The Inequality Lab.
Discussion Paper 2018-1

Felix Elwert & Fabian T. Pfeffer

The Future Strikes Back. Using Future Treatments to Detect and Reduce Hidden Bias.

July 2018
www.theinequalitylab.com

THE FUTURE STRIKES BACK: USING FUTURE TREATMENTS TO DETECT AND REDUCE HIDDEN BIAS

Felix Elwert*
University of Wisconsin-Madison

Fabian T. Pfeffer
University of Michigan

Abstract

Conventional advice discourages controlling for post-outcome variables in regression analysis. Here, we show that controlling for commonly available post-outcome (i.e. future*) values of the treatment variable can help detect, reduce, and even remove omitted variable bias (unobserved confounding). The premise is that the same unobserved confounders that affect treatment also affect future values of the treatment. Future treatments thus proxy for the unmeasured confounder, and researchers can exploit these proxy measures productively. We establish several new results: Regarding a commonly assumed data-generating process involving future treatments, we (1) introduce a simple new approach to reduce bias and show that it strictly reduces bias; (2) elaborate on existing approaches and show that they can increase bias; (3) assess the relative merits of approaches; (4) analyze true state dependence and selection as key challenges; and (5) demonstrate that future treatments can test for hidden bias, even when they fail to reduce bias. We illustrate these results empirically with an analysis of the effect of parental income on children's educational attainment.*

* This work was supported by a grant from the University of Wisconsin Graduate School Research Competition. We thank NE Barr for copy-editing and gratefully acknowledge use of the services and facilities of the Population Studies Center at the University of Michigan, funded by NICHD Center Grant R24 HD041028. The collection of data used in this study was partly supported by the National Institutes of Health under grant number R01 HD069609 and the National Science Foundation under award number 1157698. A replication package containing the data and code used for the empirical illustration in this paper is available through the PSID Public Data Extract Repository at https://www.openicpsr.org/openicpsr/psid (#104060).

INTRODUCTION

FUTURE-TREATMENT STRATEGIES

Hidden bias from unobserved confounding is a central problem in the social sciences. If unobserved variables affect the treatment and the outcome, then regression and matching estimators cannot recover causal effects (e.g., Rosenbaum 2002, Morgan and Winship 2015). One set of strategies for mitigating confounding bias that has been used in scattered contributions from sociology and economics involves *future treatments*, i.e. values of the treatment that are realized after the outcome has occurred. The basic intuition behind these strategies is that the same unobserved confounders that affect the treatment variable before the outcome often also affect future values of the treatment variable, measured after the outcome. If so, future values of the treatment are proxy measures of the unmeasured confounder and may help remove bias.

A few authors have previously appealed to this intuition and proposed a variety of different estimators. For instance, prior research has exploited future treatments in structural equation models (Mayer 1997), used future treatments to measure and subtract unobserved bias (Gottschalk 1996), and employed them as instrumental variables (Duncan et al. 1997).[1] We will critically assess some of these earlier strategies and compare them to our simpler proposal to use future treatments as control variables to remove bias by proxy.

We posit that future-treatment strategies hold significant promise for social science research for several reasons. First, future treatments can help detect, reduce, and even remove bias from unobserved confounding. Second, future values of the treatment are routinely available in panel data. Third, since future-treatment strategies require only that the *treatment variable* varies over time (i.e., not the outcome), they are available even when individual-level fixed-effects panel estimators are not. Fourth, since different future-treatment strategies impose different assumptions about the data generating process, they are applicable across a wide range of different substantive settings.

In this paper, we analyze several prior uses of future-treatment strategies and propose a new strategy. We discuss the conditions under which future values of the treatment can reduce or remove confounding bias. We also highlight the conditions under which future-treatment strategies introduce more bias than they remove. Specifically, we show the challenges of using future-treatment strategies in two scenarios: where the outcome affects future treatment (selection), and where past treatment affects future values of the treatment (true state dependence). Yet, even where future-treatment strategies fail to

[1] Other examples of research that purposefully subverts the common temporal order include the correlated random effects model proposed by Chamberlain (1982) as well as other applied contributions that consider a comparison group that only experiences treatment in the future, such as future incarceration (e.g. Grogger 1995; Wildeman 2010; Porter and King 2014;) or a future network tie (e.g. Kim et al. 2015).

reduce bias, they may still be useful for detecting the presence of bias; and we discuss a non-parametric test for bias detection using future treatments.

We investigate the performance of future-treatment strategies across a range of data generating processes, and assess their relative performance compared to regular regression estimates without corrections for unobserved confounding. We present our analysis in two complementary formats. First, we present our analysis graphically to assist applied social scientists in determining quickly whether a future-treatment strategy is appropriate for their specific substantive application. Second, we assume linearity to link our graphical results to familiar regression models and quantify biases. (Appendices discuss related approaches and instrumental variables estimation with future treatments.) Finally, we illustrate the application of future-treatment strategies with an empirical example that estimates the effect of parental income on children's educational attainment.

PRELIMINARIES: DAGS, LINEAR MODELS, AND IDENTIFICATION

In this section, we describe the tools of our formal identification analyses, following Pearl (2013). Since the causal interpretation of statistical analyses is always contingent on a theoretical model of data generation, we first review directed acyclic graphs (DAGs) to notate the assumed data-generating process (DGP). Second, we state Wright's (1921) rules, which link the causal parameters of the DGP to observable statistical associations (covariances and regression coefficients) in linear models. Readers familiar with DAGs and Wright's rules may prefer to skip this section.[2]

We use DAGs to notate the causal structure of the analyst's presumed DGPs (Pearl 2009; for an introduction, see Elwert 2013). DAGs consist of variables, and arrows that represent the direct causal effects between variables. We will focus on DAGs comprising four (vectors of) variables: a treatment, T, an outcome, Y, a future (post-outcome) value of the treatment, F, and a vector of unobserved variables, U. In keeping with convention, we assume that the DAG shows all common causes shared between variables, regardless of whether these common causes are observed or unobserved. For example, in the DAG $T \leftarrow U \rightarrow Y$, U represents all unobserved common causes between T and Y.

DAGs empower the analyst to determine whether the observed association (e.g. a regression coefficient) between treatment and outcome identifies the causal effect or is biased. The observed association between treatment and outcome *identifies* the causal effect if the only open path connecting treatment and outcome is the causal pathway, $T \rightarrow Y$. The association between treatment and outcome is spurious, or biased for the causal effect, if at least one open path does not trace the causal pathway (e.g., $T \leftarrow U \rightarrow Y$). Whether a path is open (transmits association) or closed (does not transmit association) depends on what variables are controlled in the analysis, and whether the path contains a collider variable. A collider is a variable that receives two inbound arrows, such as C in

[2] Throughout, we assume large samples in order to focus on identification.

$A \rightarrow C \leftarrow B$ (see Elwert and Winship 2014). A path is closed if it contains an uncontrolled collider, or a controlled non-collider; and is open otherwise.

Unless otherwise stated, we assume a linear DGP, the conventional workhorse of social science. The assumption of linearity may not always be terribly realistic, but it has the advantage of convenience, as it links DAGs directly to OLS regression and conventional SEM methodology. Under linearity, DAGs become linear path models, and every arrow in a linear path model is fully described by its *path parameter*, p, which quantifies its linear and homogenous direct causal effect. Since path parameters are causal effects, they cannot be observed directly.[3]

Without loss of generality, we assume standardized variables throughout (zero mean and unit variance). Standardized path parameters cannot exceed 1 in magnitude. To prevent model degeneracies, we assume that all path parameters lie strictly inside the interval $(-1, 1)$ and differ from zero: $-1 < p < 1$ and $p \neq 0$.

Wright's (1921) path rules link the unobserved path parameters of the presumed linear DGP to observable covariances.

Wright's (1921) path rule: The marginal (i.e., unadjusted) covariance between two standardized variables A and B, σ_{AB}, equals the sum of the products of the path parameters along all open paths between A and B.

That is, to calculate the marginal covariance between two variables A and B, we first compute the product of the path parameters for each of the open paths between A and B, and then sum these products across all open paths.

To link the coefficients of OLS regression (with or without control variables) to the underlying path parameters via Wright's rule, we express the coefficients in terms of marginal covariances. The regression coefficient on T in the unadjusted regression $Y = b_{YT}T + u$ with standardized variables equals the marginal covariance between Y and T,

$$b_{YT} = \sigma_{YT} \ . \qquad (1)$$

We call b_{YT} the *unadjusted coefficient* on T. The partial regression coefficient on T after controlling for F in the regression $Y = b_{YT.F}T + b_{YF.T}F + u$ is given by,

$$b_{YT.F} = \frac{\sigma_{YT} - \sigma_{YF}\sigma_{FT}}{(1-\sigma_{FT}^2)} \ . \qquad (2)$$

We call $b_{YT.F}$ the *F-adjusted coefficients* on T. Analogously, the *T-adjusted coefficient* on F is given by

$$b_{YF.T} = \frac{\sigma_{YF} - \sigma_{YT}\sigma_{FT}}{(1-\sigma_{FT}^2)} \ . \qquad (3)$$

We omit observed control variables (other than F) from the analysis because they do not

[3] Path parameters are often called "path coefficients." We write "parameter" to denote true causal effects in the DGP, and we write "coefficient" to denote statistical quantities, such as regression coefficients, which may or may not equal the desired parameter.

contribute to intuition. All of our results generalize to the inclusion of pre-treatment control variables.[4]

Putting these elements together, the subsequent analysis proceeds in four steps. First, we draw the DAG of a candidate DGP. Second, we use Wright's rule to express the marginal covariances between observed variables in terms of the true path parameters. Third, we plug these covariances into the formulas for the regression coefficients in equations 1-3. Finally, we investigate whether any of these regression coefficients, or functions of regression coefficients, equal (or "identify") the desired causal effect of the treatment on the outcome and quantify possible biases.

THE PROBLEM: UNOBSERVED CONFOUNDING

Figure 1 highlights problem of unobserved confounding and illustrates our running example.

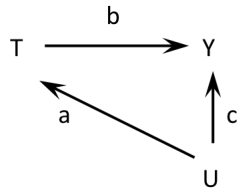


Figure 1. DAG for an observational study of parental income, T, on children's years of education, Y, with unobserved confounder(s), U.

The DAG shows the DGP for an observational study to estimate the total causal effect of a treatment, T (e.g., parental income), on an outcome, Y (e.g., children's years of completed education). Since treatment is not randomized, the effect of T on Y is usually confounded by one or more unobserved factors, U, that jointly affect T and Y (e.g., parental ambition). If so, the unadjusted association between T and Y will be biased for the causal effect of T on Y, because the association will be a combination of the association transmitted along the open causal path $T \rightarrow Y$ and the open noncausal path $T \leftarrow U \rightarrow Y$. (If all confounding variables U are measured, then controlling for the non-collider U removes all bias by closing the noncausal path $T \leftarrow U \rightarrow Y$.) Henceforth, we assume that at least some confounding factors, U, are unobserved. This mimicks the main predicament of most observational study in the social sciences.

If Figure 1 represents a linear model, then, by equation 1 and Wright's path rule, the unadjusted regression coefficient on T, equals

$$b_{YT} = \sigma_{YT} = b + ac \ . \qquad (4)$$

[4] We assume that controlling for pre-treatment variables reduces bias from unobserved confounding, as it usually does. For counterexamples, see Elwert and Winship (2014).

