

UC Riverside

UC Riverside Previously Published Works

Title

Identifying Causal Estimands for Time-Varying Treatments Measured with Time-Varying (Age or Grade-Based) Instruments

Permalink

<https://escholarship.org/uc/item/2739m2sf>

Journal

Multivariate Behavioral Research, 51(6)

ISSN

0027-3171

Authors

Steiner, Peter M
Park, Soojin
Kim, Yongnam

Publication Date

2016

DOI

10.1080/00273171.2016.1205470

Peer reviewed



Published in final edited form as:

Multivariate Behav Res. 2016 ; 51(6): 865–8780. doi:10.1080/00273171.2016.1205470.

Identifying Causal Estimands for Time-Varying Treatments Measured with Time-Varying (Age or Grade-Based) Instruments

Peter M. Steiner^a, Soojin Park^b, and Yongnam Kim^a

^aUniversity of Wisconsin—Madison;

^bUniversity of California—Riverside

Abstract

This commentary discusses causal estimands of same-age and same-grade comparisons for assessing grade-retention effects on student ability and performance. Using potential outcomes notation, we show that same-age and same-grade comparisons refer to different retention–promotion contrasts and therefore assess different causal questions. We also comment on deleting versus censoring records of students who dropped out of the study or do not belong to the treatment regimes under investigation. Whereas deleting entire student records potentially induces collider bias, censoring circumvents bias if censoring is ignorable given the observed pretreatment covariates.

Keywords

Causal inference; collider bias; same-age comparison; same-grade comparison; time-varying instruments; time-varying treatments

We appreciate the opportunity to comment on the excellent introduction on time-varying treatments by Vandecandelaere, Vansteelandt, De Fraine, and Van Damme. We congratulate the authors for writing a very accessible introduction to marginal structural models for estimating the effects of static treatment regimes. In particular, how they explain issues with respect to time-varying treatments and confounding and the assumptions required for causally identifying the treatment effects is very well done.¹ We are convinced that future

CONTACT Peter M. Steiner psteiner@wisc.edu Department of Educational Psychology, University of Wisconsin—Madison, 1025 W Johnson Street, Madison, WI 53706.

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

¹A causal estimand (defined in terms of unobserved potential outcomes) is identified if it can be computed from observed data of a hypothetically infinite population. This is the case if the observed data meet all causal assumptions (e.g., conditional independence). This is in contrast to statistical identification that refers to the estimability of a parameter from a finite sample, regardless of whether the parameter actually warrants a causal interpretation or not. For example, the unadjusted mean difference in the retained and promoted students' achievement scores is statistically identified as long as we have at least one observation per group, but causal identification requires for example that the potential outcomes of the retention and promotion condition are (conditionally) independent. In this commentary we always refer to causal identification.

research on grade retention will be guided by these authors' application of marginal structural models for estimating the effects of early grade retention on math achievement.

Identifying retention effects is particularly challenging since the test instruments for measuring student achievement vary over time; that is, the administered achievement tests depend on the time-varying retention decision. Thus, as Vandecandelaere et al. correctly point out, the choice of an appropriate comparison strategy, either a same-age or same-grade comparison, is of crucial importance. We go a step further and formalize the two comparison strategies using potential outcomes notation in order to define and discuss different causal estimands and their underlying treatment contrasts. We argue that the same-age and same-grade comparisons differ in their retention–promotion contrasts and, thus, assess different causal questions. Another issue we address in this commentary refers to handling records of students who dropped out (missing data) or do not belong to the treatment regimes under investigation (i.e., multiple retentions or retentions after a certain grade). Instead of deleting the entire student records, Vandecandelaere et al. carefully censor corresponding records and address resulting biases by using censoring weights. Here, we highlight what would have happened if Vandecandelaere et al. had deleted rather than censored the data.

