

Gain Scores Revisited: A Graphical Models Perspective

Sociological Methods & Research
2021, Vol. 50(3) 1353-1375
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0049124119826155
journals.sagepub.com/home/smr



Yongnam Kim¹  and Peter M. Steiner¹

Abstract

For misguided reasons, social scientists have long been reluctant to use gain scores for estimating causal effects. This article develops graphical models and graph-based arguments to show that gain score methods are a viable strategy for identifying causal treatment effects in observational studies. The proposed graphical models reveal that gain score methods rely on a bias-removing mechanism that is quite different to regular matching or covariance adjustment. While gain score methods offset noncausal associations via differencing, matching or covariance adjustment blocks noncausal association via conditioning. Since gain score estimators do not rely on conditioning, they are immune to measurement error in the pretest, bias amplification, and collider bias. The graph-based arguments also demonstrate that the key identifying assumption for gain score methods, the common trend assumption, is difficult to assess and justify when the pretest causally affects treatment assignment. Finally, we discuss the distinct role of pretests in the context of Lord's paradox.

Keywords

gain score, pretest, causal graphs, common trend assumption, Lord's paradox

¹ Department of Educational Psychology, University of Wisconsin–Madison, Madison, WI, USA

Corresponding Author:

Yongnam Kim, Department of Educational Psychology, University of Wisconsin–Madison, 1025 West Johnson Street, Madison, WI 53705, USA.
Email: ykim379@wisc.edu

Pretest or baseline measures of the outcome or simply *pretests* have gained much attention in the literature in the social sciences (e.g., Campbell and Stanley 1963; Cook and Steiner 2010; Shadish, Cook, and Campbell 2002). The corresponding literature and empirical evidence from meta-analyses suggest that pretest measures are the most important covariates for removing confounding bias in observational studies (Cook, Shadish, and Wong 2008; Hallberg et al. 2018; Wong, Valentine, and Miller-Bains 2017). Such pretests can be used to compute and analyze *gain scores*, also called change or difference scores, which represent the differences between the posttest and pretest scores (Allison 1990; Kenny 1975; Maris 1998). Nonetheless, gain score methods have long been criticized and frequently avoided by applied researchers and methodologists. Campbell and Erlebacher (1970:197) wrote that “gain scores are in general such a treacherous quicksand,” and Cronbach and Ferby (1970:80) even recommended researchers to “frame their questions in other ways.” This negative view is still widespread among researchers even until recently (Smolkowski 2013; Thomas and Zumbo 2012).

Instead, researchers have preferred covariance adjustment or matching methods that control for or match on the pretest (or a corresponding propensity score) in order to estimate the causal effect of an intervention (e.g., Imbens and Wooldridge 2009). We refer to these methods as *conditioning methods* because the causal effects are obtained “conditional” on the pretest (and other covariates or the corresponding propensity score). Causal identification using conditioning methods relies on the *unconfoundedness* assumption, also called strong ignorability or conditional independence assumption (Imbens 2004; Rosenbaum and Rubin 1983). Meeting the unconfoundedness assumption requires that researchers know all the confounding covariates (or a sufficient set of covariates that blocks all backdoor paths; Pearl, Glymour, and Jewell 2016) and measure them reliably (Steiner, Cook, and Shadish 2011). Since this is rarely the case, the use of conditioning methods in observational studies frequently results in biased effect estimates.

We argue that gain score methods are a viable alternative to identify causal effects when the unconfoundedness assumption is violated. Although the causal assumption underlying gain score methods, the *common trend* assumption, might not be fully met either, gain score estimators have at least three advantages over conditioning estimators (e.g., matching or covariance adjustment estimators): They are immune to (i) unreliability of the pretest, (ii) bias amplification, and (iii) collider bias. Particularly, gain score estimators’ robustness to bias amplification and collider bias has never been discussed in the literature despite the long-standing discussions about gain scores, particularly in the context of Lord’s (1967) paradox. As we will

graphically show, these comparative advantages originate from the difference in the bias-removing mechanism of gain score and conditioning estimators. While conditioning methods remove bias via *blocking* noncausal associations, gain score methods remove bias via *offsetting* noncausal associations by differencing rather than conditioning.

We use graphical models to discuss the identification strategy of gain scores and their advantages in estimating causal effects. A graphical model is a visual representation of the structural causal model of the presumed data generating process of the data at hand. This approach has been developed in computer sciences (Pearl 1988) and epidemiology (Robins 1987) and is now becoming more popular also in the social sciences (e.g., Elwert 2013; Morgan and Winship 2015; Steiner et al. 2017). With respect to causal identification, the use of graphical models has two major advantages over algebraic formulations. First, graphical models allow us to discuss causal assumptions and bias-removing mechanisms in an intuitively appealing but nonetheless formally rigorous way. With graphs and graph-based arguments, we can literally *see* the common trend assumption and the bias-offsetting mechanism of gain score methods. Second, graphical representations of subject-matter theory provide an indispensable tool for assessing the common trend assumption's plausibility in practice. This enables researchers to better defend (or reject) the rather abstract common trend assumption.

In this article, we consider a nonrandomized two-group pretest–posttest design, where the outcomes of the treatment and control groups are measured at two points in time, before and after the intervention. Given the pretest measure, researchers have two main choices to identify and estimate the treatment effect. They can use *conditioning* methods like matching or covariance adjustment, or gain score methods—the classic setting of Lord's (1967) paradox. To ease exposition, we restrict our discussion to *linear* data-generating models with constant effects across all units (extensions to non-linear or nonparametric settings is a topic for future research). In discussing gain score and conditioning estimators, we focus our attention exclusively on bias and do not discuss any efficiency or power issues relevant for significance testing. This does not mean that efficiency and power can be ignored in practice, but our major aim here is to guide researchers in choosing an identification and estimation strategy that results in the least possible bias.¹

This article is organized as follows. In the next section, we provide a brief introduction to graphical models for observational studies and gain scores and discuss the assumptions and mechanisms necessary for identifying causal treatment effects. In the following section, we highlight the three advantages of gain score estimators over conditioning estimators. The section is

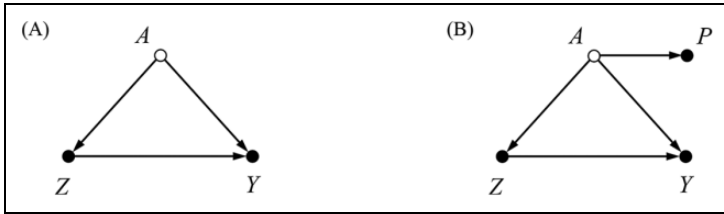


Figure 1. Graphs for observational studies. (A) Graph for an observational study without pretest. (B) Graph for an observational study with pretest.

followed by an in-depth discussion of the common trend assumption. We discuss scenarios and conditions under which the assumption is hard to assess with subject-matter knowledge. We conclude with a discussion of the distinct role of pretests in observational studies.