This regression coefficient is obviously biased for the true causal effect of T on Y, b. The bias equals $B_{OLS} = b_{YT} - b = ac$, and increases in magnitude with the effects $U \rightarrow T$, a, and $U \rightarrow Y$, c. Removing this bias from unobserved confounding is the central task of observational causal inference in the social sciences.

STRATEGIES OF BIAS CORRECTION WITH FUTURE TREATMENTS

Future treatments can be used to reduce and even remove bias from unobserved confounding, depending on both the analytic strategy (e.g., the chosen regression specification) and the DGPs. In this section, we introduce two future-treatment strategies under the assumptions of the DGP shown in Figure 2. This model represents a best-case scenario for future-treatment strategies and is commonly assumed in the literature (e.g., Mayer 1997). The model assumes that the causal effect of T on Y is confounded by one or more unobserved variables, U, and that the future value of the treatment, F, is affected by all U that affect treatment, T. In other words, F is assumed to be a proxy measure for U.

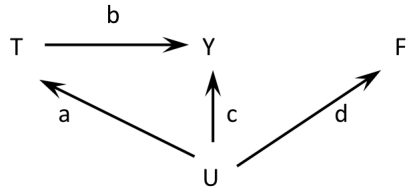


Figure 2. A confounded study where the future value of the treatment, F, is a proxy for the unobserved confounder(s), U.

The assumption that all confounders of T and Y also affect F is central for future-treatment strategies. Because the assumption cannot be tested empirically, it has to be defended on theoretical grounds. In many applications, it is eminently credible. For example, if parents' unmeasured ambition, U, affects parental income, T, prior to the child completing education, Y, it likely also affects parental income after the child has completed education, F.

Control Strategy of Future Treatments

Most future-treatment strategies in one way or another exploit the fact that F is a proxy for U. Here, we propose a simple estimator that exploits this fact directly: Since F is a proxy that carries information about U, controlling for F in the regression $Y = b_{YT.F}T + b_{YF.T}F + u$ partially controls for U and hence reduces bias in the treatment-effect estimate. We call the strategy of bias reduction by outright controlling for F the *control*

strategy of future treatments.

Definition 1 (control strategy estimator): The control-strategy estimator, b_C, for the causal effect of T on Y, b, is given by the F-adjusted regression coefficient on T,

$$b_C = b_{YT.F} = \frac{\sigma_{YT} - \sigma_{YF}\sigma_{FT}}{(1-\sigma_{FT}^2)} . \quad (5)$$

Result 1 evaluates the control strategy estimator for data generated by the DGP in Figure 2:

Result 1 (bias of the control strategy estimator in the best case): In data generated by the DGP in Figure 2, the control strategy estimator evaluates to:

$$b_C = b + ac\frac{(1-d^2)}{(1-a^2d^2)} = b + B_{OLS}M_C . \quad (6)$$

Clearly, the control estimator remains biased, because $b_C \neq b$. Result 2, however, states that the control strategy estimator always improves on the OLS estimator.

Result 2 (strict bias reduction of the control strategy estimator in the best case): In data generated by the DGP in Figure 2, the control strategy estimate is strictly less biased than the OLS estimate.

To see this, note that the control strategy estimator multiplies the OLS bias, $B_{OLS} = ac$, by the factor $M_C = \frac{(1-d^2)}{(1-a^2d^2)}$, which we call the bias multiplier of the control strategy. Since all path parameters are standardized, the magnitude of the control-bias multiplier is always less than 1, $|M_C| < 1$, and hence deflates the OLS bias, $|B_{OLS}M_C| < |B_{OLS}|$. Strict bias reduction is the key advantage of the control strategy.

Figure 3 illustrates bias reduction in the control strategy estimator compared to the unadjusted OLS estimator by graphing the absolute value of the bias multiplier of the control strategy, $|M_C|$ (dashed blue line), against the reference of no bias reduction (line at $|M_C| = 1$) as a function of the strength of the effect of U on F, d, for a moderately strong effect of U on T, $a = 0.4$.[5] Clearly, the control-bias multiplier $|M_C|$ is always between 0 and 1 and hence guarantees bias reduction regardless of sign or size of the path parameters.

The stronger the effect of U on F, $|d|$, the more bias is removed. This makes intuitive sense: the stronger the effect of U on F, the better F proxies for U. In the extreme case, where F is perfectly determined by U, $|d| = 1$, controlling for F amounts to controlling for U itself, thus removing all bias, such that $b_C = b$.

[5] We pick $a = 0.4$ for illustration. Results are qualitatively the same for other values of a.

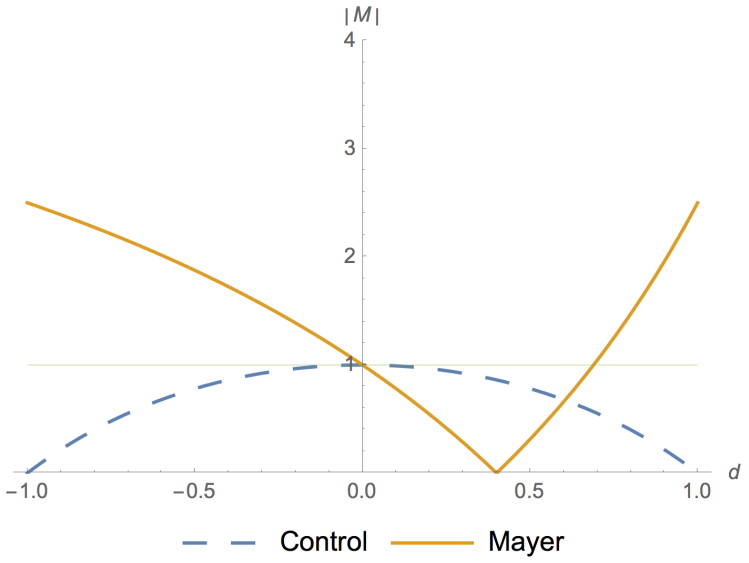


Figure 3. Absolute bias multiplier, $|M|$, for the control estimator and Mayer's estimator as a function of the effect $U \to F$, d. $|M| = 1$ indicates no change compared to OLS bias. $|M| > 1$ indicates bias amplification, $|M| < 1$ bias reduction. Graphed for a moderate effect $U \to T$, $a = 0.4$.

The control strategy gives applied social scientists a straightforward tool for reducing bias from unobserved confounding. All it takes is adding F as a regressor to the regression of Y on T. To return to our running example, under the model assumptions of Figure 2, bias in the estimated effect of parental income measured before children complete education will be strictly reduced by controlling for future parental income measured after children complete education.

Mayer's Strategy

Mayer (1997) takes a different approach to bias reduction with future treatments. Instead of simply controlling for F in a regression model, she solves the structural equations of the DGP in Figure 2 under the additional assumption that the unobserved confounder, U, affects the future treatment, F, exactly like it affects the treatment, T, $a = d$. This assumption may be defensible in some circumstances. In our running example, one might hypothesize that parental ambition is relatively time-invariant and affects parental income, T and F, similarly at all times.

Under the assumption that $a = d$, the three observable covariances between T, Y, and F in Figure 2, by Wright's rule, are functions of three unknown path parameters,

$$\begin{aligned} \sigma_{YT} &= b + ac \\ \sigma_{YF} &= a^2 b + ac \\ \sigma_{TF} &= a^2 \,. \end{aligned} \tag{7}$$

This system is solved uniquely for the desired causal effect,

$$\frac{\sigma_{YT}-\sigma_{YF}}{1-\sigma_{TF}} = \frac{b+ac-ac-a^2b}{1-a^2} = \frac{b(1-a^2)}{1-a^2} = b \text{ .}^{6} \quad (8)$$

Definition 2 (Mayer's [1997] estimator): Mayer's estimator for the causal effect of T on Y, b, is given by

$$b_M = \frac{\sigma_{YT}-\sigma_{YF}}{1-\sigma_{TF}} \text{ .} \quad (9)$$

The advantage of Mayer's estimator is that it removes all bias under the assumptions that the data are generated as in Figure 2 and that U affects T exactly as it affects F, $a = d$. However, when U affects T and F differently, $a \neq d$, then Mayer's estimator has two disadvantages. First, as Mayer (1997) notes, the estimator is biased.[7] Result 3 evaluates the bias.

Result 3 (bias of Mayer's [1997] estimator in the best case): In data generated by the DGP of Figure 2, Mayer's estimator evaluates to

$$b_M = b + ac\frac{a-d}{a-a^2d} = b + B_{OLS}M_M \text{ .} \quad (10)$$

Second, in contrast to our control strategy estimator, Mayer's estimator can increase OLS bias, as shown in Result 4.

Result 4 (bias amplification in Mayer's [1997] estimator in the best case): In data generated by the process of Figure 2, Mayer's estimator increases bias compared to the OLS estimate when $|M_M| = |\frac{a-d}{a-a^2d}| > 1$. This occurs (1) when $\frac{a}{d} < 0$ or (2) when $|\frac{2a}{1+a^2}| < |d|$.

In other words, bias amplification occurs either (1) when U affects T and F in opposite directions, or (2) when U affects T and F in the same direction but the magnitude of the effect $U \rightarrow F$, d, substantially (roughly more than twice) exceeds the magnitude of the effect $U \rightarrow T$, a.

---

[6] The first two equations are collinear if past and future values of the treatment are very similar, i.e., $a = d$ approaches 1. As a increases, the denominator of Mayer's estimator, $1 - a^2$, shrinks toward zero. Consequently, standard errors will increase with the magnitude of a.

[7] When $a \neq d$, the three observable covariances between T, Y, and U, produce three equations with four unknowns: (1) $\sigma_{YT} = b + ac$; (2) $\sigma_{YF} = abd + cd$; and (3) $\sigma_{TF} = ad$, which cannot be solved uniquely for b. Mayer (1997: p. 178) seeks to address this scenario by solving for b based on the $a = d$ assumption but then adjusting the resulting treatment effect estimate for differences in a and b assuming that they follow the same pattern as differences in the effects of observable characteristics measured before and after the outcome.

The solid orange line in Figure 3 illustrates bias reduction and bias amplification of Mayer's estimator by graphing the absolute value of the bias multiplier, $|M_M|$ across values of d for $a = 0.4$. When a and d share a sign (here, $d > 0$) and d is not much larger than a, then $|M_M| < 1$, and Mayer's estimator reduces bias. But if a and d have opposite signs (here, $d < 0$), or if $d \gg a$, then $|M_M| > 1$ and Mayer's estimator amplifies the OLS bias.

The possibility of bias amplification in Mayer's estimator has not been noted previously. Whether bias amplification occurs depends on the empirical setting and must be carefully evaluated based on sociological subject matter knowledge. We believe that bias amplification can be excluded in many settings. First, in many applications U will not affect T and F in opposite directions. In our example, it does not appear plausible that parental ambition, U, increases parental income early on, T, but decreases it later, F. Second, since the shared unobserved confounder U is by assumption a baseline characteristic that is temporally closer to T than to F, the effect of U on T will likely exceed the effect of U on F, i.e., $|a| > |d|$. In our example, we are cautiously optimistic that the effect of early parental ambition is more pronounced on parent's early income, T, than on later income, F, because other determinants of income, such as experience and seniority, may grow in importance as time passes.