Same-age and same-grade comparisons

The average causal effect (ACE) of grade retention on an achievement score Y can be defined in terms of potential outcomes as $ACE = E(Y(1) - Y(0)) = E(Y(1)) - E(Y(0))$, that is, the difference in the expected potential retention and promotion outcomes that would be observed if all at-risk students were retained, $E(Y(1))$, instead of promoted, $E(Y(0))$. Let $Z \in \{0,1\}$ be the retention indicator, then $Y(Z = z)$ represents the potential retention ($Z = 1$) or potential promotion outcome ($Z = 0$). The expectation in the ACE definition is typically taken over the population of at-risk students. For simplicity of exposition, we consider a situation where students can get retained only once and only at the same single grade $g^* \in \{0, \dots, G\}$, that is, at a specific grade g^* between the kindergarten grade 0 and a final grade G . For example, the retention condition ($Z = 1$) may refer to retention in kindergarten ($g^* = 0$), whereas the promotion condition ($Z = 0$) implies no kindergarten or later retention.

Although the ACE definition seems natural, the ACE has a meaningful interpretation only if both potential outcomes are measured with the same instrument (i.e., achievement test) at the same time. However, since achievement tests vary by grade, retention effects are not directly assessable like medication effects on blood pressure where the blood pressure monitor remains invariant to patients' treatment and age. Thus, to explicitly highlight the grade dependence of potential outcomes, we use a grade index g . Since retained and promoted students take tests of the assessment grade g (for $g > g^*$) at different ages, we also introduce an age index k . Thus, the potential outcomes, $Y_k^g(0)$ and $Y_k^g(1)$, refer to a specific assessment age and assessment grade. Using this notation, we now can discuss same-age and same-grade ACEs with respect to outcomes measured t years after the retention decision. For a given $t > 0$, intervention grade g^* , and a kindergarten entrance age of 5 years, a student's assessment age refers to $k = 5 + g^* + t$. Consider retention at the kindergarten ($g^* = 0$) and suppose we are interested in the retention effect 3 years after the retention decision

($t = 3$), then the assessment age is $k = 5 + 0 + 3 = 8$ years. Under the promotion regime, students would be in the third grade ($g^* + t = 0 + 3 = 3$); under the retention regime, they would be in the second grade ($g^* + t - 1 = 0 + 3 - 1 = 2$). We use this example ($g^* = 0$, $k = 8$, and $t = 3$) to illustrate the differences between the same-age and same-grade comparison but provide a more general and formal notation in Table 1.

Same-age comparisons

If we are interested in assessing the effect of grade g^* retention on achievement scores measured at age k , we could define the retention effect as $ACE_k = E(Y_k^{g-1}(1) - Y_k^g(0))$, which is the expected difference in the potential retention outcome assessed at age k with the grade $g-1$ test and the potential promotion outcome assessed at the same age k but with the grade g test. Thus, 3 years after kindergarten, the retention effect $ACE_{k=8} = E(Y_8^2(1) - Y_8^3(0))$ is the expected difference in the 8-year-old students' potential retention and promotion outcomes assessed at the second and third grade, respectively.

The same-age $ACE_{k=8}$ relies on a treatment contrast where retention requires repeating the kindergarten curriculum and taking the kindergarten to second grade with peers that are 1 year younger than in the promotion condition. Promotion requires taking the first to the third grade with same-age peers. Thus, retention differs from promotion in (a) repeating kindergarten with 1-year-younger peers, (b) taking the first and second grade 1 year later with 1-year-younger peers, and (c) the absence of exposure to the third grade (first column of Table 1). This treatment contrast is well defined and of primary interest in evaluations of retention effects on students. If there were no ethical issues, students could actually be randomly assigned to the corresponding retention and promotion conditions. However, the $ACE_{k=8}$ definition reveals that the retention effect is completely confounded with differences in the second- and third-grade achievement tests.