Graphical Models Perspectives on Observational Studies and Gain Scores

Graphical Models for Observational Studies

It is well-known that causal inference with observational data is challenging because the treatment and control groups are frequently not comparable at baseline (Shadish et al. 2002). In the absence of randomly assigned treatment and control conditions, the observed group difference in the outcome reflects not only potential causal effects but also spurious associations due to confounding (i.e., differential selection of units into the treatment and control groups). However, if researchers succeed to reliably measure a set of covariates that meets the unconfoundedness assumption, then the causal effect is identified and can be estimated via matching or covariance adjustment (if technical assumptions for matching and covariance adjustment are met in addition, e.g., correct specification of the functional form).

The above rationale can be *visualized* by causal graphs. They facilitate our intuitive understanding without sacrificing formal rigor. Consider an example where we are interested in evaluating the effect of participating in a summer math camp (Z) on students' math achievement (Y). Assume that participation in the math camp was not randomized, instead students or their parents decided whether to enroll or not. Further assume that we know, from subject-matter theory and empirical investigations, that students' true but latent math ability (A) is the sole confounding variable that affects both

treatment Z and outcome Y . Figure 1A shows the corresponding graphical model consisting of three nodes and three arrows. The nodes represent the variables and the arrows the causal relationships between the nodes. For instance, the arrow $A \rightarrow Z$ indicates that students' math ability causally affects participation in the math camp (e.g., high-ability students might more likely enroll than low-ability students). Since math ability A also affects math achievement Y ($A \rightarrow Y$), A is referred to as a confounding variable or confounder because A confounds the relation between treatment Z and outcome Y . Since A is unmeasured, its node is vacant; observed nodes are filled. It is important to note that the causal graph describes how the data were actually generated, regardless of whether a variable has been measured. Thus, a graphical model is a graphical representation of the presumed data-generating process, and it typically contains all observed but also unobserved variables that directly or indirectly affect both treatment and outcome.

Given the graphical model in Figure 1A, we see that treatment Z and outcome Y are connected or associated via two different *paths*²:

- (i) $Z \rightarrow Y$,
- (ii) $Z \leftarrow A \rightarrow Y$.

The first path represents the *causal* relationship of interest, while the second path represents a *noncausal* relationship between Z and Y . Both paths are naturally open and thus transmit association. The paths are “naturally” (i.e., without any other conditioning) open because they do not contain a *collider* (Elwert and Winship 2014; Pearl et al. 2016). A collider is a node at which two arrows from its adjacent nodes collide (e.g., C in $A \rightarrow C \leftarrow B$ is a collider variable). A path with a collider does not transmit association without any other conditioning because any association terminates at the collider node, that is, the path is naturally blocked. Therefore, the overall association between Z and Y in Figure 1A is a mixture of the causal and noncausal associations. Unless the noncausal association via path (ii) is stripped out, the observed marginal association between Z and Y does not correspond to the causal relationship between Z and Y via path (i).

The naturally open noncausal paths can be *blocked* by conditioning on any middle node in the paths unless it is a collider. Since A is the sole middle node and not a collider on path (ii), conditioning on A via matching or regression blocks the transmission of noncausal association. Conditional on A , the association between Z and Y is then only determined by the causal association transmitted via path (i), $Z \rightarrow Y$. Thus, the causal effect is identified conditional on A . Pearl (1993) developed a simple graphical criterion,

the *backdoor criterion*, to test whether a set of observed variables is sufficient to identify causal effects via conditioning. The backdoor criterion states that causal effects are identified if all noncausal (or backdoor) paths can be blocked. For our graph in Figure 1A, however, the noncausal path $Z \leftarrow A \rightarrow Y$ cannot be blocked because the ability A is latent and thus unavailable for conditioning. Thus, the causal effect of attending the math camp on math scores is not identifiable via matching or covariance adjustment.

Although the confounder A is unmeasured, researchers may have a pretest measure of the outcome that may serve as a proxy for A . For example, one may measure students' math achievement before the math camp starts. Let P denote such a pretest measure. Then, both pretest and posttest are likely affected by students' math ability. Figure 1B shows the graph with the added pretest: A affects both P and Y , but Z does not affect P (because P is measured before Z). Since P is measured (filled node), we can condition on it (e.g., matching on P or regressing Y on Z and P). However, conditioning on P does still not identify the causal effect because P is not a middle node on the noncausal path $Z \leftarrow A \rightarrow Y$ and thus cannot block the path. Due to the pretest's correlation with ability, conditioning on P may reduce the confounding bias but it cannot eliminate all the bias (Steiner and Kim 2016). As the graphical model in Figure 1B demonstrates, conditioning on a pretest measure is hardly sufficient to identify causal effects in observational studies.

Graphical Models for Gain Scores

In the presence of unmeasured confounding, gain score methods can be an alternative strategy to identify causal effects. Gain score methods first require computing the gain score: $G = Y - P$. In Figure 2A, the gain score G is added as a new node to the graph. Since G is determined by both Y and P , we add two arrows: $P \rightarrow G$ and $Y \rightarrow G$. Moreover, since the gain score is computed as a linear combination of P and Y with fixed coefficients of -1 and $+1$, respectively, we also add the corresponding structural coefficients to the graph in Figure 2A (see Pearl [2016] for a similar graphical representation of gain scores; Shahar and Shahar 2012).³ Assuming linear relationships and constant effects, we now label all arrows with Greek letters, which represent the unknown structural coefficients of the underlying data-generating process. For example, τ on the causal path $Z \rightarrow Y$ represents the constant causal effect of Z on Y .

Gain score methods investigate the causal effect of Z on the gain score G rather than the original outcome Y . That is, we regress G on Z or, equivalently, compare the group mean difference in the gain score G using a two-

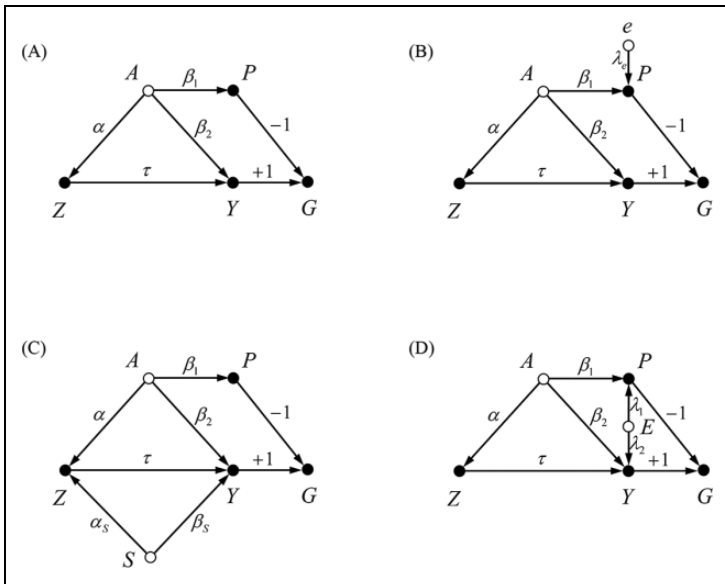


Figure 2. Graphs for gain scores. (A) Simple gain score graph. (B) Graph showing how the independent measurement error e affects the pretest. (C) Graph with two confounders but the pretest is only affected by confounder A . (D) Graph with a common measurement error that affects both the pretest and posttest.

sample t test. This is justified because the causal effect of Z on G is the mediated effect via the causal path $Z \rightarrow Y \rightarrow G$, which is given by the product of the two path coefficients, $\tau \times (+1) = \tau$, and thus identical to the causal effect of Z on Y .⁴

However, with regard to the causal relationship between Z and G in Figure 2A, we now have three noncausal paths:

- (i) $Z \leftarrow A \rightarrow P \rightarrow G$,
- (ii) $Z \leftarrow A \rightarrow Y \rightarrow G$,
- (iii) $Z \rightarrow Y \leftarrow A \rightarrow P \rightarrow G$.