On the other hand, we cannot entirely rule out the possibility of bias amplification, even in our running example. Suppose, for example, that we analyze the effect of parental income on children's educational outcomes among young parents. Young parents with high ambition may still be enrolled in college and hence earn little compared to their lower-ambition counterparts who already have jobs. Later, however, these highly ambitious parents may become high-earning professionals, whereas their lower-ambition counterparts may remain in lower paying jobs. Hence, the effects $U \rightarrow T$ and $U \rightarrow F$ could have opposite signs, such that Mayer's estimator would increase rather than decrease bias. And even if the effects share the same sign, $U \rightarrow F$ may strongly exceed $U \rightarrow T$. That is, using our example, if the returns to parental ambition compound as employees climb up the corporate ladder, early ambition may have a relatively modest effect on early income but a large effect on later income via successive promotions. If the effect of early ambition on later income sufficiently exceeds its effect on early income, then Mayer's estimator would also increase rather than decrease bias.

Implementing Mayer's Strategy as a Difference Estimator

The original presentation of Mayer's estimator required customized programming. Next, we show that Mayer's estimator can be expressed straightforwardly as the difference between two regression coefficients. This enables estimation via all standard statistical software packages and provides additional intuition.

Definition 3 (difference estimator): The difference estimator for the effect of T on Y is

the difference between the adjusted coefficients on T and F in the regression $Y = b_{YT.F}T + b_{YF.T}F + u$,

$$b_D = b_{YT.F} - b_{YF.T} = \frac{\sigma_{YT} - \sigma_{YF}\sigma_{FT}}{(1-\sigma_{FT}^2)} - \frac{\sigma_{YF} - \sigma_{YT}\sigma_{FT}}{(1-\sigma_{FT}^2)}\,. \qquad (11)$$

Result 5 (equivalence of the difference and Mayer's estimators):

$$b_D = \frac{\sigma_{YT} - \sigma_{YF}\sigma_{FT}}{(1-\sigma_{FT}^2)} - \frac{\sigma_{YF} - \sigma_{YT}\sigma_{FT}}{(1-\sigma_{FT}^2)} = \frac{(1+\sigma_{FT})(\sigma_{YT} - \sigma_{YF})}{(1+\sigma_{FT})(1-\sigma_{FT})} = \frac{(\sigma_{YT} - \sigma_{YF})}{(1-\sigma_{FT})} = b_M \;\square \qquad (12)$$

The equivalence between Mayer's estimator and the difference estimator holds for all DGPs—not just the DGP of Figure 2—, because the definition of the estimators only draws on empirical covariances and does not appeal to the structure of the DGP.

Equating Mayer's estimator with the difference estimator provides additional insight: The idea behind the difference estimator is to use future treatments first to measure and then to remove the spurious association between T and Y.

This fact is best appreciated by investigating the difference estimator under the assumption that the effect of U on T equals the effect of U on F, $a = d$, in data generated by Figure 2. First, the coefficient $b_{YT.F}$ is biased for b by the confounding path $T \leftarrow U \rightarrow Y$, less whatever part of confounding is removed by controlling for F (recall that F is a proxy for U). Specifically, $b_{YT.F} = b + ac\frac{(1-a^2)}{(1-a^4)}$, where $0 < \frac{(1-a^2)}{(1-a^4)} < 1$ is the deflation factor by which confounding along $T \leftarrow U \rightarrow Y$, ac, is diminished by controlling for F. Second, the coefficient $b_{YF.T}$ captures the association flowing along the path $T \leftarrow U \rightarrow F$, less whatever part of this association is removed by controlling for T (like F, T is a proxy for U). Specifically, $b_{YF.T} = ac\frac{(1-a^2)}{(1-a^4)}$, which equals the association flowing along $F \leftarrow U \rightarrow Y$, diminished by the same deflation factor $0 < \frac{(1-a^2)}{(1-a^4)} < 1$ due to controlling for T. Third, clearly, $b_{YF.T}$ equals the bias in $b_{YT.F}$; hence subtracting one from the other yields an unbiased estimate for b.

Expressing Mayer's estimator as a difference estimator helps explicate the properties that we claimed for it above. First, the Mayer/difference estimator removes all bias only if $a = d$, because only then does $b_{YF.T}$ exactly measure the bias in $b_{YT.F}$. More generally, by Result 3, the estimator equals $b_M = b_D = b + ac\frac{a-d}{a-a^2d}$ and is biased to the extent that a and d differ.

Second, if $a > d$, the estimator is biased because $b_{YF.T}$ understates the bias in $b_{YT.F}$: the association captured by the path $Y \leftarrow U \rightarrow F$ understates the bias flowing along $Y \leftarrow U \rightarrow T$.

Third, if $a < d$, the estimator is biased because $b_{YF.T}$ overstates the bias in $b_{YT.F}$: the association captured by the path $Y \leftarrow U \rightarrow F$ overstates the bias flowing along $Y \leftarrow U \rightarrow T$.

Fourth, if d is more than twice as large as a, then $b_{YF.T}$ may overstate the bias in $b_{YT.F}$

more than twofold, so that the Mayer/difference estimator $b_{YT.F} - b_{YF.T}$ first subtracts all bias and then more than adds it back, resulting in absolute bias amplification.

Fifth, if a and d have different signs, then $b_{YF.T}$ measures the negative of the bias in $b_{YT.F}$ such that the difference estimator $b_D = b_{YT.F} - b_{YF.T}$ adds rather than removes bias, also resulting in bias amplification.

We note that the difference estimator for future treatments has some history in social science methodology. Versions of this differencing logic are discussed by Gottschalk (1996), who explicitly uses future treatments (but not this exact estimator, see Appendix A), and by DiNardo and Pischke (1997) and Elwert and Christakis (2008), who analyze structurally similar models without future treatments.

The Difference-Strategy of Future Treatments Is Different from Difference-In-Difference

Despite superficial similarities, the Mayer/difference strategy of future treatments differs from conventional difference-in-difference (DiD), or gain score, estimation. While both approaches assume the same qualitative causal structure for the DGP, shown in Figure 2, they impose different parametric constraints on this structure. Mayer's approach interprets F as a future (post-outcome) value of the treatment and assumes that U affects T and F equally, $a = d$. By contrast, DiD interprets F as a lagged (pre-treatment) value of the outcome and assumes that U affects Y and F equally, $c = d$. As a result of these different constraints, the two methods lead to different estimators. As is easily verified against the graph, with F as future treatment, the spurious association between Y and T is measured and removed by the conditional covariance between Y and F given T. Hence, the Mayer/difference estimator is $b = b_{YT.F} - b_{YF.T}$. With F as lagged outcome, the spurious association between Y and T equals the marginal covariance between F and T, and the DiD estimator is $b = b_{YT} - b_{TF}$.

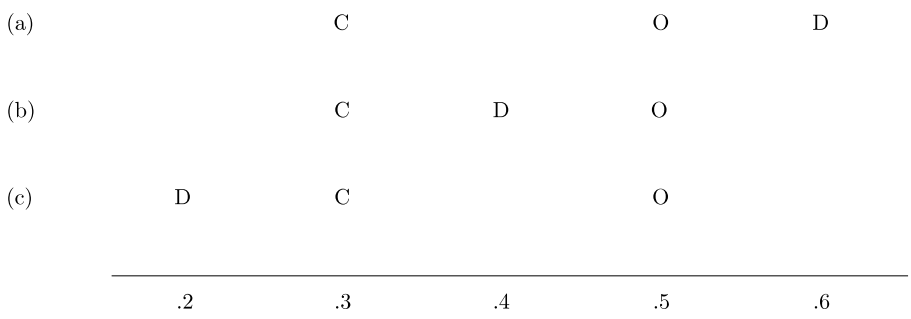
Choosing Between Future Treatment Estimators

Next, we compare the performance of the two future-treatment strategies and provide guidance for choosing between them. We continue to assume that the data are generated by the model in Figure 2.

Obviously, maximally cautious analysts should always prefer the control estimator, because, in contrast to the Mayer/difference estimator, it guarantees bias reduction when the data are produced by Figure 2, regardless of the relative size of the path parameters. Bias reduction with the control estimator, however, is often quite modest. For most values of the effect $U \rightarrow T$, a, the control estimator will remove less than half of the OLS bias unless the effect $U \rightarrow F$, d, is large, $d > 0.7$. In many cases, Mayer's estimator will thus remove more bias than the control estimator.

Analysts can sometimes decide between the two future treatment estimators by comparing the relative positions of the OLS, control, and Mayer/difference estimates.

Figure 4 illustrates the decision process. Since the control estimate, in expectation, is closer to the true treatment effect than the OLS estimate, the difference between the control and the OLS estimate reveals the direction of the OLS bias. For example, if the OLS estimate is $b_{OLS} = 0.5$ and the control estimate is $b_C = 0.3$, then the true treatment effect should be no larger than the control estimate, $b \leq 0.3$. The first decision rule thus states that if the control and Mayer/difference estimates change the OLS estimate in different directions (Figure 4-a), then the analyst should choose the control estimate as bias reducing and eschew the Mayer/difference estimate as bias increasing. Second, the control estimator is preferred as long as the Mayer/difference estimator does not differ more strongly from the OLS estimator in the same direction. For example, if the OLS estimate is $b_{OLS} = 0.5$, the control estimate is $b_C = 0.3$, and the Mayer/difference estimate is $b_M = 0.4$ (Figure 4-b), then the control estimate is preferred.



Notes: C = control estimate; D = Mayer/difference estimate; O = OLS estimate

Figure 4. Illustration of the heuristic for choosing between estimates. The difference between the control and OLS estimates indicates the direction of OLS bias in data generated by Figure 2. The relative position of control, difference, and OLS estimates can help the analyst decide between alternative estimates. In scenarios (a) and (b), the control estimate is preferred. In (c), additional assumptions are needed to decide between the control and difference estimates.

If the control and Mayer/difference estimators change the unadjusted OLS estimate in the same direction but the Mayer/difference estimator is farther away from the OLS estimate than is the control estimate (Figure 4-c), then it does not follow that the Mayer/difference estimator is automatically preferred. For example, with $b_{OLS} = 0.5$, $b_C = 0.3$, and $b_M = 0.2$, then the true effect could be closer to either the control estimate or the Mayer/difference estimate. Thus, the analyst would require additional knowledge about the relative size of effects $U \rightarrow T$, a, and $U \rightarrow F$, d, to decide between the control and Mayer/difference estimates. Two rules from Result 4, illustrated in Figure 3, help with this decision. First, if the analyst can argue that a and d share the same sign and that

the magnitude of a does not considerably exceed the magnitude of d, then the analyst should choose the Mayer/difference estimator because it will remove more bias than the control estimator. Second, if $|d| \gg |a|$ or if d and a have opposite signs, then the Mayer/difference estimator will increase the OLS bias, and the analyst should chose the control estimator.

CHALLENGES TO BIAS CORRECTION WITH FUTURE TREATMENTS

The DGP of Figure 2, analyzed so far, provides a best-case scenario for future-treatment strategies to reduce confounding bias in OLS regressions because it guarantees bias reduction for the control estimator and provides bias removal in the Mayer/difference estimator if $a = d$. In this section, we explain that both future-treatment strategies can increase bias in the presence of either (1) true state dependence, where past treatment causally affects future treatment, or (2) selection, where the outcome causally affects future treatment, or both. We demonstrate this failure by showing that both future-treatment strategies can produce bias even when the unadjusted OLS estimate is unconfounded and hence unbiased. With either true state dependence or selection, choosing the best future-treatment strategy becomes a matter of carefully weighing prior knowledge about the underlying path parameters in the DGP.

