also associated with the treatment condition and censoring were adjusted for in the respective propensity scores and censoring models. Our findings are unbiased only to the extent that the treatment and censoring models were correctly specified and included all these prognostic factors. An important challenge for future research is to explore ways to examine the sensitivity of parameters estimated in a MSM to possible violations of the exchangeability assumption that result from non-availability of relevant prognostic factors.

Due to the remaining imbalance of some covariates, it is possible that the negative effect of retention in second grade is overestimated. In particular, the weighting substantially reduced the standardized mean difference in pre-treatment math, yet, second-grade repeaters, compared to promoted children, still scored lower on mathematics before they were retained. Caution must be exercised in interpreting any outcome differences that might be due to covariates that remain imbalanced after weighting. This finding underscores the vital importance of evaluating covariate balance, something that has regularly been ignored in previous studies using MSMs. Indeed, one drawback of using observational data is the untestable, and often unrealistic assumption of exchangeability. Although we cannot completely rule out the existence of bias due to the possible violation of this assumption, it is important to evaluate and report the extent

to which, after weighting, balance - and thus exchangeability - has been attained in terms of the measured covariates. This way, estimated causal effects can be interpreted more correctly, and the research community can weigh the validity of the findings.

Recent work of Imai and Ratkovic (2014) offers promising perspectives with respect to improved covariate balance. The authors demonstrate an alternative approach to the estimation of the weights. By means of the generalized method of moments, their estimation procedure produces covariate balancing propensity scores (CBPS) which incorporate the key covariate balancing property of propensity scores (i.e., the mean independence between the treatment and covariates after weighting). Software to apply this approach in a longitudinal setting is quickly developing and might enhance the use of CBPS with MSMs in the future.

Finally, it should be noted that because GEE analyses do not specify the full conditional likelihood, they assume that possibly incomplete outcomes are MCAR (Ghisletta & Spini, 2004). The bias produced under MAR mechanisms may often not be large in magnitude, however (Fitzmaurice, Laird, & Rotnitzky, 1993). Furthermore, valid GEE analyses under MAR can be obtained via the use of censoring weights, as illustrated in this article. The assumption underlying censoring weights is comparable to the MAR assumption.

This study demonstrated the usefulness of MSMs in investigating questions in the context of time-varying grade retention. We believe that this approach is promising for social and behavioral researchers who wish to investigate interventions that may occur at multiple points in time. An important advantage of MSMs is its straightforward implementation as it is a weighted version of standard approaches. In this article, we also demonstrated the complications that may arise, and the need (a) to carefully explore the bias-variance trade-off and (b) to evaluate the balancing properties in order to allow valid interpretations of the estimated parameters in the MSM. The challenge is to select potential confounders that reduce

substantial bias, but which do not lead to high variance, and, at the same time, that satisfy the exchangeability assumption (Robins & Hernán, 2008).

Reference List

Allen, C. S., Chen, Q., Willson, V. L., & Hughes, J. N. (2009). Quality of research design moderates effects of grade retention on achievement: A meta-analytic, multilevel analysis. *Educational Evaluation and Policy Analysis 31*(4), 480-499. doi:10.3102/0162373709352239

Austin, P. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399-424. doi:10.1080/00273171.2011.568786

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research, 20*(1), 40-49. doi:10.1002/mpr.329

Barber, J. S., Murphy, S. A., & Verbitsky, N. (2004). Adjusting for time-varying confounding in survival analysis. *Sociological Methodology, 34*(1), 163-192. doi:10.1111/j.0081-1750.2004.00151.x

Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45*(3), 1-67. doi: 10.18637/jss.v045.i03

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

Cham, H., & West, S. G. (in press). Propensity score methods with missing data. *Psychological Methods*.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.).* New Jersey: Lawrence Erlbaum.