Since the noncausal paths (i) and (ii) are naturally *open* (they do not contain any collider variable), they transmit associations and confound the relation between Z and G . In contrast, path (iii) is naturally *blocked* by the collider Y and thus does not transmit any association. According to the backdoor criterion, the causal effect of Z on G is identified when the two open noncausal paths (i) and (ii) are blocked. Although conditioning on both P and

Y would block the noncausal paths (note that A is unmeasured),⁵ conditioning on Y would also block the causal path $Z \rightarrow Y \rightarrow G$. Since the causal path must remain unblocked, conditioning on both P and Y is not a viable identification strategy.

Gain score methods eliminate the confounding bias by *offsetting* rather than *blocking* the noncausal associations. The association transmitted via the noncausal path (i) can be quantified by the product of the corresponding structural path coefficients on the path: $\alpha \times \beta_1 \times (-1) = -\alpha\beta_1$.⁶ Analogously, we can quantify the associations via the other noncausal paths:

- (i) $Z \leftarrow A \rightarrow P \rightarrow G : -\alpha\beta_1,$
- (ii) $Z \leftarrow A \rightarrow Y \rightarrow G : +\alpha\beta_2,$
- (iii) $Z \rightarrow Y \leftarrow A \rightarrow P \rightarrow G : 0.$

Note that path (iii) transmits no association because this path is naturally blocked by collider Y . If the sum of all the noncausal associations is zero,

$$\alpha\beta_2 - \alpha\beta_1 = \alpha(\beta_2 - \beta_1) = 0,$$

the noncausal associations offset each other, and all the confounding bias is eliminated. Because α is assumed to be nonzero,⁷ the confounding bias cancels out if the unmeasured math ability A affects the pretest math score P and the posttest math score Y to the same extent, $\beta_1 = \beta_2$. This equality condition is frequently referred to as the *common trend* (Lechner 2011) or *time-invariant confounding* assumption.⁸ If the common trend assumption holds, the relation between the treatment and the gain score is free of any confounding and the overall association between Z and G is solely due to the causal effect of Z on G . Importantly, the bias-removing mechanism of offsetting confounding bias does not require any conditioning to block noncausal paths. Therefore, the causal effect can be identified by gain score methods despite the presence of unmeasured confounders (i.e., violation of the unconfoundedness).

It is interesting to investigate what happens if we were to condition on the pretest P in a gain score analysis, regressing G on Z and P . As the gain score graph directly reveals, conditioning on P blocks the noncausal path (i) $Z \leftarrow A \rightarrow P \rightarrow G$ while the noncausal path (ii) $Z \leftarrow A \rightarrow Y \rightarrow G$ remains open. This results in losing the bias offsetting effect of gain score methods and we have the same bias as in a standard matching or covariance adjustment with respect to Y (Allison 1990; Jamieson 2004; Kenny 1975; Laird 1983; Lechner 2011). This is so because the association transmitted via the remaining open path (ii) $Z \leftarrow A \rightarrow Y \rightarrow G$ is identical to the noncausal association via Z

$\leftarrow A \rightarrow Y$. Thus, conditioning on the pretest in a gain score analysis turns the analysis into a standard conditioning method.

Advantages of Gain Score Estimators

Unreliability of Pretest

In practice, pretests are often contaminated with random measurement error. To highlight the impact of measurement error on both conditioning and gain score estimators, we now explicitly add an independent error term e to the graph in Figure 2B (by convention, such random disturbance terms are usually omitted from graphs). The new structural parameter λ_e represents the impact of the measurement error e on the pretest P ($e \rightarrow P$).

Given the graph in Figure 2B, conditioning on P will remove a major part of the confounding bias if P closely resembles A , which is the case whenever measurement error is very small. However, as measurement error e increases, P becomes a weaker proxy for A , resulting in more bias in conditioning estimators. This intuition is confirmed when we consider the regression of Y on Z and P and express the expectation of Z 's partial regression coefficient, $b_{YZ \bullet P}$, in terms of the structural parameters in the graph in Figure 2B (for derivations of all estimator formulae hereafter, see the Appendix):

$$b_{YZ \bullet P} = \tau + \frac{\alpha\beta_2(1 - r)}{\text{Var}(Z) - \alpha^2r}, \tag{1}$$

where r denotes the reliability of P , $r = \frac{\beta_1^2}{\beta_1^2 + \lambda_e^2}$.⁹ The regression estimator consists of the true causal effect (τ) and an additive bias term. The bias term shows that the regression coefficient varies with the impact of the measurement error, λ_e . For example, if measurement error is large (i.e., $|\lambda_e|$ is large in comparison to $|\beta_1|$), the reliability r decreases and the bias term $|\alpha\beta_2(1 - r)|$ increases. That is, $(1 - r)\%$ of the confounding bias induced by A (i.e., $\alpha\beta_2$) is remaining. In addition, the remaining bias is amplified by the factor $\frac{1}{\text{Var}(Z) - \alpha^2r}$ (we discuss bias amplification in the next section). If the pretest is measured without error (i.e., $\lambda_e = 0$ and $r = 1$), then the partial regression estimator is unbiased, $b_{YZ \bullet P} = \tau$. Thus, measurement error attenuates the bias-removing potential of the pretest (see Aiken and West 1991; Steiner et al. 2011).

In comparison to conditioning estimators, gain score estimators are insensitive to measurement error in the pretest (Maris 1998). This is so because the association transmitted via the noncausal path $Z \leftarrow A \rightarrow P \rightarrow G$ does not involve λ_e . According to the path-tracing rule, the association along the path is simply

given by the product of the three path coefficients of α , β_1 , and -1 . In regressing the gain score G on the treatment indicator Z , we can write the expectation of the gain score estimator, b_{GZ} , in terms of structural parameters τ , α , β_1 , and β_2 :

$$b_{GZ} = \tau + \frac{\alpha(\beta_2 - \beta_1)}{\text{Var}(Z)}. \quad (2)$$

The formula clearly shows that the gain score estimator is not a function of λ_e (or the reliability r), revealing its insensitivity to measurement error in the pretest.¹⁰ Suppose that the math pretest is a highly unreliable measure of students' true math ability. In this case, conditioning methods are not able to remove all the bias. Depending on the unreliability, only a minor fraction of the bias might be removed. However, gain score methods' accuracy is unaffected by measurement error and, as long as the common trend assumption holds, $\beta_1 = \beta_2$, they estimate the causal effect without any bias.