To avoid a confounding due to instrumentation differences, we can define two alternative ACEs: $ACE_k^g = E(Y_k^g(1) - Y_k^g(0))$ and $ACE_k^{g-1} = E(Y_k^{g-1}(1) - Y_k^g(0))$, or with respect to our example, $ACE_{k=8}^g = E(Y_8^3(1) - Y_8^3(0))$ and $ACE_{k=8}^{g-1} = E(Y_8^2(1) - Y_8^3(0))$. That is, under both the retention and promotion condition, we assess the 8-year-old students with the same achievement test, either with the third-grade test or the second-grade test (in practice, this would require an additional testing of either retained or promoted students). However, both ACEs are presumably not very interesting: With respect to $ACE_{k=8}^g$, we would expect that the potential third-grade retention scores ($Y_8^3(1)$) are lower than the third-grade promotion scores ($Y_8^3(0)$), just because the third-grade test requires problem-solving skills taught in the third grade but not in the second grade. And for $ACE_{k=8}^{g-1}$, we might expect that the potential second-grade retention scores ($Y_8^2(1)$) are at least as good as the second-grade promotion scores ($Y_8^2(0)$), just because of the more recent exposure to problem sets of the second grade. Thus, even if retention would have no differential effect on students' *ability growth*, we would very likely obtain a negative $ACE_{k=8}^g$ but likely a nonnegative $ACE_{k=8}^{g-1}$. Nonetheless,

both ACEs are well-defined causal estimands for evaluating retained and promoted students' *performance* on grade-specific problem sets at age k .

Since researchers using same-age comparisons are often not directly interested in grade-specific problem-solving skills of retained and promoted students but instead in students' ability growth, they aim at assessing the retention effect on students' latent ability (rather than the observed performance score). Since ability is not directly observed and standard achievement scores cannot directly be compared across grades, researchers use vertically equated scores (e.g., Hong & Raudenbush, 2006; Vandecandelaere et al. also use this strategy). With perfectly equated scores for grades $g-1$ and g assessed at ages k , $\tilde{Y}_k^{g-1}(z)$ and $\tilde{Y}_k^g(z)$, we can establish for each student the following two equalities: $\tilde{Y}_k^{g-1}(0) = \tilde{Y}_k^g(0)$ and $\tilde{Y}_k^{g-1}(1) = \tilde{Y}_k^g(1)$. With respect to our example, the equalities imply that it does not matter whether we assess a promoted or retained student's ability with the second- or third-grade test. After vertical equating, both instruments measure student ability equally well. If this is true, then the ACE with scaled scores,

$$ACE_k^s = E(\tilde{Y}_k^{g-1}(1) - \tilde{Y}_k^g(0)) = E(\tilde{Y}_k^{g-1}(1) - \tilde{Y}_k^{g-1}(0)) = E(\tilde{Y}_k^g(1) - \tilde{Y}_k^g(0)),$$

no longer depends on the choice of a grade-specific achievement test. An $ACE_{k=8}^s$ of zero would imply that kindergarten retention has the same effect as being promoted on the ability of 8-year-old students. But it does not imply that retention and promotion necessarily provide the same skills for solving problem sets of the second or third grade.

The interpretability of ACE_k^s crucially depends on the characteristics of the vertically equated scores, particularly when (a) the scores of retained and promoted at-risk students come from different tails of the grade-specific distribution of achievement scores and (b) assessments involve not only tests of two neighboring grades but tests across three or more grades (in case of multiple retentions). If the equated scores do not meet the equalities stated in the preceding paragraph, then the meaning of ACE_k^s becomes opaque and may lead to invalid inferences. For example, if the equated potential outcomes assessed with the second-grade test are systematically lower than the equated outcomes assessed with third-grade test, $\tilde{Y}_8^2(0) < \tilde{Y}_8^3(0)$ or $\tilde{Y}_8^2(1) < \tilde{Y}_8^3(1)$, the estimated retention effect would be negative even if retention has no effect on ability growth. Moreover, since the curricula for the first to second grade are taught 1 year later in the retention condition than in the promotion condition, curricula changes from one year to the next (but also other grade-specific changes) affect the treatment contrast and the generalizability of results.