Cole, S. R., & Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology, 168*(6), 656-664. doi:0.1093/aje/kwn164 t

Dong, Y. Y. (2010). Kept back to get ahead? Kindergarten retention and academic performance. *European Economic Review, 54*(2). doi:10.1016/j.euroecorev

Eurydice. (2014). *Eurypedia: The European encyclopedia on national education systems.* Retrieved from https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/

Eurydice network. (2011). *Grade Retention during Compulsory Education in Europe: Regulations and Statistics*. Retrieved from: http://eacea.ec.europa.eu/education/eurydice/documents/thematic_reports/126EN.pdf

Fitzmaurice, G. M., Laird, N. M., & Rotnitzky, A. G. (1993). Regression models for discrete longitudinal responses. *Statistical Sciences, 8*, 284-309. doi: 10.1214/ss/1177010903

Ghisletta, P., & Spini, D. (2004). An introduction to generalized estimating equations and an application to assess selectivity effects in a longitudinal study on very old individuals. *Journal of Educational and Behavioral Statistics, 29*(4), 421-437. doi:10.3102/10769986029004421

Goos, M., Van Damme, J., Onghena, P., Petry, K., & de Bilde, J. (2013). First-grade retention in the Flemish educational context: Effects on children's academic growth, psychosocial growth, and school career throughout primary education. *Journal of School Psychology, 51*(3), 323-347. doi:10.1016/j.jsp.2013.03.002

Halekoh, U., Hojsgaard, S., & Yan, J. (2006). The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software, 15*(2), 1-11. doi: 10.18637/jss.v015.i02

Harder, V., Stuart, E., & Anthony, J. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15*(3), 234-249. doi:10.1037/a0019623

Hernán, M. A. (2004). A definition of causal effect for epidemiological research. *Journal of Epidemiology and Community Health, 58*(4), 265-271. doi:10.1136/jech.2002.006361

Hernán, M. A., Hernández-Diaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology, 15*(5), 615-625. doi: 10.1097/01.ede.0000135174.63482.43

Hernán, M. A., Brumback, B. A., & Robins, J. M. (2002). Estimating the causal effect of zidovudine on CD4 count with a marginal structrural model of repeated measures. *Statistics in Medicine, 21,* 1689-1709. doi:10.1002/sim.1144

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis,15*(3), 199-236. doi:10.1093/pan/mpl013

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945-960. doi:10.1080/01621459.1986.10478354

Holmes, C. T. (1989). Grade-level retention effects: A meta-analysis of research studies. In L.A. Shepard & M. L. Smith (Eds.), *Flunking grades: Research and policies on retention* (pp. 16-33). London, United Kingdom: The Falmer Press.

Hong, G. (2015). *Causality in a social world: Moderation, mediation, and spill-over.* West Sussex, UK: Wiley-Blackwell.

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis, 27*(3), 205-224. doi:10.3102/01623737027003205

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association, 101*(475), 901-910. doi:10.1198/016214506000000447

Hong, G., & Raudenbush, S. W. (2008). Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics, 33*(3), 333-362. doi: 10.3102/1076998607307355

Hong, G., & Yu, B. (2007). Early-grade retention and children's reading and math learning in elementary years. *Educational Evaluation and Policy Analysis, 29*(4), 239-261. doi:10.3102/0162373707309073

Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76*(1), 243-263. doi:10.1111/rssb.12027

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

Jacob, B. A., & Lefgren, L. (2009). The effect of grade retention on high school completion. *American Economic Journal: Applied Economics, 1*(3), 33-58. doi:10.1257/app.1.3.33

Jimerson, S. R. (2001). Meta-analysis of grade retention research: Implications for practice in the 21st century. *School Psychology Review, 30*(3), 420-437.

Joffe, M. M., Ten Have, T. R. T., Feldman, H. I., & Kimmel, S. E. (2004). Model selection, confounder control, and Marginal Structural Models: Review and new applications. *The American Statistician, 58*(4), 272-279. doi:10.1198/000313004X5824.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*(3), 337-396. doi:10.1002/sim.3782

Lorence, J. (2006). Retention and academic achievement research revisited from a United States perspective. *International Education Journal, 7*(5), 731-777. Retrieved from http://ehlt.flinders.edu.au/education/iej/articles/v7n5/Lorence/paper.pdf

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research (2nd Ed., Chapter 4).* Cambridge, UK: Camridge University Press.