Bias Amplification

Steiner and Kim (2016) showed that any remaining bias in conditioning estimators is amplified (also see Pearl 2010, 2011). Bias amplification is a phenomenon that occurs with conditioning methods whenever the conditioning covariates (a) fail to remove the entire bias *and* (b) causally determine treatment selection. The denominator in the bias term of equation (1) contains the amplification factor $\frac{1}{\text{Var}(Z) - \alpha^2 r}$, which is always greater than the factor without conditioning on P , $\frac{1}{\text{Var}(Z)}$. Thus, the subtraction of $\alpha^2 r$ in the denominator determines the extent of bias amplification due to controlling for P in the conditioning estimator. The stronger the ability's effect on treatment selection and the higher the pretest's reliability, the stronger the bias-amplifying effect. However, the term $-\alpha^2 r$ does not occur in the denominator of the gain score estimator in equation (2).

To better see this, consider two unobserved confounders A and S as depicted by the graph in Figure 2C. In addition to the confounder A , the variable S also confounds the relation between the treatment Z and the outcome Y . Since the relations $S \rightarrow Z$ and $S \rightarrow Y$ are described by the structural parameters α_s and β_s , respectively, the confounding bias induced by S is given by $\alpha_s \beta_s$. The graph also shows that the pretest P is affected by A while unaffected by S , indicating that P can serve as a proxy for A but not for S . Hence, conditioning on P does not eliminate any bias induced by the confounder S ; on the contrary, it amplifies the bias due to S .

Given the graph in Figure 2C, the expected partial regression coefficient $b_{YZ\bullet P}$ (i.e., conditioning estimator) is

$$b_{YZ\bullet P} = \tau + \frac{\alpha\beta_2(1-r)}{\text{Var}(Z) - \alpha^2r} + \frac{\alpha_S\beta_S}{\text{Var}(Z) - \alpha^2r}. \tag{3}$$

The first bias term represents, as already discussed, the remaining bias due to P s unreliability with respect to A , while the second bias term shows the hidden bias due to S , $\alpha_S\beta_S$, which is amplified by the factor $\frac{1}{\text{Var}(Z) - \alpha^2r}$. In order to see that bias amplification only occurs if we condition on the pretest P , compare equation (3) to the expected regression estimator without conditioning on P (i.e., regression of Y on Z):

$$b_{YZ} = \tau + \frac{\alpha\beta_2}{\text{Var}(Z)} + \frac{\alpha_S\beta_S}{\text{Var}(Z)}. \tag{4}$$

It becomes clear that, without conditioning on P , the hidden biases due to A and S are not amplified because α^2r is not subtracted from $\text{Var}(Z)$ in the denominators.

Since bias amplification is a phenomenon that only occurs if one *conditions* on covariates, gain score estimators are immune to bias amplification (provided one does not condition on any other covariates).¹¹ Regressing G on Z , the expectation of the gain score estimator is given by:

$$b_{GZ} = \tau + \frac{\alpha(\beta_2 - \beta_1)}{\text{Var}(Z)} + \frac{\alpha_S\beta_S}{\text{Var}(Z)}. \tag{5}$$

Note that the third term, the bias due to S (i.e., $\alpha_S\beta_S/\text{Var}(Z)$), is identical to the third term in equation (4), which is the bias in the unadjusted effect estimate of Z on Y . Although gain score methods do not eliminate the bias due to S , at least they do not amplify the remaining bias.

Collider Bias

Since pretest and posttest are typically measured with the same or a very similar instrument (e.g., same or same type of test or questionnaire items, or interviewers) in the same or a similar setting (e.g., lab or classroom, lab personnel, or teachers), the error terms of the pretest and posttest are very likely correlated. Zimmerman and Williams (1982:153, emphasis added) wrote, “correlated errors [between pretests and posttests] are probably the *rule* rather than the *exception* in pretest-posttest measurements.” The graph in Figure 2D represents such a correlated error structure. The exogenous variable E represents a common source of the correlated measurement errors,

that is, E simultaneously affects the pretest and posttest with structural parameters λ_1 and λ_2 for the causal relations $E \rightarrow P$ and $E \rightarrow Y$, respectively.

Given the data-generating model in Figure 2D, conditioning methods now face the issue of *collider bias*. Compared to the graph in Figure 2A, where the error terms are independent, the correlated error structure in Figure 2D creates an additional noncausal path between Z and Y :

$$Z \leftarrow A \rightarrow P \leftarrow E \rightarrow Y.$$

Since this path contains the collider P , it is naturally blocked at the collider node P . However, once we condition on P , the path becomes unblocked and transmits spurious association (Ding and Miratrix 2015; Elwert and Winship 2014). This spurious association between Z and Y is referred to as *collider bias*. Thus, with correlated errors, conditioning estimators are biased due to the unreliable measurement of A and the collider bias induced by conditioning on the collider P .

The graph in Figure 2D reveals that the correlated error structure via E creates three additional noncausal paths between Z and G :

- (i) $Z \leftarrow A \rightarrow P \leftarrow E \rightarrow Y \rightarrow G$,
- (ii) $Z \leftarrow A \rightarrow Y \leftarrow E \rightarrow P \rightarrow G$,
- (iii) $Z \rightarrow Y \leftarrow E \rightarrow P \rightarrow G$.

However, gain score methods are robust against collider bias because all new noncausal paths via E are naturally blocked either at P or Y because one of them is always a collider on the paths. Thus, no noncausal association is transmitted through these three noncausal paths. Since gain score methods condition neither on P nor on Y , the noncausal paths remain naturally blocked such that collider bias is not an issue for gain score estimators.

This can also be seen from algebraic expressions of the conditioning and gain score estimators. According to the data-generating model in Figure 2D, the expectation of the conditioning estimator is given by:

$$b_{YZ \bullet P} = \tau + \frac{\alpha\beta_2(1-r)}{\text{Var}(Z) - \alpha^2r} - \frac{\alpha\beta_1\lambda_1\lambda_2}{\{\text{Var}(Z) - \alpha^2r\}\text{Var}(P)}. \tag{6}$$

Compared to equation (1), the correlated error structure results in an additional subtractive bias term—the collider bias. This new bias term corresponds to the unblocked collider path $Z \leftarrow A \rightarrow P \leftarrow E \rightarrow Y$, given by the product of the four structural path coefficients of the path: α , β_1 , λ_1 , and λ_2 .