TRUE STATE DEPENDENCE: WHEN TREATMENT AFFECTS FUTURE TREATMENT

Past and future values of the treatment are typically correlated over time. One reason for this association could be mutual dependence of T and F on the unmeasured confounder U along the path $T \leftarrow U \rightarrow F$, as in Figure 2, which would justify the future-treatment strategies discussed above. Another reason for a correlation between T and F could be true state dependence, where past states of the treatment cause future states of the treatment (Bates and Neyman 1951; Heckman 1981a; Heckman 1981b).[8] True state dependence is captured by the arrow $T \rightarrow F$ in Figure 5. Sociologists are amply familiar with cumulative advantage and cumulative disadvantage as important special cases of true state dependence. DiPrete and Eirich (2006:272) explain that cumulative (dis)advantage "becomes part of an explanation for growing inequality when current levels of accumulation have a direct causal relationship on future levels of accumulation." For instance, individuals with higher incomes will be able to accumulate more financial assets that in turn generate asset income returns that grow at a higher rate than earnings (Piketty 2014). In this example, as in others, a causal story of true state dependence involves a causal mediator (here: income → financial asset acquisition →

[8] True state dependence is also a central challenge in the literature on dynamic treatment effects (Robins 1994; Wodtke and Almirall 2017).

income), and the strength of state dependence may be quite limited (e.g., for most people, asset income plays no appreciable role in determining their total annual income).

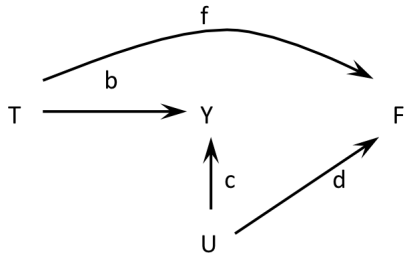


Figure 5. An unconfounded study with true state dependence of treatment, $T \rightarrow F$.

To build intuition for the problem of true state dependence, we first analyze the performance of future-treatment strategies when the effect of T on Y is not confounded (no arrow $U \rightarrow T$), as in Figure 5. Here, the marginal association between T and Y identifies the causal effect of T on Y, because the causal effect $T \rightarrow Y$ is the only open path between them. Hence, the unadjusted OLS estimate equals the true causal effect, $b_{OLS} = b_{YT} = b$, and the OLS estimator is unbiased.

Future-treatment strategies are vulnerable to true state dependence because needlessly controlling for F introduces bias. Since F is a collider variable on the noncausal path $T \rightarrow F \leftarrow U \rightarrow Y$, controlling for F opens this path and induces a spurious association between T and Y. Controlling for F in the regression of Y on T would therefore create bias where none existed before. This intuition is confirmed algebraically using Wright's rules. The control estimator for data generated by Figure 5, with true state dependence and without confounding, evaluates to

$$b_C = \hat{b}_{YT.F} = b - \frac{cdf}{(1-f^2)}\,. \tag{13}$$

Note that the control estimator in this scenario is biased even though the OLS estimator is not. As expected, the bias in the control estimator under true state dependence is a function of the path parameters on the noncausal path $T \rightarrow F \leftarrow U \rightarrow Y$; $f, d, and\ c$. The bias in b_C increases with the strength of confounding between Y and F, cd, in the numerator of the bias; and the bias increases especially strongly with the strength of state dependence, f, which increases the numerator and decreases the denominator of the bias.

The Mayer/difference estimator for data generated by Figure 5, with true state dependence and without confounding, evaluates to

$$b_{M/D} = b - \frac{cd(f-1)}{(1-f^2)}\,. \tag{14}$$

Note that the Mayer/difference estimator is also biased in this scenario, even though the OLS estimator is not. Comparing expressions [13] and [14] shows that true state dependence introduces less bias into the control strategy estimator than into the Mayer/difference estimator, unless true state dependence is strongly positive, $f > 0.5$. In

sum, both future treatment estimators can increase bias under true state dependence, but the control estimator will be less biased as long as true state dependence is not too large.

Next, we analyze the empirically more interesting DGP of Figure 6, which combines Figure 2 with Figure 5 to form a scenario of true state dependence with unobserved confounding. Here, U is a confounder of T and F, which motivates the use of F as a proxy control to reduce bias in the OLS estimator, but T also directly causes F via true state dependence, thus introducing bias into both future treatment estimators. Without further restrictions, the analytic expressions for the control and difference estimators are unwieldy and scarcely informative (not shown). Depending on the exact parameter constellation, both future-treatment strategies could reduce bias or increase bias in the OLS estimator. Hence, analysts must carefully consider existence, direction, and size of true state dependence in their empirical applications.

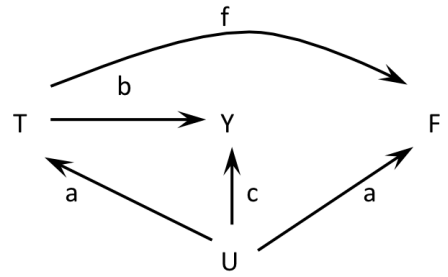


Figure 6. A confounded study with true state dependence of treatment (combination of Figures 2 and 5)

Nonetheless, future-treatment strategies remain promising if the analyst can defend certain parametric restrictions on the relative size of the path parameters. Consider, for example, the restriction that U affects T to the same extent as it affects F, $a = d$, as Mayer (1997) proposed for the effect of parent income on child outcomes.

Result 6 (bias of the control estimator with true state dependence): In data generated by the model in Figure 6 with the constraint $a = d$, the control estimator evaluates to

$$b_C = b + ac\frac{-f-f^2-a^2f-a^2+1}{1-(f+a^2)^2} = b + B_{OLS}R_C \ . \qquad (15)$$

Result 7 (bias of the Mayer/difference estimator with true state dependence): In data generated by the model in Figure 6 with the constraint $a = d$, the Mayer/difference estimator evaluates to

$$b_{M/D} = b + ac\frac{-f-f^2-a^2f}{1-(f+a^2)^2} = b + B_{OLS}R_M \ . \qquad (16)$$

The bias multipliers of the control and Mayer/difference estimators, R_C and R_M, are obviously closely related, though their behavior is somewhat surprising. Simulations (not

shown) reveal several facts, summarized in Table 1:

Result 8 (relative performance of the control and Mayer/difference estimators under confounding and true state dependence): In data generated by the model in Figure 6 with the constraint $a = d$, the following five facts hold:

With (unrealistic) negative state dependence, $f < 0$,

(1) The Mayer/difference estimator is strictly bias reducing, $0 < R_M < 1$, and strictly dominates the performance of the control estimator, $R_M < R_C$.

(2) The control estimator becomes increasingly bias amplifying as true state dependence becomes increasingly negative.

With (realistic) positive state dependence, $f > 0$,

(3) The Mayer/difference estimator reduces bias less than the control estimator, $|R_C| < |R_M|$, under moderately strong state dependence, $0.37 \lesssim f \lesssim 0.5$,

(4) The Mayer/difference estimator is strictly bias amplifying, $|R_M| > 1$, under strong positive state dependence, $f \gtrsim 0.5$.

(5) The control estimator is strictly bias reducing up to moderately strong positive state dependence, $0 < f \lesssim .06$, and moderate confounding ($|a| \lesssim 0.5$).

State Dependence (f)	Bias with Control Estimator	Bias with Mayer/Difference Estimator
Negative ($f < 0$)	Amplified	Reduced
Positive, moderate ($0.37 \lesssim f \lesssim 0.5$)	Reduced	Weakly reduced
Positive, strong ($f \gtrsim 0.5$)	Reduced *(up to $f \lesssim .06$ & $\|a\| \lesssim 0.5$)*	Strictly amplified

Table 1. Performance of the control estimator and the Mayer/difference estimator in the presence of state dependence and assuming $a = d$.

Table 1 underlines that true state dependence, which is a common concern in sociology, ruins the strict bias-reduction property of the control estimator. Nonetheless, under realistic values of mild positive true state dependence, both the control and the Mayer/difference estimators are bias reducing. For weak positive true state dependence, the Mayer/difference estimator removes more bias than the control estimator; and for

moderate and strong positive state dependence, the control estimator outperforms the Mayer/difference estimator and remains (strongly) bias reducing as long as the effects of U on T and F are not too large.

SELECTION BIAS: WHEN THE OUTCOME AFFECTS FUTURE TREATMENT

Selection also complicates future-treatment strategies for unobserved confounding. We say that selection is present when the outcome exerts a causal effect on the future value of the treatment, as captured by the arrow $Y \rightarrow F$ in Figure 7. Selection is a concern in many situations. For example, in a study of the effect of parental income on educational attainment, college enrollment may affect parents' income if parents adjust their labor supply to the financial needs of the child. In other scenarios, selection may be absent. For example, when studying the effect of parental income on children's test scores, it is implausible to believe that children's test scores affect future values of parental income (except, perhaps, when a child's abysmal test scores inspire a parent to quit her job to tutor the child).

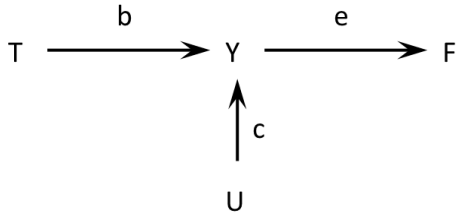


Figure 7. An unconfounded study with selection, $Y \rightarrow F$.

Figure 7 isolates the problems of selection. By assuming that the effect of T on Y is unconfounded. In this scenario, the unadjusted OLS estimator again recovers the true causal effect, $b_{YT} = b$: Since the unadjusted OLS estimator does not involve F, OLS does not suffer from selection bias. The control and difference estimators, however, do involve F and hence suffer selection bias, because controlling for F amounts to selecting on the outcome.[9] Algebraic derivation shows that the control strategy estimator without confounding but with selection evaluates to

$$b_C = b_{YT.F} = b\frac{(1-e^2)}{1-e^2b^2} = bP_C \ , \tag{17}$$

and the difference strategy estimator evaluates to

$$b_{M/D} = b\frac{b-e}{b-b^2e} = bP_D \ . \tag{18}$$

[9] Figure 7 presents an example of post-outcome endogenous selection bias (Elwert and Winship 2014). In the language of DAGs, Y is a collider variable on the path $T \rightarrow Y \leftarrow U$, and F is a descendant of Y. Conditioning on a descendant of a collider induces an association between the collider's immediate causes, i.e. between T and U. Hence, conditioning on F induces a non-causal association between T and Y via U, which is the bias in the F-adjusted analysis.

Since $P_C \neq 1$ and $P_D \neq 1$ are pure bias terms, neither the control estimator nor the difference estimator recovers the true causal effect. It can be shown, however, that $|P_C| \ll |P_D|$; that is, selection (without confounding) introduces far less bias into the control estimator than into the Mayer/difference estimator, especially for small treatment effects $T \rightarrow Y$. Note that bias in the control and Mayer/difference estimators with selection depends on the size of the treatment effect, b.

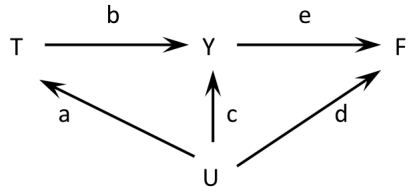


Figure 8. A confounded study with selection, $Y \rightarrow F$.