Moser, S. E., West, S. G., & Hughes, J. N. (2012). Trajectories of math and reading achievement in low-achieving children in elementary school: Effects of early and later retention in grade. *Journal of Educational Psychology, 104*(3), 603-621. doi:10.1037/a0027571

Robins, J. M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In L. Sechrest, H. Freeman, & A. Mulley (Eds.), *Health Service Research Methodology: A Focus on AIDS* (pp. 113-159). Washington D.C.: National Center for Health Services Research.

Robins, J. M. (2000). Marginal Structural Models versus Structural Nested Models as tools for causal inference. In M.E. Halloran & D. Berry (Eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (116 ed., pp. 95-133). New York, NY: The IMA Volumes in Mathematics and its Applications Springer.

Robins, J. M., & Hernán, M. A. (2008). 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). New York, NY: Chapman and Hall/CRC Press.

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

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, 147*(5), 656-666. doi: 10.2307/2981697

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. doi:10.2307/2335942

Rubin, D. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, *Journal of Educational Psychology, 66*(5), 688–701. doi:10.1037/h0037350

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley. doi:10.1002/9780470316696

Rubin, D. (2001). Using propensity scores to help design observational studies: Application to the Tobacco Litigation. *Health Services & Outcomes Research Methodology, 2*(3-4), 169-188. doi:10.1023/A:1020363010465

Schafer, J. L., & Kang, J. D. Y. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods, 13*(4), 279-313. doi:10.1037/a0014268

van der Wal, W. M., & Geskus, R. B. (2011). ipw: An R package for Inverse Probability Weighting. *Journal of Statistical Software, 43*(13), 1-23. doi:10.18637/jss.v043.i13

Vandecandelaere, M., Schmitt, E., Vanlaar, G., De Fraine, B., & Van Damme, J. (2014). Effects of kindergarten retention for at-risk children's psychosocial development. *Educational Psychology.* Advance online publication. doi:10.1080/01443410.2014.950194

Vandecandelaere, M., Schmitt, E., Vanlaar, G., De Fraine, B., & Van Damme, J. (2015a). Effects of kindergarten retention for at-risk children's mathematics development. *Research Papers in Education, 30*(3), 305-326. doi:10.1080/02671522.2014.919523

Vandecandelaere, M., Vansteelandt, S., De Fraine, B., & Van Damme, J. (2015b). The effects of early grade retention: Effect modification by prior achievement and age. *Journal of School Psychology*. Advance online publication. doi:10.1016/j.jsp.2015.10.004

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. doi:10.1037/a0022610

Vansteelandt, S. (2007). On confounding, prediction and efficiency in the analysis of longitudinal and cross-sectional clustered data. *Scandinavian Journal of Statistics, 34*(3), 478-498. doi:10.1111/j.1467-9469.2006.00555.x

Vansteelandt, S. & Daniel, R. M. (2014). On regression adjustment for the propensity score. *Statistics in Medicine, 33(*23), 4053-4072. doi:10.1002/sim.6207

Vlaams ministerie van onderwjis en vorming (2014). *Voorpublicatie Statistisch jaarboek van het Vlaams onderwijs - schooljaar 2013-2014 [Preliminary publication yearbook of education statistics in Flanders 2013-2014]*. Retrieved from http://www.ond.vlaanderen.be/onderwijsstatistieken/2013-2014/statistischjaarboek2013-2014/publicatiestatistischjaarboek2013-2014.htm

Wu, W., West, S. G., & Hughes, J. N. (2008a). Effect of retention in first grade on children's achievement trajectories over 4 years: A piecewise growth analysis using propensity score matching. *Journal of Educational Psychology, 100*(4), 727-740. doi:10.1037/a0013098

Wu, W., West, S. G., & Hughes, J. N. (2008b). Short-term effects of grade retention on the growth rate of Woodcock-Johnson III broad math and reading scores. *Journal of School Psychology, 46*(1), 85-105. doi:10.1016/j.jsp.2007.01.003

Zeger, S., & Liang K. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics, 42*(1), 121-130.

i Of the 460 children who were retained in kindergarten, first grade or second grade, 43 children were retained a second time. These children were censored from the second retention year onwards.