In comparison to conditioning methods, gain score methods are unaffected by collider bias. The expectation of the gain score estimator for the

graph in Figure 2D is identical to equation (2). Although the common cause E generates a correlated error structure, it does not affect the bias in gain score estimators.

The Common Trend Assumption under Different Data-Generating Models

When the Pretest Affects Treatment Selection

The common trend assumption requires that the pretest–posttest change in the outcome Y does not differ between the treatment and control groups in the absence of a treatment effect. Given the previous data-generating models (except for Figures 1A and 2C), the common trend assumption implies that the impact of the unmeasured confounder A on P and on Y is identical, $\beta_1 = \beta_2$. However, whether the equality establishes the common trend assumption strongly depends on the actual data-generating model. The models thus far assumed that the pretest P has no causal effect on treatment Z and posttest Y . This might often be unrealistic in practice. For example, the math pretest score may be known to students and their parents before they decide whether to attend the math camp or not. Then, parents of students with a low pretest score may encourage their children to take the camp, that is, the pretest causally affects treatment selection ($P \rightarrow Z$).

The graph in Figure 3A describes this scenario. The graph has four naturally open noncausal paths between Z and G with the following transmitted associations:

- (i) $Z \leftarrow A \rightarrow P \rightarrow G: -\alpha\beta_1,$
- (ii) $Z \leftarrow A \rightarrow Y \rightarrow G: +\alpha\beta_2,$
- (iii) $Z \leftarrow P \rightarrow G: -\gamma_1\text{Var}(P),$
- (iv) $Z \leftarrow P \leftarrow A \rightarrow Y \rightarrow G: +\gamma_1\beta_1\beta_2.$

Note that the association transmitted via path (iii) depends on $\text{Var}(P)$ because P is the “root” node of this path (see Note 6). To obtain an unbiased estimate of the causal effect using gain score methods, the common trend assumption requires that the sum of the four noncausal associations must be zero. Using $\text{Var}(P) = \text{Var}(\beta_1A + e_P) = \beta_1^2 + \text{Var}(e_P)$, where e_P is the independent (measurement) error of P , the common trend assumption is met if

$$\alpha(\beta_2 - \beta_1) + \gamma_1\{\beta_1\beta_2 - \text{Var}(P)\} =$$

$$(\alpha + \gamma_1\beta_1)(\beta_2 - \beta_1) - \gamma_1\text{Var}(e_P) = 0.$$

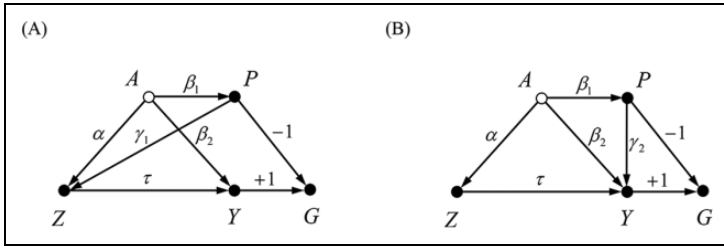


Figure 3. Graphs where the pretest directly affects (A) the treatment or (B) the posttest.

One obvious case that meets this condition is $\beta_1 = \beta_2$ and $\text{Var}(e_P) = 0$, that is, the unmeasured confounder A affects P and Y to the same extent *and* the pretest is measured without error. However, since the pretest is rarely measured without error, the common trend assumption only holds if the first bias term $(\alpha + \gamma_1\beta_1)(\beta_2 - \beta_1)$ and the second bias term $\gamma_1 \text{Var}(e_P)$ perfectly offset each other. But, it is not clear under which conditions such an offsetting can happen. Consequently, when the pretest affects treatment assignment, the common trend assumption is generally hard to assess on subject-matter grounds. Also note, since the association via the noncausal path (iii) depends on the variance of the pretest, and thus on the measurement error in the pretest, the gain score estimator is no longer robust to measurement error in the pretest. For this reason, some authors have advocated against the use of gain scores when the pretest causally determines treatment selection (Allison 1990; Imai and Kim 2019; Maris 1998).

When the Pretest Affects the Posttest Measure

The graph in Figure 3B describes a situation where the pretest causally affects the posttest ($P \rightarrow Y$). Such a situation occurs if the pretest scores are unknown prior to camp enrollment, but after students and parents learn about the pretest score, it may stimulate students' motivation or parents' engagement to organize private tutoring, for instance. With respect to the gain score G , we have three naturally open noncausal paths and corresponding associations:

- (i) $Z \leftarrow A \rightarrow P \rightarrow G : -\alpha\beta_1,$
- (ii) $Z \leftarrow A \rightarrow Y \rightarrow G : +\alpha\beta_2,$
- (iii) $Z \leftarrow A \rightarrow P \rightarrow Y \rightarrow G : +\alpha\beta_1\gamma_2.$

In order to identify the causal effect, the common trend assumption requires that

$$\alpha\{\beta_2 + \beta_1(\gamma_2 - 1)\} = 0.$$

Given $\alpha \neq 0$, the equality holds when $\beta_1 = \beta_2 + \beta_1\gamma_2$. Note that $\beta_2 + \beta_1\gamma_2$ represents the total effect of A on Y (except for the effect via Z): the direct effect of A on Y ($A \rightarrow Y$) plus the mediated effect via P ($A \rightarrow P \rightarrow Y$). Thus, the common trend assumption requires that the impact of A on P is the same as the total impact of A on Y (again, not via Z).

In contrast to the previous case where the pretest affects treatment selection, the implication of the common trend assumption has a clearer substantive interpretation here. Researchers need to assess whether the impact of the unmeasured confounder on the pretest is identical to the cumulative impact of the same confounder on the posttest (i.e., the direct and indirect effect via the pretest). If the two impacts are identical, gain score methods identify the causal effect (and are insensitive to measurement error in P). Note that we do not claim that the common trend assumption is more likely met if the pretest does not affect treatment selection. Rather, we argue that the common trend assumption is easier to assess because the causal graphs of data-generating processes that do not have an effect of P on Z (Figure 2A or Figure 3B) allow for a more meaningful interpretation of the assumption.

Discussion

The widespread reservations about gain scores are partly due to the lack of understanding about how gain score methods actually remove bias. A few methodological articles argued that gain score methods can be effective for causal inference with observational studies (e.g., Allison 1990; Maris 1998; Van Breukelen 2006). However, most of these articles rely exclusively on algebra, which is not easily accessible to many applied researchers. This article revisited the topic with a graphical models approach. The graphical representations visualize the process of how gain score methods identify causal effects and help in understanding and assessing the common trend assumption. For example, our graphical discussion of the common trend assumption provides a clear explanation for why the assumption is hard to assess when the pretest directly affects treatment selection. Corresponding graphical discussions can be easily extended to more complex data-generating models, for instance, when the pretest simultaneously affects both treatment selection and the posttest. Also, if there are multiple unmeasured confounders, more noncausal paths between treatment Z and gain score G

need to be considered, and the common trend assumption holds only if the noncausal associations transmitted along these paths offset each other.