Finally, Figure 8 shows the empirically important scenario with both selection and confounding (combining Figures 7 and 2, respectively). The corresponding analytic expressions for the control and Mayer/difference estimators are highly non-linear.

Result 9 (bias in the control estimator with selection): In data generated by Figure 8, the control estimator evaluates to

$$b_C = b_{YT.F} = \frac{(b + ac) - (bad + e + cd)\,(be + ace + ad)}{1 - (be + ace + ad)^2} = b + B_{OLS}S_C\ , \qquad (19)$$

Result 10 (bias in the Mayer/difference estimator with selection): In data generated by Figure 8, the Mayer/difference estimator evaluates to

$$b_D = \frac{(b + a\,c) - (b\,a\,d + e + c\,d)}{1 - (b\,e + a\,c\,e + a\,d)} = b + B_{OLS}S_D\ . \qquad (20)$$

The bias-reduction properties of both future treatment estimators with confounding and selection strongly depend on the underlying path parameters. Simulations (not shown) suggest that the Mayer/difference estimator is usually performing worse, and often dramatically so, than the control estimator as long as the path parameters, p, are not too large, $|p| < 0.5$. Specifically, any hint of selection, $e \neq 0$, threatens to turn the Mayer/difference estimator into a bias amplifier. By contrast, as long as selection is mild, $e \lesssim 0.3$, the control estimator remains bias reducing, though bias reduction can be small in absolute terms.[10]

Table 2 summarizes the divergent performance of the control and the Mayer/different estimator. The upshot is that for scenarios in which path parameters are at most moderately strong (≤ 0.5), the control strategy generally carries the day. Since the

[10] When the control estimator increases bias, it does so negligibly.

control strategy without selection is strictly bias reducing and only minimally biased by selection, it tends to remove some bias overall. By contrast, the Mayer/difference strategy is strongly bias reducing without selection, but can induce heavy bias with selection, and so it is to be used with caution.

Selection (e)	Bias with Control Estimator	Bias with Mayer/Difference Estimator
Selection ($e \neq 0$)	Negligibly amplified or weakly reduced (see below)	Mostly amplified
Mild Selection ($e \lesssim 0.3$)	Weakly reduced	Mostly amplified

Table 2. Performance of the control estimator and the Mayer/difference estimator in the presence of selection and with $p < 0.5$

A FUTURE-TREATMENTS TEST FOR UNOBSERVED CONFOUNDING

Importantly, beyond bias reduction and bias removal, we can also employ future treatments to test for the absence of unobserved confounding in the causal effect of T on Y. T and Y are unconfounded if no cause of T also causes Y. The Null hypothesis of no confounding is thus encoded by the absence of the arrow $U \rightarrow Y$ in Figure 9a.

The test is straightforward. If we assume, as we have before,

A1: All factors U that affect T also affect F,

then conditional independence between F and Y given T, $Y \perp F|T$, implies the absence of unobserved confounding between T and Y.

If, furthermore, we assume

A2: Y does not cause F (no selection),

and

A3: F and Y share no common causes outside, possibly, of U,

then non-independence between F and Y given T, $Y \not\perp F|T$, implies confounding between T and Y. (Proofs follow directly from d-separation [Pearl 2009].)

Figure 9 illustrates the logic of the test, and the importance of the three assumptions A1-A3 for various DGPs.

We note that rejecting the Null of no unobserved confounding in the causal effect $T \rightarrow Y$ in favor of the alternative of unobserved confounding requires weaker

assumptions than removing the bias. Specifically, the test of no-confounding does not require (a) the absence of true state dependence, $T \rightarrow F$ or (b) linearity, or, for that matter, any parametric assumptions. Hence, we can test for unobserved confounding even if we cannot identify and consistently estimate the causal effect of T on Y using the control or Mayer/difference strategies, as in Figure 9.

Under assumptions A1-A3, any non-parametric or parametric test for conditional independence between F and Y given T is a valid test of no confounding. For example, in a linear model, testing the T-adjusted regression coefficient H_0: $b_{YF.T} = 0$, against H_A: $b_{YF.T} \neq 0$, using a conventional two-sided t-test, is a valid test of the null hypothesis of no confounding.[11]

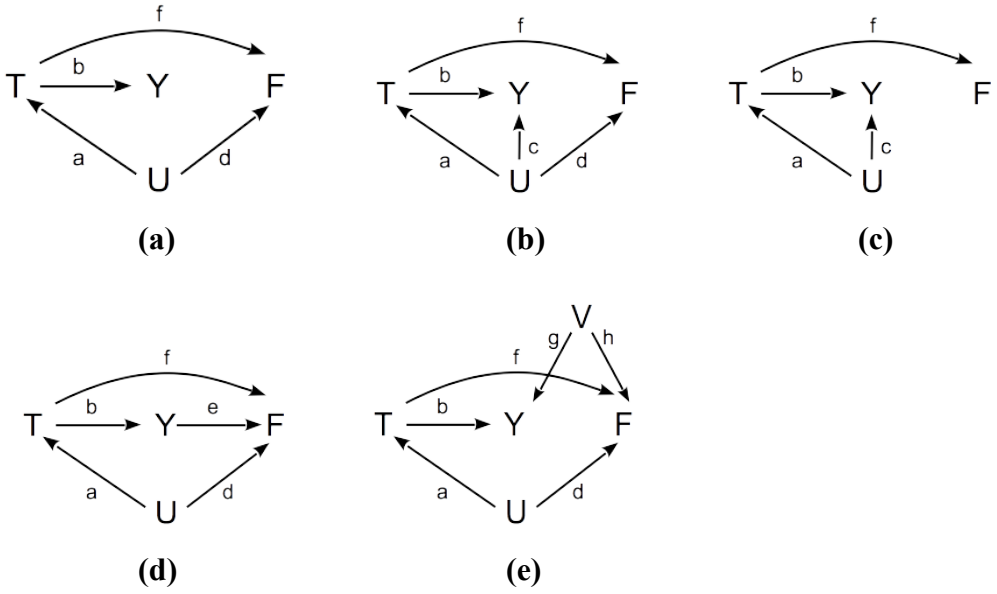


Figure 9. (a) The absence of the arrow $U \rightarrow Y$ encodes the Null hypothesis of no-confounding between T and Y. Since A1 is met, conditional independence between Y and F given T implies the absence of confounding between T and Y. (b) Since A1-A3 are met, a conditional association between Y and F given T implies the existence of the arrow $U \rightarrow Y$, i.e. confounding between T and Y. (c) Assumption A1 is violated by the absence of the arrow $U \rightarrow F$. The conditional independence between Y and F given T does not imply the absence of confounding between T and Y. (d) Assumptions A2 is violated by the existence of the arrow $Y \rightarrow F$. The conditional association between Y and F given T does not imply confounding between T and Y. (e) Assumption A3 is violated by the existence of the arrows $Y \leftarrow V \rightarrow F$. The conditional association between Y and F given T does not imply confounding between T and Y.

[11] Mayer (1997) originally suggested this test for the linear DGP of Figure 2. Here, we add, first, that the logic of this test generalizes non-parametrically, i.e. is valid for all functional forms, and, second, that it also holds for DGPs other than Figure 2, as long as assumptions A1-A3 hold.

In sum, even where future treatment strategies may fail in reducing unobserved bias, a simple, non-parametric test for the absence of unobserved bias is available and holds promise across a wide field of applications.

EMPIRICAL ILLUSTRATION

MOTIVATION

We illustrate the utility of future-treatment strategies by elaborating on Mayer's (1997) original empirical example of the effect of parental income on children's educational attainment. More than merely of historical interest, the example remains salient for contemporary debates on intergenerational transmission, which are plagued by concerns about unobserved confounding (Sobel 1998; Morgan and Winship 2015).[12]

Suppose that an analyst wants to know whether an observed association between parents' income and the educational attainment of their children suffers from unobserved confounding bias. An argument in favor of a causal effect of parental income could be made as follows (see also Mayer 1997: 45ff): High income allows parents to make higher monetary investments in their children's education (Kornrich and Furstenberg 2012; Schneider et al. 2018), for instance by providing private tutors (Buchmann et al. 2010), which ultimately leads them to educational success. On the other hand, the association between parental income and children's educational success could also be due to unobserved confounding. Parents' ability, attitudes, and behaviors could determine not only their own income but also directly influence the educational success of their children.

DATA

We analyze data from the Panel Study of Income Dynamics (PSID). In an effort to replicate the estimates provided by Mayer (1997), we closely follow her decisions in the construction of the analytic samples (covering birth cohorts 1954 through 1968) and variables as described there (in particular, ibid: 161ff). The main outcome of interest, Y, is children's years of education completed by age 24 ($mean = 12.9$, $sd = 2.0$; see also descriptive statistics in Appendix Table C.1). The treatment variable, T, is logged family income in 1992 dollars, measured at children's ages 13 through 17 (5-year averages). The future treatment variable, F, is logged family income in 1992 dollars, measured at children's ages 25 through 29 (5-year averages). The list of observed confounders, X, includes logged family size, whether the child's household head is black, parental age of the younger parent, the highest years of education attained by either parent, and whether

[12] In fact, at the time of this writing, an ambitious randomized control trial is about to be implemented to directly adjudicate the causal effects of parental income on child development (Duncan et al. 2017).

the child is male. The analytic sample size for the future treatment regressions is N=1,513. All analyses are weighted based on the child's individual survey weight in 1989. All variables are standardized (mean zero and variance one). A replication package containing the data and code used for this analysis is available online at [HIDDEN].

RESULTS

Our analyses replicate Mayer's published results almost perfectly. For instance, in her main analyses (based on cohorts born between 1954 and 1968), Mayer estimates the unstandardized coefficient of logged family income on children's years of education to be 0.78 ($se = 0.07$), compared to our estimate of 0.76 ($se = 0.07$). We observe an even closer correspondence in the analytic subsample (for which future income measures are available; birth cohorts 1954-1964) with a standardized coefficient estimate of 0.19 in both hers and our analysis (for full results see also Appendix Table C.2).

Our empirical illustration of future treatment strategies applies OLS regression models to the same data, predicting the outcome (Y, years of education) based on different combinations of the regressors of interest: the treatment (T, parental income), the future treatment (F, future parental income), and all observed control variables mentioned above (X). Table 3 reports four different model specifications that we draw on to demonstrate the use of different future-treatment strategies under various assumptions about the DGP. Model 1 displays the unadjusted association between parental income (T) and offspring's educational attainment (Y). Without controlling for any observed confounders (X), we expect the association between T and Y to provide a biased estimate of the causal effect. For illustration purposes, it is helpful to first apply future-treatment strategies to a treatment-effect estimate that we know to be biased. Therefore, in model 2, we add the future treatment (F), but no further controls, to the model. The comparison between model 1 and model 2 will thus be used to show how future-treatment strategies produce expected answers in a situation of bias. The more realistic scenario encountered in empirical applications, of course, is that the analyst has already exhausted the options to adjust for observable differences, reflected in model 3, which includes all control variables used in the original analyses by Mayer (1997). In model 4, we then add a control for the future treatment to illustrate the conclusions drawn from future-treatment strategies in the typical empirical setting without prior knowledge about the existence and direction of unobserved bias.