Same-grade comparisons

To avoid potentially unfair comparisons of retained and promoted students due to changes in curricula and imperfect equating procedures of grade-specific achievement scores, researchers also use same-grade comparisons for assessing the effects of grade retention (e.g., Moser, West, & Hughes, 2012; Wu, West, & Hughes, 2008). Using the same promotion regime and the respective potential promotion outcomes as in the same-age comparison, we can define the average causal retention effect for the same-grade

comparison as $ACE_g = E(Y_{k+1}^g(1) - Y_k^g(0))$, that is, the expected difference in the potential retention scores assessed with the grade g test at age $k + 1$ and the potential promotion scores assessed with the same grade g test but at age k . Suppose we want to assess the effect of kindergarten retention at the third grade, then $ACE_{g=3} = E(Y_9^3(1) - Y_8^3(0))$ compares 9-year-old students in the retention condition to the very same but 1-year-younger students (8-year-old) in the promotion condition. The difference in age arises because promotion to the third grade takes an additional year under the retention condition.

It is harder in the case of ACE_g than in the case of ACE_k to conceive of a realistic intervention that actually produces the treatment contrast. If we define the promotion condition as taking the first to third grade with same-age peers, one can think of the retention condition as a sudden hypothetical intervention after the end of first kindergarten year that lets a 6-year-old student instantaneously age for 1 year and simultaneously experience the kindergarten curriculum with 5-year-old peers again. After the intervention, the now 7-year-old student attends the first to third grade just like in the promotion condition but with 1-year-younger peers (due to the sudden aging). An instantaneous intervention is necessary because the grade-specific conditions (curricula, peers, achievement tests, and other grade-specific factors) need to be identical in the retention and promotion schemes, otherwise the potential retention and promotion outcomes would differentially depend on historical events. In addition, all prekindergarten grades need to be taken at the same age in both conditions; otherwise, the potential outcomes would not refer to the same individual. Thus, the treatment contrast of the same-grade comparison is given by (a) 1 year of aging, (b) the second exposure to the kindergarten curriculum with 1-year-younger peers, and (c) taking the first to third grade with 1-year-younger peers—due to the retained students' sudden aging (second column in Table 1).

Although this is an interesting retention–promotion contrast, it is hard to imagine how it could actually be established in a credible experiment. We would need to randomly retain 6-year-old at-risk students that just finished their kindergarten year and match them to 5-year-old students that just finished the last prekindergarten year (in order to “simulate” the aging). Matching has to guarantee that the 5- and 6-year-old students are comparable in all aspects except for age. Then, the retention effect captures the double exposure to kindergarten and 1 year of aging. Note that aging is an integral part of the retention intervention.

In this sense, $ACE_{g=3}$ is a legitimate estimand for assessing the *performance* or even *ability* of retained students on the third-grade test in comparison to the 1-year-younger promoted students (Wu et al., 2008). However, $ACE_{g=3}$ is less appropriate for evaluating the retention effect that is exclusively due to the double dose of kindergarten (but not due to aging) because $ACE_{g=3}$ confounds the double-dose effect with the effect of 1 year of aging. Only if the age effect would be zero, $ACE_{g=3}$ can be interpreted as the double-dose effect. However, the absence of age effects seems implausible given that student abilities naturally grow from one year to the next—even without any exposure to a grade-specific curriculum.

Summary

We think that both the same-age and same-grade comparisons are informative and useful for practice. However, one needs to be aware of the respective treatment contrasts and the different assumptions required for identifying ACEs. Same-age comparisons using ACE_k^s are better suited for assessing the double-dose effect of repeating a grade that is not confounded with the naturally occurring ability growth due to aging. However, if vertical equating does not produce valid and reliable scores with respect to the underlying ability, one needs to be very cautious in interpreting the ACE_k^s estimates. Same-grade comparisons (ACE_g) are particularly appropriate for comparing the performance of retained and promoted students at a specific grade, particularly so when grade-specific performance builds the basis for future promotions, course or school placements, or admissions to other educational programs. The results of Vandecandelaere et al. essentially support what we can expect from the different comparisons. However, a more thorough discussion of whether the assumptions underlying the same-age and same-grade comparison are met would help in assessing the meaning and credibility of the estimated effects.