In this article, we also showed that gain score estimators are robust against the unreliability of pretests, bias amplification, and collider bias—issues that may strongly affect conditioning estimators. Nonetheless, gain score methods do not always work and not necessarily remove more bias than conditioning methods like regression or matching adjustments. It is possible that conditioning on the pretest yields less biased or even unbiased effect estimates while gain score estimators might be seriously biased. One of the main messages of this article is that researchers need subject-matter knowledge about the data-generating process to select an appropriate method. Without strong subject-matter knowledge, an informed choice of an appropriate identification strategy and the corresponding estimator is impossible despite the long-standing discussions and investigations since Lord's (1967) seminal article. However, if subject-matter knowledge is available, graphical models are a useful tool to incorporate such knowledge into causal investigations and to choose an identification strategy that has the best chances to remove confounding bias. Graphical models help in understanding gain score methods and Lord's paradox, just as they are useful for discussing missing data problems (Thoemmes and Mohan 2015) and quasi-experimental designs (Steiner et al. 2017).

This article suggests a distinct role of pretests in observational studies. Although the literature has emphasized the importance of pretests (Cook and Steiner 2010; Shadish et al. 2002), it has been unclear what a good pretest is. Our comparison of gain score and conditioning estimators revealed that different methods exploit different characteristics of the pretests. For conditioning estimators, a good pretest must be a close *proxy* of the unmeasured confounders, that is, the pretest and the unobserved confounders should be *nearly perfectly associated*. For gain score estimators, however, a good pretest is affected by the unobserved confounders *to nearly the same extent* as the posttest. When planning an observational study, researchers need to assess whether it is easier to meet the unconfoundedness or common trend assumption with pretest measures. If they think that the pretest may be a good proxy for all unobserved confounders, then they need to put considerable effort into the reliable measurement of a single or multiple pretests. Moreover, in taking repeated measures they should try to avoid correlated errors because of the possibility of collider bias. In contrast, if researchers believe that the unobserved confounders affect the pretest and posttest to almost the same extent, the reliable measurement and independent error structure are less important. Instead, researchers might put more effort into using the same instrument at the pretest and posttest, or into an adequate calibrating or equating of scores from different instruments.

Finally, the graphical discussion of gain scores in this article is a first step in developing graphical models for a broader class of methods including fixed effects models and comparative interrupted time series designs. They belong to the same class of methods as gain score methods because they rely on the bias offsetting mechanism of differencing instead of the bias-blocking mechanism of conditioning methods like matching or covariance adjustment. This distinction (conditioning methods vs. differencing methods) has not been clearly made in the previous literature. For example, it is not rare to find a comparative interrupted time-series design with an additional regression adjustment for or matching on multiple pretests (e.g., St. Clair, Cook, and Hallberg 2014; Wong et al. 2017; also see Abadie, Diamond, and Hainmueller (2010), for synthetic control methods). Our graphs show that conditioning on the pretest when using a gain score regression automatically turns into standard covariance adjustment, which then relies on the unconfoundedness rather than the common trend assumption. More research is needed to reveal similarities but also differences among those methods.

Appendix

Regression Estimator Formula

The linear structural causal model corresponding to the graph in Figure 2D is given by $A = \varepsilon_A$, $Z = \alpha A + \varepsilon_Z$, $Y = \tau Z + \beta_2 A + \lambda_2 E + \varepsilon_Y$, $P = \beta_1 A + \lambda_1 E + \varepsilon_P$, and $G = Y - P$, where ε_A , ε_Z , ε_P , and ε_Y are mutually independent random disturbance terms (omitted from the graph). Without loss of generality, we assume $\text{Var}(A) = \text{Var}(E) = 1$. Then, the expectation of the partial regression coefficient of Z , $b_{YZ \bullet P}$, from the regression of Y on Z and P can be written in terms of bivariate correlations as $b_{YZ \bullet P} = \frac{\rho_{YZ} - \rho_{YP}\rho_{ZP}}{1 - \rho_{ZP}^2} \times \frac{SD(Y)}{SD(Z)}$, where the correlation coefficients are given by

$$\begin{aligned} \rho_{YZ} &= \text{Cov}(\tau Z + \beta_2 A + \lambda_2 E + \varepsilon_Y, \alpha A + \varepsilon_Z) / \{SD(Y)SD(Z)\} \\ &= \{\text{Var}(Z) \tau + \alpha\beta_2\} / \{SD(Y)SD(Z)\}, \end{aligned}$$

$$\begin{aligned} \rho_{YP} &= \text{Cov}(\tau Z + \beta_2 A + \lambda_2 E + \varepsilon_Y, \beta_1 A + \lambda_1 E + \varepsilon_P) / \{SD(Y)SD(P)\} \\ &= (\tau\alpha\beta_1 + \beta_1\beta_2 + \lambda_1\lambda_2) / \{SD(Y)SD(P)\}, \end{aligned}$$

$$\begin{aligned} \rho_{ZP} &= \text{Cov}(\alpha A + \varepsilon_Z, \beta_1 A + \lambda_1 E + \varepsilon_P) / \{SD(Z)SD(P)\} \\ &= \alpha\beta_1 / \{SD(Z)SD(P)\}. \end{aligned}$$

Plugging the population correlations into the formula for $b_{YZ\bullet P}$, we obtain equation (6), $b_{YZ\bullet P} = \tau + \frac{\alpha\beta_2(1-r)}{\text{Var}(Z) - \alpha^2 r} - \frac{\alpha\beta_1\lambda_1\lambda_2}{\{\text{Var}(Z) - \alpha^2\tau\}\text{Var}(P)}$, where r is the reliability of the pretest P , $r = \beta_1^2/\text{Var}(P)$.

Because the structural causal model for Figure 2B is a restricted model of the structural causal model in Figure 2D, we obtain equation (1) by setting either $\lambda_1 = 0$ or $\lambda_2 = 0$: $b_{YZ\bullet P} = \tau + \frac{\alpha\beta_2(1-r)}{\text{Var}(Z) - \alpha^2 r}$.