We contrast the conclusions drawn based on the control strategy, the Mayer/difference strategy, and our non-parametric test under various assumptions about the DGPs.

Table 3: Estimating the causal effect of parental income on children's years of education with and without future treatments

Standardized OLS regression coefficients (standard errors in parentheses); weighted

	(1)	(2)	(3)	(4)
Coefficients				
T: Parental Income	0.448 ***	0.319 ***	0.185 ***	0.118 **
	(0.037)	(0.039)	(0.038)	(0.039)
F: Future Parental Income		0.274 ***		0.202 ***
		(0.031)		(0.029)
X: Controls			Yes	Yes
Difference in coefficients				
T minus F		0.045		-0.084
		(0.059)		(0.054)
Equivalence of T coefficients: p-values				
Model (1) vs. (2): 0.0006				
Model (1) vs. (3): 0.0000				
Model (3) vs. (4): 0.0006				
Model (1) vs. (4): 0.0000				
N	1,513	1,513	1,513	1,513

Statistical significance at + p<.10, * p<.05, ** p<.01, and *** p<.001 (two-tailed test)

BEST CASE SCENARIO

We begin by assuming what we have described as the best-case scenario, the DGP of Figure 2. This scenario starts from the central assumption that all confounders of T also affect F. Furthermore, it assumes the absence of true state dependence (no arrow $T \rightarrow Y$) – that is, changes in parental income during middle childhood (ages 13-17) have no causal impact on parental income during offspring's young adulthood (ages 24-29).

Finally, it assumes the absence of selection (no arrow $Y \rightarrow F$) – that is, children's years of education do not cause changes in their parents' income.

We begin with the test for absence of unobserved bias in the unadjusted association between parental income, T, and child's educational attainment, Y. Model 1 gives this unadjusted association as $b_{YT} = 0.448$ ($p < .001$). We have argued that the test of no unobserved confounding amounts to testing the null hypothesis that the T-adjusted regression coefficient on F is zero, $b_{YF.T} = 0$. In model 2, this null hypothesis is safely rejected ($p < .001$). Hence, we conclude that the naïve, unadjusted, estimate of model 1 suffers from unobserved bias, which appears plausible.

Now that we believe in the existence of bias, we use the control strategy to reduce it. The control strategy calls for a focus on the F-adjusted treatment effect, estimated as $b_{YT.F} = 0.319$ ($p < .001$) in model 2, which is significantly less than the unadjusted association between parental income and child's educational attainment in model 1 ($b_{YT} - b_{YT.F} = 0.448 - 0.319 = 0.129$ [$p < .001$]). By Result 2, we know that the control strategy is strictly bias reducing under the DGP of Figure 2. Thus, we conclude that the F-adjusted estimate from model 2 is closer to the true treatment effect than the naïve estimate without F-adjustment of model 1; the naïve treatment effect estimated in model 1 is upwardly biased.

The Mayer/difference method, applied to model 2, estimates the treatment effect as the difference between the partial coefficients on T and F, $b_{YT.F} - b_{YF.T} = 0.319 - 0.274 = 0.045$ ($p = 0.445$). We note that this estimate is again lower than the naïve estimate of the treatment effect ($b_{YT} = 0.448$) and also lower than the control estimate ($b_{YT.F} = 0.319$). Earlier, we showed that the Mayer/difference estimate is potentially more powerful in reducing bias than the control strategy, but that – unlike the control strategy – it may also amplify bias.

From the application of the control strategy we learned that the naïve estimate of model 1 is upwardly biased. If the Mayer/difference estimator had yielded a higher estimated treatment effect than the naïve estimate (cf. Figure 4a), we would have concluded that the Mayer/difference strategy amplifies rather than reduces existing bias. If, by contrast, the Mayer/difference estimator had fallen between the naïve and the F-adjusted estimate of the treatment effect (Figure 4b), we would have concluded that the difference strategy is less effective in reducing bias than the control strategy. In both instances, we would have preferred the control estimate to the Mayer/difference estimate.

In the case of our specific application, however, the Mayer/difference estimate corrects the naïve estimate in the same direction as, but more strongly than, the control estimate (Figure 4c). Yet, without further assumptions about the strength of the path parameters, we do not know whether the Mayer/difference estimate is closer to the true causal effect than the control estimate. The most conservative analyst may therefore prefer the estimate provided by the control strategy in this empirical application noting, however, that bias reduction may be relatively modest unless the effect $U \rightarrow F$ is very

large. In some applications, the analyst may have reasonable expectations about the direction and sign of the effects $U \rightarrow T$, a, and $U \rightarrow F$, d. In our example, an analyst may assume that the multifold characteristics of parents that are not controlled for in these models, U, impact parental income in the same direction at T and F, i.e., a and d have the same sign. Then, unless the effect $U \rightarrow F$, d, is much larger than the effect $U \rightarrow T$, a, the analyst should prefer the Mayer/difference estimator as the strategy to reduce the most bias. In sum, in this empirical example, the decision between the control and the difference estimator depends on how defensible the analyst's additional assumptions about the relative size of the effects $U \rightarrow T$ and $U \rightarrow F$ are. If the analyst prefers the Mayer/difference strategy estimates, then one should note that this estimate is not statistically different from zero ($p = 0.445$). This would cast doubt on the proposition that an increase in parental income causes any improvement in children's educational attainment.

Next, we turn to the more common scenario in which we estimate a treatment effect controlling for observables (model 3). This covariate-adjusted model estimates the treatment effect as $b_{YT.X} = 0.185$ ($p < .001$). This estimate is much lower ($p < .001$) than the unadjusted association of model 1 and indicates that the reduced control strategy estimate of model 2 correctly determined the positive direction of the bias.

Although, in model 3, we draw on a number of important control variables, the critical analyst will still worry about unobserved bias. These concerns are addressed in the future-adjusted model 4. Again, the null hypothesis of no unobserved bias cannot be rejected since the coefficient on F, $b_{YF.TX} = 0.202$ is significantly different from zero ($p < .001$). Confirming the presence of unobserved bias certainly provides license to apply future treatment adjustments that may reduce bias.[13]

The control estimate of model 4 is again lower than the baseline treatment effect of model 3 ($b_{YT.FX} = 0.118$ vs. $b_{YT.X} = 0.185$ [$p < 0.001$]), again indicating that the former corrects for remaining upward bias in the latter. The correction is more modest than before but statistically significant ($b_{YT.X} - b_{YT.FX} = 0.185 - 0.118 = 0.067$ [$p < .001$]). Between models 1 and 2, the correction was larger – about twice the size – since there we put a greater burden on the future treatment control to reduce bias in the absence of any control variables. By contrast, in model 3, there should be less unobserved bias to control for. The Mayer/difference method estimates the treatment effect to be yet smaller and even negative, at $b_{YT.FX} - b_{YF.TX} = -0.084$ ($p = 0.119$), though statistically

[13] If our test had not detected bias (i.e. if the estimate of $b_{YF.TX}$ on F was indistinguishable from zero), bias could still be present: Failure to detect bias is different from confirming the absence of bias. A conservative analyst would then be yet more sensitive to the possibility of bias amplification in a situation that potentially does not suffer from existing bias.

indistinguishable from zero at customary levels of statistical significance.[14] As before, absent additional assumptions about the relative strength of the effects $U \rightarrow T$ and $U \rightarrow F$, we cannot be certain that the Mayer/difference estimate is closer to the true treatment effect than the control estimate.

For a final step of this illustration, let us put aside the concerns about remaining unobserved bias and instead accept the control estimate of model 4 ($b_{YT.FX} = 0.118$) as the true treatment effect.[15] In this scenario, we would accept and expand some of the conclusions drawn from model 2. The model without adjustments for observables (model 1) is upwardly biased (0.448 vs. 0.118 [$p < 0.001$]), as previously concluded based on the control estimate from model 2 (0.319), which removed some but not all upward bias in the treatment effect of model 1. The Mayer/difference method of model 2 eliminated all upward bias but introduced downward bias (providing an estimate lower than the hypothetical true treatment effect). Overall, though, the Mayer/difference estimator of model 2 is closer to the assumed true treatment effect ($|0.045 - 0.118| = 0.073$) than the control estimator ($|0.319 - 0.118| = 0.201$) and thus preferred in the setting of a model without further observed controls. The unlikely conditions needed for either future-treatment strategy to fully eliminate bias ($d = 1$ for the control method, and $a = d$ for the difference method) do not hold in this application.

TRUE STATE DEPENDENCE AND SELECTION

Next, we revisit the interpretation of the results presented in Table 3 under different assumptions about the DGP – true state dependence and selection – which we have shown to pose challenges to future-treatment strategies.

As mentioned, sociologists are well accustomed to cumulative advantage arguments. In our empirical example, one may suspect that parents' income growth depends on their baseline income. One scenario of such true state dependence was discussed above as higher income enabling access to financial assets and their returns. Even under true state dependence, however, the test for the absence of unobserved bias is valid. The conclusions remain the same as those discussed above: We would rule out that the treatment effect estimate is unbiased in all models shown. With state dependence,

[14] Interestingly, while we successfully replicate Mayer's main estimates, our estimate of the Mayer/difference estimator is quite different. Her analyses suggests a quite modest drop from 0.186 in model 3 to 0.168 in model 4 (ibid: pp. 92), ours show a much larger drop from 0.185 to a statistically insignificant estimate of -0.084. The conclusions that may be drawn from our estimate –no causal relationship between parental income and children's educational attainment – are in fact more supportive of the general conclusions drawn by Mayer (1997).

[15] In this example, and conditional on the assumed DGP, the substantive interpretation of this estimate (unstandardized effect size of 0.4282) is that an increase in parental income by 10% leads to an increase in children's educational attainment by about half a month ($0.4282 * log(1.1) * 12 = 0.49$).

however, using the control or Mayer/difference estimator to reduce this bias in model 1 requires new assumptions about the size of certain path parameters, because now even the control estimator is not strictly bias reducing anymore. Most important, this includes assumptions about confounding itself. For example, we could assume that the effects $U \rightarrow T$ and $U \rightarrow F$ are of the same size ($a = d$) and that confounding is of, at most, moderate size ($|a| \lesssim 0.5$) –the latter assumption is less credible for model 1. Second, our choice between the control and Mayer/difference estimator is dictated by assumptions about the direction and degree of state dependence. That is, if state dependence is negative or at best weakly positive, the Mayer/difference estimator is the better choice. However, if state dependence is moderately ($f \gtrsim 0.37$) or strongly positive ($f \gtrsim 0.5$), the control estimator is the better choice. The stakes involved in making these assumptions are quite high. If they are wrong, future-treatment strategies may amplify bias (namely, the control estimator if state dependence is negative and the difference estimator if state dependence is strongly positive). Existing empirical work on the dynamics of income poverty (in essence, a dummy variable version of our treatment variable) suggest state dependence to be positive and large (e.g. Cappellari and Jenkins 2004, Biewen 2009), which would lead one to prefer the control estimator.

In our empirical example, selection, $Y \rightarrow F$, may be of concern, for instance if the children's decision not to enroll in college and instead begin work may cause parents to reduce their labor supply as the need for intergenerational transfers declines. While such selection story may be plausible for certain subgroups of the population, we are not aware of well-identified estimates of large selection effects. Still, what does the worry about selection imply for the utility of different future-treatment strategies for detecting and removing bias? Unfortunately, the test for the absence of unobserved bias is no longer valid under selection. The attractiveness of the difference method is drastically reduced as its bias-amplification property becomes more pronounced. Those who assume selection to be a concern in our empirical example should refrain from both an interpretation of the test and the difference estimator. However, the control estimator would remain useful because if there is confounding, and if selection is mild, then the control estimator remains bias-reducing. Hence, the control estimator would remain the preferred estimate.