Collider bias due to deleting records based on (post)-outcome variables

In analyzing their data, Vandecandelaere et al. deleted records of early dropouts (Year 0 or Year 1) and censored records of students that dropped out in Year 2 or later (attrition) or were retained after Year 3 or more often than once. Vandecandelaere et al. handle these issues with great care. For the deleted early dropouts, they plausibly argue that this type of attrition is MCAR (missing completely at random). Attrition from Year 2 onward they addressed with censoring weights. Vandecandelaere et al. apply the same strategy to the deliberately censored data of students who do not belong to the four treatment regimes (i.e., students who got retained twice or got retained after Year 3). Censoring weights work very well provided MAR (missing at random) holds. It is important to realize that deleting instead of censoring the corresponding student records would have caused serious biases, even if one were to weight cases with respect to their baseline covariates in Year 0 and Year 1.

Here we briefly demonstrate the consequences of deleting student records according to (post)-outcome variables with a simple data-generating scenario with 3 years of schooling as depicted in Figure 1. The time-varying treatment variable T_t ($t = 1, \dots, 3$) represents the sequence of grades taken by a student over 3 years. If a student gets promoted in all 3 years, the treatment variables take on values $T_1 = g1$, $T_2 = g2$, and $T_3 = g3$, where $g1$ corresponds to grade 1, $g2$ to grade 2, and $g3$ to grade 3. For a student who gets retained only in grade $g1$, the treatment sequence is $T_1 = g1$, $T_2 = g1$, and $T_3 = g2$. Similar sequences can be constructed for all possible retention regimes. The outcome measures Y_t refer to vertically equated achievement scores measured at the end of year t and reflect the underlying student abilities A_t . The treatment status T_t —that is, whether a student is retained or promoted to the next grade in year t —depends only on the observed achievement score Y_{t-1} and the treatment status T_{t-1} of the previous year (but it does not depend on ability because no arrows directly connect A_{t-1} and T_t). The short arrows pointing into all the variables indicate independent but unobserved factors or error terms.

Assume that we want to estimate the effect of being retained in T_2 (i.e., repetition of grade g_1) on the outcomes Y_2 and Y_3 . From the graph in Figure 1, we directly see that the ACE on Y_2 (i.e., $T_2 \rightarrow A_2 \rightarrow Y_2$) is identified because conditioning on the observed pretest Y_1 and treatment T_1 blocks all three confounding paths $T_2 \leftarrow Y_1 \leftarrow A_1 \rightarrow A_2 \rightarrow Y_2$, $T_2 \leftarrow T_1 \rightarrow A_1 \rightarrow A_2 \rightarrow Y_2$, and $T_2 \leftarrow T_1 \leftarrow Y_0 \leftarrow A_0 \rightarrow A_1 \rightarrow A_2 \rightarrow Y_2$. Similarly, one effect of static treatment regimes (T_2, T_3) on Y_3 are sequentially identified (Pearl, 2009).

Now assume that we deleted the records of students with missing values in Y_2 and Y_3 (with nonresponse or attrition being MAR or NMAR) and the records of students who got retained in T_3 because they do not belong to the treatment regime of interest (i.e., we are only interested in evaluating the effect of being retained in year $t = 2$ without being retained in year $t = 3$).² Deleting entire student records implies that the ACE needs to be identified from the subsample of remaining students that we visualized by putting dashed boxes around the conditioning variables (Figure 2).³ Since the conditioning (post)-treatment variables Y_2, Y_3 , and T_3 are collider variables but also descendants of other collider variables ($A_1, A_2, A_3, Y_0, Y_1, Y_3, T_1$, and T_2), deleting records induces a series of spurious relations between the unobserved abilities, retention statuses, and outcomes (Elwert & Winship, 2014). If the observed variables fail to block these spurious associations, collider bias distorts the effect estimates of interest.