The structural causal model corresponding to Figure 2C is given by $A = \varepsilon_A$, $S = \varepsilon_S$, $Z = \alpha A + \alpha_S S + \varepsilon_Z$, $Y = \tau Z + \beta_2 A + \beta_S S + \varepsilon_Y$, $P = \beta_1 A + \varepsilon_P$, and $G = Y - P$. The correlation coefficients, based on this model, are given by:

$$\begin{aligned} \rho_{YZ} &= \text{Cov}(\tau Z + \beta_2 A + \beta_S S + \varepsilon_Y, \alpha A + \alpha_S S + \varepsilon_Z) / \{SD(Y)SD(Z)\} \\ &= \{\text{Var}(Z) \tau + \alpha\beta_2 + \alpha_S\beta_S\} / \{SD(Y)SD(Z)\}, \\ \rho_{YP} &= \text{Cov}(\tau Z + \beta_2 A + \beta_S S + \varepsilon_Y, \beta_1 A + \varepsilon_P) / \{SD(Y)SD(P)\} \\ &= (\tau\alpha\beta_1 + \beta_1\beta_2) / \{SD(Y)SD(P)\}, \\ \rho_{ZP} &= \text{Cov}(\alpha A + \alpha_S S + \varepsilon_Z, \beta_1 A + \varepsilon_P) / \{SD(Z)SD(P)\} \\ &= \alpha\beta_1 / \{SD(Z)SD(P)\}. \end{aligned}$$

Plugging the correlation terms into $b_{YZ\bullet P} = \frac{\rho_{YZ} - \rho_{YP}\rho_{ZP}}{1 - \rho_{ZP}^2} \times \frac{SD(Y)}{SD(Z)}$, we obtain equation (3)

$$b_{YZ\bullet P} = \tau + \frac{\alpha\beta_2(1-r)}{\text{Var}(Z) - \alpha^2 r} + \frac{\alpha_S\beta_S}{\text{Var}(Z) - \alpha^2 r}.$$

Relying on the same structural causal model, we can also obtain equations (2), (4), and (5). First, equation (4) is the regression coefficient of Z of the regression of Y on Z , b_{YZ} . Since the coefficient can be written as $b_{YZ} = \text{Cov}(Y, Z)/\text{Var}(Z)$ and using

$$\begin{aligned} \text{Cov}(Y, Z) &= \text{Cov}(\tau Z + \beta_2 A + \beta_S S + \varepsilon_Y, \alpha A + \alpha_S S + \varepsilon_Z) \\ &= \text{Var}(Z) \tau + \alpha\beta_2 + \alpha_S\beta_S, \end{aligned}$$

we obtain $b_{YZ} = \tau + \frac{\alpha\beta_2}{\text{Var}(Z)} + \frac{\alpha_S\beta_S}{\text{Var}(Z)}$. Similarly, for equation (5), using

$$\begin{aligned} \text{Cov}(G, Z) &= \text{Cov}(\tau Z + \beta_2 A + \beta_S S + \varepsilon_Y - \beta_1 A - \varepsilon_P, \alpha A + \alpha_S S + \varepsilon_Z), \\ &= \text{Var}(Z) \tau + \alpha(\beta_2 - \beta_1) + \alpha_S\beta_S, \end{aligned}$$

we obtain the regression coefficient for Z of the regression of G on Z , $b_{GZ} = \tau + \frac{\alpha(\beta_2 - \beta_1)}{\text{Var}(Z)} + \frac{\alpha_S\beta_S}{\text{Var}(Z)}$ (equation [5]). Since the structural causal model

in Figure 2A is a restricted model of the structural causal model in Figure 2C, we obtain equation (2) by setting $\alpha_S = 0$ and $\beta_S = 0$, that is, $b_{GZ} = \tau + \frac{\alpha(\beta_2 - \beta_1)}{\text{Var}(Z)}$.

Authors' Note

The opinions expressed are those of the authors and do not represent views of National Science Foundation, Institute of Education Sciences, or the U.S. Department of Education.


Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was partially supported by a collaborative grant from the National Science Foundation (#2015-0285-00) and a grant from the Institute of Education Sciences (#R305D120005), U.S. Department of Education.

ORCID iD

Yongnam Kim  <https://orcid.org/0000-0001-6731-7123>

Notes

1. Van Breukelen (2006) discusses efficiency issues of gain score estimators.
2. A path is a sequence of adjacent nodes without visiting a node more than once. The directions of the arrows do not matter.
3. An alternative graphical representation, called the latent change score (LCS) model, describes the change score as a latent variable that is affected by the pretest but directly causes the posttest: $P \rightarrow C \rightarrow Y$ (together with $P \rightarrow Y$), where C is the latent change score (Coman et al. 2013; McArdle 2009). Although the LCS model can represent statistical relations among variables, we think that it is limited in representing causal relations. For example, since the LCS model requires that P directly affects Y , but also indirectly via C , it cannot correctly represent the causal model we present in Figure 2A which assumes that the pretest does not causally affect the posttest. Other critical aspects of the LCS model are discussed in Shahar and Shahar (2012).
4. The path-tracing rules, developed by Wright (1921), are applicable to all our graphs because we assume linear causal relationships. For more information about Wright's path-tracing rules in linear models, see Pearl (2013).

5. Conditioning on Y alone would open path (iii) because Y is a collider on the path. However, the additional conditioning on P blocks the open path and thus does not transmit any association conditional on Y and P .
6. To be precise, the product should be multiplied by the variance of the “root” node of the path, such that $-\alpha\beta_1 \times \text{Var}(A)$. Throughout this article, we assume that unmeasured confounders (i.e., vacant nodes) such as A have a unit variance, $\text{Var}(A) = 1$.
7. If $\alpha = 0$, we would have no arrow $A \rightarrow Z$, implying that A is not a confounder.
8. Note that we present here the common trend assumption for linear data-generating models with constant effects. For models with nonlinear relations or effect heterogeneity, the common trend assumption refers to the overall effects transmitted via $Z \leftarrow A \rightarrow P$ and $Z \leftarrow A \rightarrow Y$.
9. Note that we assume $\text{Var}(A) = \text{Var}(e) = 1$.
10. However, the gain score estimator’s variance will be affected by the unreliability in P .
11. It is possible to condition on covariates (other than the pretest) in a gain score analysis. This may be desirable because the common trend assumption can be met only after conditioning on some covariates. Although this strategy may introduce bias amplification in gain score estimators, in this article, we consider the basic gain score estimator, which does not require any other conditioning.

References

- Abadie, A., A. Diamond, and J. Hainmueller. 2010. “Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program.” *Journal of the American Statistical Association* 105:493-505.
- Aiken, L. S. and S. G. West. 1991. *Multiple Regression: Testing and Interpreting Interactions*. Newbury Park, CA: Sage.
- Allison, P. D. 1990. “Change Scores as Dependent Variables in Regression Analysis.” *Sociological Methodology* 20:93-114.
- Campbell, D. T. and A. Erlebacher. 1970. “How Regression Artifacts in Quasi-experimental Evaluations Can Mistakenly Make Compensatory Education Programs Look Harmful.” Pp. 185-210 in *The Disadvantaged Child: Vol. 3. Compensatory Education: A National Debate*, edited by J. Hellmuth. New York: Bruner/Mazel.
- Campbell, D. T. and J. C. Stanley. 1963. “Experimental and Quasi-experimental Designs for Research on Teaching.” Pp. 171-246 in *Handbook of Research on Teaching*, edited by N. L. Gage. Chicago, IL: Rand McNally.
- Coman, E. N., K. Picho, J. J. McArdle, V. Villagra, L. Dierker, and E. Iordache. 2013. “The Paired T-test as a Simple Latent Change Score Model.” *Frontiers in Psychology* 4:738.