CONCLUSION

The problem of unobserved confounding is profound. Most research in the social sciences is observational and observational studies cannot rule out bias from unobserved confounding. The direction and especially the size of the bias are often difficult to gauge, in part because the bias could originate in confounders that are as yet unknown to science.

In this paper, we have discussed future values of the treatment variable as a tool

for detecting, reducing, and removing bias from unobserved confounding. Future treatments have occasionally been used for bias removal in prior research. Here, we have subjected several easily computed future-treatment strategies to a detailed analysis, introduced a new strategy, and compared the relative strengths and weaknesses of these estimators to each other and to baseline conventional regression estimates. While we identify challenges to future-treatment strategies, we do not stop there. To maximize the usefulness of future treatment estimators in applied research, we also demonstrate how additional assumptions about effect sizes can help choose between them and inform their interpretation.

The idea behind future-treatment strategies is intuitive: any variable that affects the treatment variable before the outcome likely also affects it after the outcome has been measured. In other words, future treatments can proxy for unobserved confounding. We have used this insight directly and proposed controlling for future treatments using what we have termed the control estimator. This estimator has the great advantage of being strictly bias reducing for some linear data generating processes.

Analyzing important prior future-treatment strategies, we have noted that Mayer's (1997) estimator is not strictly bias reducing even in the best-case scenario, and may in fact amplify OLS bias. The same is true of Gottschalk's (1996) future treatment estimator (Appendix A). Nonetheless, Mayer's estimator holds promise because, in certain situations, it reduces bias more than the control estimator.

Future-treatment strategies have several advantages over other strategies for dealing with unobserved confounding. One advantage lies in the ready availability of future treatment measures in most panel data. Another is the ease of implementation – including future treatments as control variables in a conventional regression analysis. In contrast to fixed-effects estimation, future-treatment strategies to reduce unobserved bias do not require repeated measures of the outcome, nor do they require long panels (three periods suffice; see also Vaisey and Miles 2017 for a critical discussion of fixed-effects estimation based on three observation points). As such, several recent survey innovations provide attractive data for the application of the future-treatment strategy, e.g., the recent longitudinal extension of the General Social Survey (GSS) to three-wave panels (Hout 2017) or the newly redesigned Survey of Income and Program Participation (SIPP) as a four-wave panel. Finally, future-treatment strategies can be used for the dual purpose of detecting and reducing—sometimes even removing—unobserved confounding. Indeed, we have shown that future treatments can detect the presence of bias even in situations in which they cannot reduce this bias, and without any parametric assumptions.

A limitation shared with all strategies to remove bias from unobserved confounding is that causal treatment effect estimation based on observational data requires detailed knowledge of the data-generating process. We have highlighted two conditions that pose particular challenges for future treatment estimators: true state dependence (when prior treatment causally affects future treatment) and selection (when

the outcome causally affects future treatment). In both scenarios, all future treatment estimators can increase rather than decrease bias in OLS estimates. And whereas selection may be ruled out in many substantive applications, true state dependence often remains a credible threat. Based on our analytic results, however, we have argued that the control estimator remains bias reducing for moderate confounding under moderate true state dependence, and is surprisingly robust to selection as well.

Since future-treatment strategies make different demands on the data-generating process than fixed-effects or instrumental variables estimators, and because measures of future treatment measures are widely available in panel data, future-treatment strategies promise help where other popular strategies fail.

REFERENCES

Bates and Neyman. 1951. Contributions to the Theory of Accident Proneness. An Optimistic Model of the Correlation Between Light and Severe Accidents. *University of California Publications in Statistics* 1(9): 215–54.

Biewen, Martin. 2009. Measuring State Dependence in Individual Poverty Histories When There Is Feedback to Employment Status and Household Composition. *Journal of Applied Econometrics* 24(7):1095–1116.

Brito, Carlos, and Judea Pearl. 2002. Generalized Instrumental Variables. *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence* 85–93.

Buchmann, Claudia, Dennis Condron, and Vincent Roscigno. 2010. Shadow Education, American Style. Test Preparation, the SAT and College Enrollment. *Social Forces* 89(2):435–62.

Cappellari, Lorenzo and Stephen P. Jenkins. 2004. Modelling Low Income Transitions. *Journal of Applied Econometrics* 19(5):593–610.

Chamberlain, Gary. 1982. Multivariate Regression Models For Panel Data. *Journal of Econometrics* 18(1): 5–46.

Chan, Hei, and Manabu Kuroki. 2010. Using Descendants as Instrumental Variables for the Identification of Direct Causal Effects in Linear SEMs. *International Conference on Artificial Intelligence and Statistics* 73–80.

Deluca, Stefanie. 2012 What Is the Role of Housing Policy? Considering Choice and Social Science Evidence. *Journal of Urban Affairs* 34(1):21–28.

DiNardo, John, and Jorn-Steffen Pischke. 1997. The Returns to Computer Use Revisited: Have Pencils Changed the Wage Structure Too? *Quarterly Journal of Economics* 112(1): 291–303.

DiPrete, Thomas A, and Gregory M Eirich. 2006. Cumulative Advantage as a Mechanism for Inequality: A Review of Theoretical and Empirical Developments. *Annual Review of Sociology* 32: 271–97.

Duncan, Greg J, James P Connell, and Pamela K Klebanov. 1997. Conceptual and Methodological Issues in Estimating Causal Effects of Neighborhoods and Family Conditions on Individual Development. In: *Neighborhood Poverty, Volume 1: Context and Consequences for Children*: 219–50.

Duncan, Greg J. 2017. Household Income and Child Development in the First Three Years of Life. NIH Grant 1R01HD087384-01A1.

Elwert, Felix, and Nicholas A. Christakis. 2008. Wives and Ex-Wives: A New Test for Homogamy Bias in the Widowhood Effect. *Demography* 45(4): 851–73.

Elwert, Felix. 2013. Graphical Causal Models. *Handbook of Causal Analysis for Social Research*: 245–73.

Elwert, Felix and Christopher Winship. 2014. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology* 40(1):31–53.

Gottschalk, Peter. 1996. Is the Correlation in Welfare Participation across Generations Spurious? *Journal of Public Economics* 63: 1–25.

Grogger, J. 1995. The Effect of Arrests on the Employment and Earnings of Young Men. *The Quarterly Journal of Economics* 110(1):51–71.

Heckman, James J. 1981a. Statistical Models for Discrete Panel Data. In *Structural Analysis of Discrete Data and Econometric Applications*, edited by Charles F. Manski and Daniel L. McFadden, Cambridge: MIT Press, 114–78.

Heckman, James J. 1981b. Heterogeneity and State Dependence. Pp. 91–140 in Studies in *Labor Markets*, edited by S. Rosen. Chicago: University of Chicago Press.
Hout, Michael. 2017. Models for Three-Wave Panel Data: Examples Using the General Social Survey Panels. *Sociological Methods & Research* 46(1):41–43.
Kim, Jerry W., Bruce Kogut, and Jae-Suk Yang. 2015. "Executive Compensation, Fat Cats, and Best Athletes*." American Sociological Review* 80(2):299–328.
Kornrich, Sabino and Frank Furstenberg. 2012. Investing in Children. Changes in Parental Spending on Children, 1972-2007. *Demography* 50(1):1–23.
Mayer, Susan E. 1997. What Money Can't Buy. Family Income and Children's Life Chances. Cambridge: Harvard University Press.
Morgan, Stephen L, and Christopher, Winship. 2015. Counterfactuals and Causal Inference. Methods and Principles for Social Research. Second Edition. Cambridge: Cambridge University Press.
Pearl, Judea. 2009. Causality: Models, Reasoning, and Inference. Second Edition. Cambridge: Cambridge University Press.
Pearl, Judea. 2013. Linear Models: A Useful 'Microscope' for Causal Analysis. *Journal of Causal Inference* 1(1): 155–170.
Piketty, Thomas. 2014. *Capital in the Twenty-First Century*. Cambridge: Belknap Press.
Porter, Lauren C., and Ryan D. King. 2014. Absent Fathers or Absent Variables? A New Look at Paternal Incarceration and Delinquency. *Journal of Research in Crime and Delinquency* 52 (3): 414–43.
Robins, James M. 1994. "Correcting for Non-Compliance in Randomized Trials Using Structural Nested Mean Models." *Communications in Statistics-Theory and Methods* 23(8):2379–2412.
Rosenbaum, Paul R. 2002. Observational Studies. Second Edition. New York: Springer.
Schneider, Daniel, Orestes P. Hastings, and Joe LaBriola. 2018. Income Inequality and Class Divides in Parental Investments. *American Sociological Review* 83(3):475–507.
Sobel, Michael E. 1998. Causal Inference in Statistical Models of the Process of Socioeconomic Achievement. *Sociological Methods & Research* 27(2): 318–48.
Vaisey, Stephen and Andrew Miles. 2017. What You Can—and Can't—Do With Three-Wave Panel Data. *Sociological Methods & Research* 46(1):44–67.
Verma, Tom S., and Judea Pearl. 1988. Causal Networks: Semantics and Expressiveness. *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence*
Wildeman, Christopher. 2010. Paternal Incarceration and Children's Physically Aggressive Behaviors: Evidence from the Fragile Families and Child Wellbeing Study. *Social Forces* 89(1):285–309.
Wodtke, Geoffrey T. and Daniel Almirall. 2017. Estimating Moderated Causal Effects with Time-Varying Treatments and Time-Varying Moderators: Structural Nested Mean Models and Regression with Residuals. *Sociological Methodology* 47(1):212–245.
Wright, Sewall. 1921. Correlation and Causation. *Journal of Agricultural Research*: 557–585

APPENDIX A. GOTTSCHALK'S FUTURE-TREATMENT STRATEGY

A third future treatment estimator was introduced by Gottschalk (1996). Like Mayer (1997), Gottschalk (1996) premises his analysis on the DGP of Figure 2 and derives a future treatment estimator from its covariance structure. Unlike Mayer, Gottschalk explicitly motivates his estimator with an argument that resembles our difference logic: to use the association between F and Y first to measure and then to subtract bias in the association between T and Y.

Definition 4 (Gottschalk's estimator[16]): Gottschalk's estimator for the causal effect of T on Y, b, is given by

$$b_G = b_{YT} - \sigma_{YF.T} = \sigma_{YT} - (\sigma_{YF} - \sigma_{YT}\sigma_{TF}) \,. \tag{A.1}$$

This estimator is similar, but not identical, to the Mayer/difference estimator. Whereas the difference estimator subtracts two partial regression coefficients, $b_D = b_{YT.F} - b_{YF.T}$, Gottschalk subtracts a conditional covariance from the unadjusted regression of Y on T.

Like Mayer's (1997) estimator, Gottschalk's estimator is biased when U affects T and F differently, $a \neq d$.