The dashed edges in Figure 2 only show the bias-inducing spurious relations that cannot be blocked by sequentially conditioning on the observed achievement scores (Y_0, Y_1 , and Y_3) or treatment variables (T_1 and T_2). Thus, the retention effects on Y_2 and Y_3 are no longer identified because all noncausal paths from T_2 to Y_2 (and Y_3) that contain dashed edges induce collider bias. These unblockable noncausal paths confound the causal relation between T_2 and Y_2 either directly ($T_2 \text{ --- } Y_2$) or indirectly via the unobserved abilities (e.g., $T_2 \text{ --- } A_1 \rightarrow A_2 \rightarrow Y_2$). The same holds for the sequential identification of the (T_2, T_3) regimes' effect on Y_3 . It is important to realize that the noncausal paths cannot be blocked even if the nonresponse process or retention in T_3 is exclusively determined by the observed variables (MAR). This is so because the collider bias operates predominantly via the unobserved abilities.

However, censoring instead of deleting the corresponding student records is able to circumvent these bias issues but only if nonresponse is MAR and censoring T_3 -retained student records is ignorable given the observed pretreatment covariates, precensoring treatment, and outcome history (as pointed out by Vandecandelaere et al.).⁴ First, consider the effect of T_2 on Y_2 . Censoring based on T_3 cannot cause any bias because no data are deleted for timepoints 1 and 2; thus, the complete data set is analyzed. Censoring based on Y_2 and Y_3 does not induce collider bias either because no entire student records are deleted—all data up to T_2 remain complete. However, the missingness in Y_2 causes nonresponse bias, but if the nonresponse mechanism can be assumed to be MAR, censoring weights

²Note that the very same issues occur if we would delete records only according to the outcomes Y_2 and Y_3 or treatment T_3 alone.

³The boxes are a simplified way to indicate that we select a subsample of students according to the corresponding conditioning variables (Elwert & Winship, 2014; Steiner, Kim, Hall, & Su, 2015). However, conditioning on nonmissing data usually requires a more complex graphical representation that also shows the nonresponse mechanism (Thoemmes & Kohan, 2015).

⁴MAR and ignorability refer essentially to the same set of assumptions (MAR is typically used for nonresponse processes only).

adjust for the bias. Thus, the causal effect of T_2 on Y_2 is identified. Now consider the effect of T_2 on Y_3 where censoring in T_3 actually induces confounding bias (but no collider bias). Since censoring is ignorable conditional on the observed T_2 and Y_2 (at least in our simple data-generating model), sequential censoring weights are able to take care of the confounding bias.

However, if nonresponse is NMAR or censoring T_3 -retained students is not ignorable (because not all treatment-outcome confounders are observed), censoring weights will fail to fully remove nonresponse and censoring bias. Vandecandelaere et al. have been fully aware of these issues; instead of deleting student records, they relied on MAR and ignorability assumptions for the censored and addressed the biases via time-varying censoring weights. Although the assumptions need to be justified on subject-matter knowledge, we believe that Vandecandelaere et al. set new standards for future research on grade retention.

Acknowledgments:

Funding: This work was supported by Grant R305D120005 from the Institute of Education Sciences.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions or the Institute of Education Sciences is not intended and should not be inferred.