- Cook, T. D., W. R. Shadish, and V. C. Wong. 2008. "Three Conditions under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-study Comparisons." *Journal of Policy Analysis and Management* 27:724-50.
- Cook, T. D. and P. M. Steiner. 2010. "Case Matching and the Reduction of Selection Bias in Quasi-experiments: The Relative Importance of Pretest Measures of Outcome, of Unreliable Measurement, and of Mode of Data Analysis." *Psychological Methods* 15:56-68.
- Cronbach, L. J. and L. Furby. 1970. "How We Should Measure "Change": Or Should We?" *Psychological Bulletin* 74:68-80.
- Ding, P. and L. W. Miratrix. 2015. "To Adjust or Not to Adjust? Sensitivity Analysis of M-bias and Butterfly-bias." *Journal of Causal Inference* 3:41-57.
- Elwert, F. 2013. "Graphical Causal Models." Pp. 245-73 in *Handbook of Causal Analysis for Social Research*, edited by S. Morgan. Dordrecht, the Netherlands: Springer.
- Elwert, F. and C. Winship. 2014. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable." *Annual Review of Sociology* 40: 31-53.
- Hallberg, K., T. D. Cook, P. M. Steiner, and M. H. Clark. 2018. "Pretest Measures of the Study Outcome and the Elimination of Selection Bias: Evidence from Three within Study Comparisons." *Prevention Science* 19:1-10.
- Imai, K. and I. S. Kim. 2019. "When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" Princeton, NJ: Princeton University. Retrieved January 21, 2019 from (<https://imai.princeton.edu/research/files/FEmatch.pdf>).
- Imbens, G. W. 2004. "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review." *The Review of Economics and Statistics* 86:4-29.
- Imbens, G. W. and J. M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47:5-86.
- Jamieson, J. 2004. "Analysis of Covariance (ANCOVA) with Difference Scores." *International Journal of Psychophysiology* 52:277-83.
- Kenny, D. A. 1975. "Cross-lagged Panel Correlation: A Test for Spuriousness." *Psychological Bulletin* 82:887-903.
- Laird, N. 1983. "Further Comparative Analyses of Pretest-posttest Research Designs." *The American Statistician* 37:329-30.
- Lechner, M. 2011. "The Estimation of Causal Effects by Difference-in-difference Methods." *Foundations and Trends® in Econometrics* 4:165-224.
- Lord, F. M. 1967. "A Paradox in the Interpretation of Group Comparisons." *Psychological Bulletin* 68:304-05.

- Maris, E. 1998. "Covariance Adjustment versus Gain Scores—Revisited." *Psychological Methods* 3:309-27.
- McArdle, J. J. 2009. "Latent Variable Modeling of Differences and Changes with Longitudinal Data." *Annual Review of Psychology* 60:577-605.
- Morgan, S. L. and C. Winship. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2nd Ed. New York, NY: Cambridge University Press.
- Pearl, J. 1988. *Probabilistic Inference in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann.
- Pearl, J. 1993. "Comment: Graphical Models, Causality, and Intervention." *Statistical Science* 8:266-69.
- Pearl, J. 2010. "On a Class of Bias-amplifying Variables That Endanger Effect Estimates." Pp. 425-32 in *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, edited by Grunwald, P. and P. Spirtes Retrieved January 29, 2019 from (https://ftp.cs.ucla.edu/pub/stat_ser/r356.pdf).
- Pearl, J. 2011. "Understanding Bias Amplification [Invited Commentary]." *American Journal of Epidemiology* 174:1223-27.
- Pearl, J. 2013. "Linear Models: A Useful "Microscope" for Causal Analysis." *Journal of Causal Inference* 1:155-70.
- Pearl, J. 2016. "Lord's Paradox Revisited—(Oh Lord! Kumbaya!)." *Journal of Causal Inference* 4. doi:10.1515/jci-2016-0021.
- Pearl, J., M. Glymour, and N. P. Jewell. 2016. *Causal Inference in Statistics: A Primer*. New York: Wiley.
- Robins, J. M. 1987. "A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect." *Mathematical Modelling* 7:1393-512.
- Rosenbaum, P. R. and D. B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41-55.
- Shadish, W. R., T. D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton-Mifflin.
- Shahar, E. and D. J. Shahar. 2012. "Causal Diagram and Change Variable." *Journal of Evaluation in Clinical Practice* 18:143-48.
- Smolkowski, K. 2013, September 26. *Gain Score Analysis*. Retrieved December 15, 2017, from (http://homes.ori.org/keiths/Tips/Stats_GainScores.html).
- St. Clair, T. S., T. D. Cook, and K. Hallberg. 2014. "Examining the Internal Validity and Statistical Precision of the Comparative Interrupted Time Series Design by Comparison with a Randomized Experiment." *American Journal of Evaluation* 35:311-27.

- Steiner, P. M., T. D. Cook, and W. R. Shadish. 2011. "On the Importance of Reliable Covariate Measurement in Selection Bias Adjustments Using Propensity Scores." *Journal of Educational and Behavioral Statistics* 36:213-36.
- Steiner, P. M. and Y. Kim. 2016. "The Mechanics of Omitted Variable Bias: Bias Amplification and Cancellation of Offsetting Biases." *Journal of Causal Inference* 4. doi:10.1515/jci-2016-0009.
- Steiner, P. M., Y. Kim, C. E. Hall, and D. Su. 2017. "Graphical Models for Quasi-experimental Designs." *Sociological Methods & Research* 46:155-88.
- Thoemmes, F. and K. Mohan. 2015. "Graphical Representation of Missing Data Problems." *Structural Equation Modeling: A Multidisciplinary Journal* 22: 631-42.
- Thomas, D. R. and B. D. Zumbo. 2012. "Difference Scores from the Point of View of Reliability and Repeated-measures ANOVA in Defense of Difference Scores for Data Analysis." *Educational and Psychological Measurement* 72:37-43.
- Van Breukelen, G. J. 2006. "ANCOVA versus Change from Baseline: More Power in Randomized Studies, More Bias in Nonrandomized Studies." *Journal of Clinical Epidemiology* 59:920-25.
- Wong, V. C., J. C. Valentine, and K. Miller-Bains. 2017. "Empirical Performance of Covariates in Education Observational Studies." *Journal of Research on Educational Effectiveness* 10:207-36.
- Wright, S. 1921. "Correlation and Causation." *Journal of Agricultural Research* 20: 557-85.
- Zimmerman, D. W. and R. H. Williams. 1982. "Gain Scores in Research Can Be Highly Reliable." *Journal of Educational Measurement* 19:149-54.

Author Biographies

Yongnam Kim earned his PhD in Educational Psychology (Quantitative Methods) at the University of Wisconsin–Madison and is currently a post-doctoral fellow at the Center for Demography and Ecology at the same university. His research focuses on causal inference, including quasi-experimental designs, graphical models, and causal discovery.

Peter M. Steiner is an Associate Professor in the Department of Educational Psychology, University of Wisconsin–Madison. His research interests are in causal inference with experimental and quasi-experimental designs, graphical models, and causal replication.