[ii] For clarity, note that the most recent measurement of time-varying covariates that predict treatment in Year *t* in fact took place in the preceding school year (i.e., Year *t*-1). For example, the most recent score for wellbeing in the prediction model for retention in Year 1, was measured six months before the start of Year 1. In other words, although we refer to this covariate as Year 1-wellbeing, this variable was in fact measured at the end of Year 0.

[iii] Note that the means and ranges of the weights slightly changed in Year 4 through Year 6 because each year, the group of censored students expanded (as explained later in the article, we dealt with censored students by means of censoring weights). The means of the weights in Year 4, 5 and 6 were equal to 0.88, 0.87 and 0.87 respectively. The ranges in these years were equal to 0.19 – 18.92, 0.19 – 11.20, and 0.20 – 11.10.

Appendix. Assumptions underlying weighting

There are several key assumptions underlying the use of inverse probability weighting in a time-varying setting.

1. Exchangeability assumption

Exchangeability between treatment groups, or the assumption of no unmeasured confounding, requires that the treated, had they been untreated, would have experienced the same average outcome as did the untreated, and vice versa. In the same way, the exchangeability assumption between censored and uncensored subjects, requires that the censored subjects, had they been uncensored, would have experienced the same average outcome as did the uncensored subjects, and vice versa. Exchangeability holds when treatment assignment or censoring is independent of both potential outcomes, as in a randomized controlled trial (RCT). In a time-varying setting, this assumption implies that at each time *t*, there are no prognostic factors of the outcome that have different distributions in the treatment and the control groups, given the treatment history $\bar{Z}_{t\text{-}1}$, the baseline covariates *X*, and the covariate history $\bar{L}_t$. For censoring, the assumption implies that at each time *t*, there are no prognostic factors of the outcome that have different distributions in the censored and the uncensored groups, given the censoring history, baseline covariates, and the covariate history. This assumption is also called the sequential randomization assumption. The assumption would hold if at each time, treatment or censoring

were randomly assigned with randomization probabilities that are possibly depending on the treatment/censoring and confounder history (Robins & Hernán, 2008).

In the empirical example, we additionally assumed that once a student is delayed for one year, the student stays delayed. As a consequence, we did not need to assume that children who remain and do not remain delayed were exchangeable. Hence, our estimates did not require the assumption of exchangeability after grade retention.

The exchangeability assumption implies that we assume that the measured covariates are sufficient to adjust for both confounding and for selection bias due to loss to follow-up measurement. Unfortunately, as is the case in all observational studies, this assumption cannot be tested based on the data. Results will be unbiased only to the extent that the treatment and censoring models included all relevant confounders. An important challenge for future research is to explore ways to examine the sensitivity of estimated parameters in a MSM (Marginal Structural Model) to possible violations of the exchangeability assumption.

2. Consistency assumption

The consistency assumption requires that the counterfactual outcome of a subject, under the observed treatment or censoring history, is precisely the observed outcome (Cole & Hernán, 2008).

3. Positivity assumption

The positivity assumption refers to the condition that treatment or censoring is possible at every level of the covariates. This is also referred to as the experimental treatment assumption. When there is a covariate combination at which it is impossible to be treated or not treated, a

structural zero probability of receiving treatment will occur. One way to deal with nonpositivity is to restrict the sample to subjects who meet a minimum probability of being treated or not treated on the basis of baseline covariate information (Cole & Hernán, 2008). In the empirical example included in this article, we limited the analytic sample to children who had at least a 5% predicted probability of being retained in kindergarten.

4. No model misspecification

The fourth assumption requires that the series of propensity score models used to estimate the treatment and the censoring weights are correctly specified. A necessary condition for correct model specification is that the stabilized weights have a mean of one (Cole & Hernán, 2008).

Under these assumptions, Robins (1999) demonstrated that inverse probability weighting (IPW) can be used to consistently estimate the mean potential outcome, allowing researchers to compute the average outcome under any treatment pattern.

It should be noted that the assumptions of MSMs are less restrictive than those of standard methods. For example, MSMs do not require the absence of time-dependent confounding by variables affected by previous exposure (Hernán, Brumback, & Robins, 2002).