References

- Elwert F , & Winship C (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40, 31–53. 10.1146/annurev-soc-071913-043455
- Hong G , & Raudenbush SW (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. 10.1198/016214506000000447
- Moser S , West SG , & Hughes JN (2012). Trajectories of math and reading achievement in low achieving children in elementary school: How are they affected by retention in first and later grades? *Journal of Educational Psychology*, 104, 603–621. 10.1037/a002757123335818
- Pearl J (2009). *Causality: Models, reasoning, and inference*. (2nd ed.). New York, NY: Cambridge University Press.
- Steiner PM , Kim Y , Hall C , & Su D (2015). Graphical models for quasi-experimental designs. *Sociological Methods and Research*, 51(6), 865–870. 10.1177/0049124115582272
- Thoemmes F , & Mohan K (2015). Graphical representation of missing data problems. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(4), 631–642. 10.1080/10705511.2014.937378
- Wu W , West SG , & Hughes JN (2008). 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. 10.1037/a001309819337582

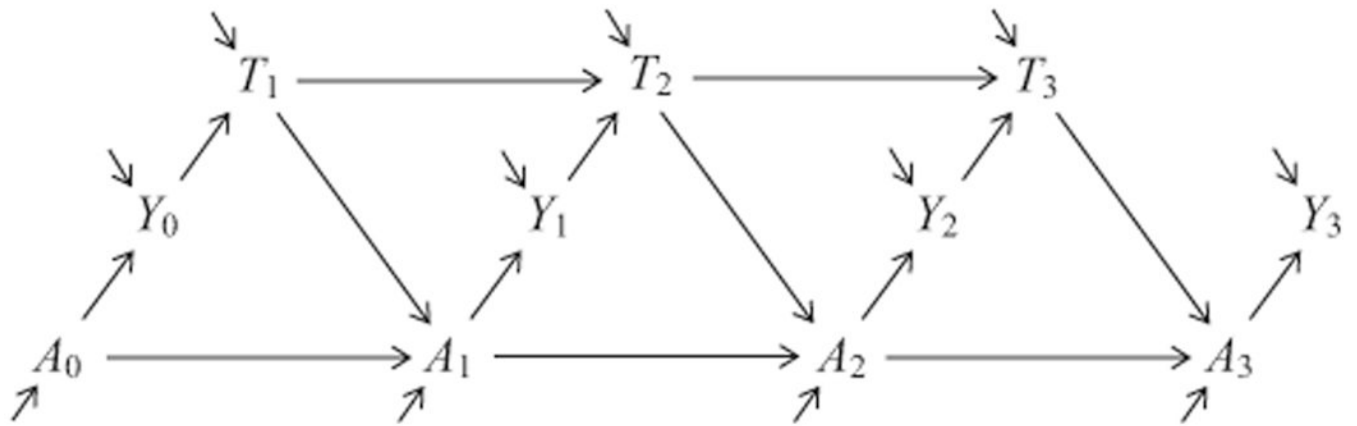


Figure 1. Data-generating graph. T_t = the treatment variables; Y_t = the achievement scores; A_t = the unobserved abilities (for $t = 0, \dots, 3$).