Result A.1 (bias of Gottschalk's [1996] estimator in the best case): Gottschalk's estimator is biased when data are generated by the model in Figure 2,

$$b_G = b + ac\left(1 - \frac{d}{a} + ad\right) = b + B_{OLS}M_G, \tag{A.2}$$

But contrary to Gottschalk's claim (his equation 4c), and in contrast to Mayer's (1997) estimator, this estimator is not unbiased in the best-case model of Figure 2 when $a = d$.

Corollary A.1: Gottschalk's estimator remains biased when data are generated by the model in Figure 2 and U affects T and F in the same way, $a = d$,

$$b_G = b + a^3c \neq b. \tag{A.3}$$

Like the Mayer/difference estimator, but unlike our control estimator, Gottschalk's estimator can increase rather than decrease the bias from unobserved confounding when $a \neq d$. Like the Mayer/difference estimator, Gottschalk's estimator strictly increases bias when a and d have opposite signs. Interestingly, however, unlike Mayer's estimator, Gottschalk's estimator is mostly bias reducing when a and d share the same sign and a is strong or moderately strong. Indeed, for magnitudes of $|a|$ larger than about 0.42 (regardless of the value of d), Gottschalk's estimator is strictly bias-reducing.

[16] Our notation is superficially different from Gottschalk's original notation since we assume standardized variables (without loss of generality).

APPENDIX B. FUTURE TREATMENTS AS INSTRUMENTAL VARIABLES

This appendix evaluates the circumstances under which future treatments can, or cannot, serve as instrumental variables (IV). Instrumental variables analysis is a popular strategy for removing bias from unobserved confounding. With a valid IV, F, the causal effect of treatment T on outcome Y in linear DPGs is consistently estimated by the covariance ratio

$$b_{IV} = \frac{\sigma_{FY}}{\sigma_{FT}} . \tag{B.1}$$

IV analysis in linear models requires two assumptions: (1) the instrumental variable must be associated with T ("relevance"); and (2) the IV must be associated with the outcome only via paths that include the causal effect of the treatment on the outcome ("exclusion") (Brito and Pearl 2002). If both assumptions are met, we say that the instrumental variable is valid.

Future treatments are not valid instrumental variables in any of the DGPs considered in the main body of this paper. The key assumption motivating our future-treatment strategies–that F is a proxy for the unobserved confounder, U–violates the exclusion assumption because it induces an association between F and Y via the open path $F \leftarrow U \rightarrow Y$.

For example, the instrumental variables estimator, using F as instrumental variable, in data generated by Figure 2 would evaluate to

$$b_{IV} = \frac{\sigma_{FY}}{\sigma_{FT}} = \frac{bad+cd}{ad} = b + \frac{c}{a} \neq b . \tag{B.2}$$

Recalling that all path parameters lie in the interval $(-1, 1)$, it is obvious that the instrumental variables estimator in this case is strictly more biased than the OLS estimator because

$$|B_{OLS}| = |ac| < |\frac{a}{c}| = |B_{IV}|, \text{ for all } a, c \neq 0 . \tag{B.3}$$

Nonetheless, future treatments have previously been used as instrumental variables, when F was assumed *not* to be a proxy for the unobserved confounders U. For example, Duncan et al. (1997) cautiously defend such a scenario for the estimation of causal neighborhood effects. In their application, Y is children's test scores, T is parents' neighborhood environment while living with the child, and F is parents' neighborhood environment after the child has moved out. Their central assumption is that U can be partitioned into two independent components, as shown in Figure B.1: U_1 represents unobserved parenting quality, which affects child test scores and neighborhood choice while the child lives at home; and U_2 represents parent's residential preferences aside from child rearing considerations.

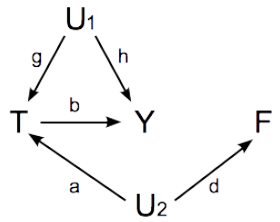


Figure B.1. Model in which future treatments, F, are a valid instrumental variable for the effect $T \rightarrow Y$, because the unobservables, U, are suitably partitioned.

If this model is true, then F is indeed a valid instrument for the effect of T on Y, and the instrumental variables estimator evaluates to

$$b_{IV} = \frac{\sigma_{FY}}{\sigma_{FT}} = \frac{abd}{ad} = b. \qquad \text{(B.4)}$$

As Duncan et al. (1997) have noted, this model may not be especially robust. Instrumental variables estimation would fail under small modifications of the original model, e.g., if parenting, U_1, is associated with future neighborhood conditions (ibid: p. 249), perhaps because concerned parents move to better neighborhoods, or if parent's neighborhood preferences, U_2, are associated with other unobserved factors, such as parental ability, that also affect child test scores (ibid: p. 230). We capture these scenarios in Figures B.2a and B.2b, in which the instrumental variable estimator evaluates to $b_{IV} = b + \frac{ih}{gi+ad} \neq b$ and $b_{IV} = b + \frac{c}{a} \neq b$, respectively. In both of these more elaborate scenarios, F is not a valid instrumental variable because it is a proxy for one or another unobserved confounder, U_1 or U_2, of T and Y, and hence violates the exclusion condition via the open paths $F \leftarrow U_1 \rightarrow Y$ and $F \leftarrow U_2 \rightarrow Y$, respectively.

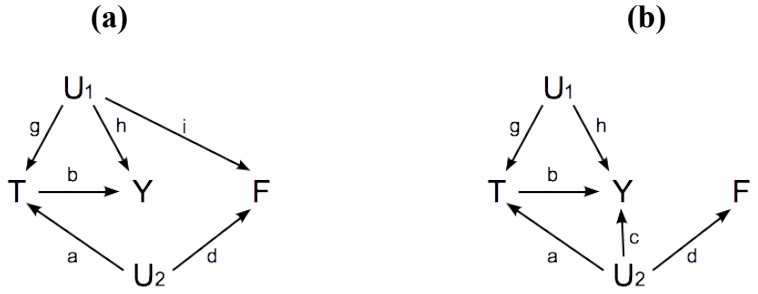


Figure B.2. Two models in which future treatments are not valid instrumental variables for the effect $T \rightarrow Y$.

We further note that F also fails as an instrumental variable even if F is not a proxy for unobserved confounders of T and Y, namely in the presence of true state dependence or selection. True state dependence would occur in Duncan et al.'s (1997) scenario if parents develop a taste for the kind of neighborhood they live in (Deluca 2012), as shown

in Figure B.3a. In this scenario, the exclusion assumption is violated because F is associated with Y via the open path $F \leftarrow T \leftarrow U_1 \rightarrow Y$ (i.e. via a path that does not include the causal effect of T on Y). Consequently, the instrumental variables estimator is biased, $b_{IV} = b + \frac{fgh}{ad+f} \neq b$.

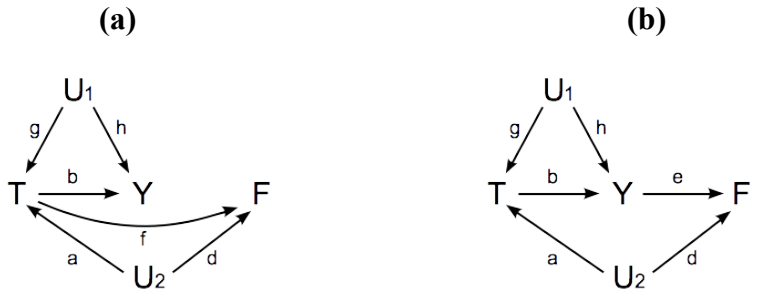


Figure B.3. True state dependence and selection invalidate future treatments as instrumental variables for the effect $T \rightarrow Y$.

Selection would occur if children's test scores affect parents' future residential choice, as shown in Figure B.3b (an admittedly far-fetched proposal, unless, e.g., families relocate in response to children experiencing academic difficulties at a local school). Here, the exclusion condition would be violated because F is directly associated with Y, and the instrumental variable estimator evaluates to

$$b_{IV} = b + \frac{e}{ad+e(b+gh)} \neq b \; . \qquad \text{(B.5)}$$

In a final twist, although true state dependence ($T \rightarrow F$) and selection ($Y \rightarrow F$) invalidate the use of future treatments as instrumental variables, Chan and Kuroki (2010) have shown that descendants of T and Y (which could include future values of the treatment) can sometimes be used to remove unobserved confounding in linear models if true state dependence and selection are suitably mediated in more complicated DGPs. Their results are akin, but not identical, to instrumental variables analysis. To the best of our knowledge, Chan and Kuroki's methodological results have not yet been used in empirical applications.

APPENDIX C. REPLICATION OF MAYER

Table C.1. Descriptives
Means (standard deviations in parentheses); weighted

	Main Sample		Analytic Sample	
	Mayer (1997)	Replication	Mayer (1997)	Replication
Years of Education	12.793	12.838	(a)	12.886
	(1.940)	(1.928)	(a)	(1.957)
Log family income	10.687	11.840	(a)	11.938
	(0.572)	(0.447)	(a)	(0.357)
Log family size	1.647	1.576	(a)	1.609
	(0.331)	(0.331)	(a)	(0.338)
Parent is black	0.141	0.151	(a)	0.170
	(0.347)	(0.358)	(a)	(0.376)
Parent's age	40.127	40.691	(a)	40.899
	(6.163)	(5.908)	(a)	(5.646)
Parent's years of education	12.590	12.593	(a)	12.663
	(2.722)	(2.768)	(a)	(2.859)
Child is a boy	0.481	0.494	(a)	0.717
	(0.498)	(0.500)	(a)	(0.450)
Observations	3,275	3,357	1,853	1,513

Estimates as reported in Mayer (1997), Table A.2 (pp.162-163); (a) Estimates not reported

Table C.2. Full Regression Results
OLS coefficient estimates (standard errors in parentheses); weighted

	Main Sample Unstandardized Coefficients		Analytic Sample Standardized Coefficients	
	Mayer (1997)	Replication	Mayer (1997)	Replication
Log family income	0.784	0.749	0.186	0.185
	(0.065)	(0.074)	(a)	(0.038)
Log family size	-0.714	-0.700	(a)	-0.157
	(0.091)	(0.092)	(a)	(0.025)
Parent is black	0.257	0.276	(a)	0.031
	(0.091)	(0.088)	(a)	(0.034)
Parent's age	0.023	0.033	(a)	0.115
	(0.005)	(0.005)	(a)	(0.025)
Parent's years of education	0.235	0.293	(a)	0.476
	(0.013)	(0.011)	(a)	(0.026)
Child is a boy	-0.032	-0.181	(a)	0.008
	(0.059)	(0.057)	(a)	(0.026)
Constant	1.651	0.082	(a)	-0.042
	(0.652)	(0.824)	(a)	(0.030)
N	3,275	3,357	1,853	1,513
R^2	0.265	0.274	(a)	0.301

Estimates as reported in Mayer (1997), Table B.6 (p.174) for main sample, Table 5.3 (p. 92) for analytic sample; (a) Estimates not directly reported in Mayer (1997)

The Inequality Lab.

Discussion Paper Series

The Inequality Lab at the University of Michigan is a research and training laboratory that investigates the dynamics of social inequality and trains the next generation of inequality scholars. The lab opened in the fall of 2017 to support the study of social inequality, its change across time, and its maintenance across generations. Current projects are focused on wealth inequality and its intergenerational consequences, the determinants and effects of social mobility, and the development of new data and methods to address these topics.

The Discussion Paper series serves to distribute ongoing work by members and affiliates of the Inequality Lab.

The Inequality Lab, University of Michigan
www.theinequalitylab.com