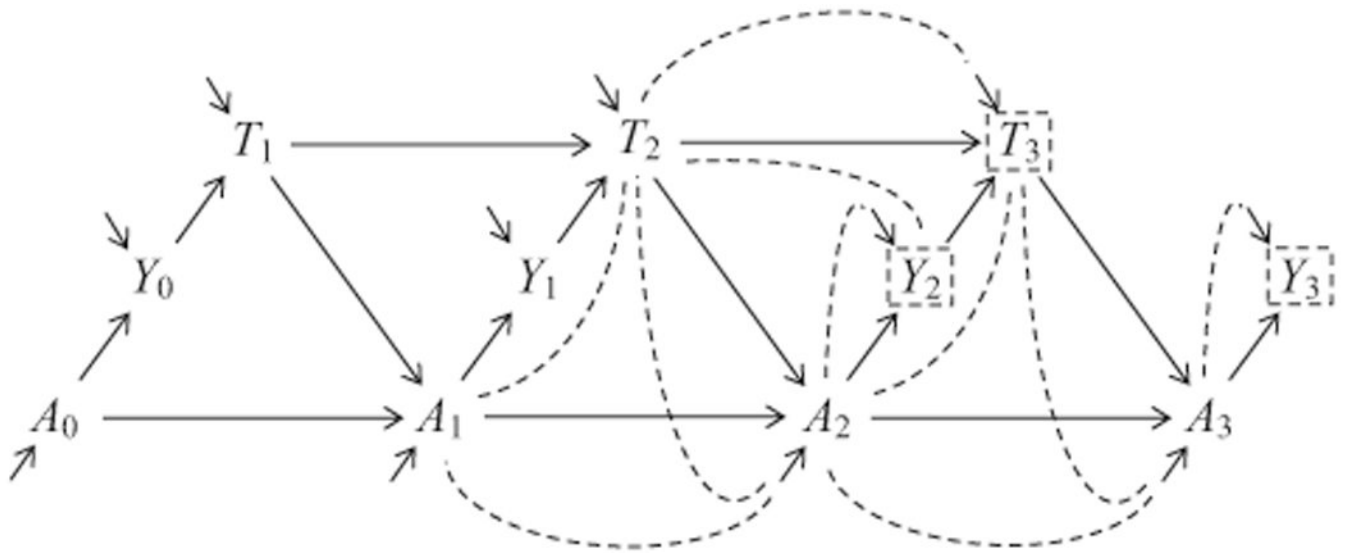


Figure 2. Unblockable spurious associations (dashed edges) that bias the effect of retention in T_2 on Y_2 and Y_3 due to deleting entire student records based on Y_2 , Y_3 , and T_3 (as indicated by the dashed boxes). T_t = the treatment variables; Y_t = the achievement scores; A_t = the unobserved abilities (for $t = 0, \dots, 3$).

Table 1.

Treatment conditions, measurements, and causal estimands of same-age and same-grade comparisons of repeating grade g^* .

	Same-age comparison	Same-grade comparison
	Intervention	
Type of intervention	actual implementable intervention	sudden hypothetical intervention
Control condition (promotion)	grade $g^* + 1$ to $g^* + t$ curricula with same-age peers	grade $g^* + 1$ to $g^* + t$ curricula with same-age peers
Treatment condition (retention)	grade g^* to $g^* + t - 1$ curricula 1 year later with 1-year-younger peers	1 year of aging; grade g^* to $g^* + t$ curricula with 1-year-younger peers
Treatment contrast (retention vs. promotion)	(a) second exposure to grade g^* curriculum with 1-year-younger peers; (b) grades $g^* + 1$ to $g^* + t - 1$ one year later with 1-year-younger peers; (c) absence of grade $g^* + t$ curriculum	(a) 1 year of aging; (b) second exposure to grade g^* curriculum with; (c) 1-year-younger peers 1-year-younger peers from grade $g^* + 1$ to $g^* + 1$
	Instruments and measurements	
Treatment condition	grade $g-1$ test and scores	grade g test and scores
Control condition	grade g test and scores	grade g test and scores
	Causal estimands	
Main estimand	$ACE_k^s = (\tilde{Y}_k^{g-1}(1) - \tilde{Y}_k^g(0))$	$ACE_g = E(Y_{k+1}^g(1) - Y_k^g(0))$
Alternative estimands	$ACE_k^g = E(Y_k^g(1) - Y_k^g(0))$ $ACE_k^{g-1} = E(Y_k^{g-1}(1) - Y_k^{g-1}(0))$	

Note. Grade $g \in \{0, \dots, G\}$ refers to a grade between kindergarten ($g = 0$) and a final grade G ($g = G$). Students can get retained once at grade g^* but at no other grade ($0 < g^* < G$). We are interested in the average causal effects (ACEs) t years after the retention decision. For a given $t > 0$, intervention grade g^* , and a kindergarten entrance age of 5 years, a student's assessment age is given by $k = 5 + g^* + t$. Potential promotion and retention outcomes, $Y_k^g(0)$ and $Y_k^g(1)$, refer to student achievements assessed with the instrument of grade g at age k . Potential outcomes $\tilde{Y}_k^g(0)$ and $\tilde{Y}_k^g(1)$ denote the vertically equated scores.