Causal Graphical Views of Fixed Effects and Random Effects Models

Yongnam Kim[1] & Peter M. Steiner[2]

[1] Department of Educational, School & Counseling Psychology, University of Missouri–Columbia, Hill Hall, Columbia, MO 65211. Email: ykcpb@missouri.edu

[2] Department of Human Development and Quantitative Methodology, University of Maryland, 3942 Campus Drive, College Park, MD 20742. Email: psteiner@umd.edu

June 6, 2020

ACKNOWLEDGEMENTS

The authors thank Felix Elwert for helpful comments on the previous version of this manuscript.

ABSTRACT

Despite the long-standing discussion on fixed effects (FE) and random effects (RE) models, how and under which conditions both methods can eliminate unmeasured confounding bias have not yet been widely understood in practice. Using a simple pretest-posttest design in a linear setting, this article translates the conventional algebraic formalization of FE and RE models into causal graphs and provides intuitively accessible graphical explanations about their data-generating and bias-removing processes. The proposed causal graphs highlight that FE and RE models consider different data-generating models. RE models presume a data-generating model that is identical to a randomized controlled trial while FE models allow for unobserved time-invariant treatment-outcome confounding. Augmenting regular causal graphs that describe data-generating processes by adding the computational structures of FE and RE estimators, the article visualizes how FE estimators (gain score and deviation score estimators) and RE estimators (quasi-deviation score estimator) offset unmeasured confounding bias. In contrast to standard regression or matching estimators that reduce confounding bias by blocking non-causal paths via conditioning, FE and RE estimators offset confounding bias by deliberately creating new non-causal paths and associations of opposite sign. Though FE and RE estimators are similar in their bias-offsetting mechanisms, the augmented graphs reveal their subtle differences that can result in different biases in observational studies.

KEYWORDS: fixed effect, random effect, causal graph, bias-offsetting, gain score, demeaning

INTRODUCTION

Fixed effects (FE) and random effects (RE) models are often used in the social sciences for studies with longitudinal or panel data. Using repeated measures of the outcome, researchers try to evaluate the causal impact of a program or policy in observational settings where treatment assignment is non-randomized. However, the current literature on FE and RE models is often too technical or math-intensive and, thus, despite the voluminous literature (e.g., Allison, 2009; Bell & Jones, 2015; Bollen & Brand, 2010; Clark & Linzer, 2015; Halaby, 2004; Lockwood & McCaffrey, 2007; Wooldridge, 2010, 2012), some applied researchers, especially outside of econometrics and statistics, still find it hard to understand the basic principles how these methods can produce unbiased causal effect estimates in the presence of unmeasured confounding. Strong exclusive preferences for either FE or RE models found in the methodological literature (e.g., Allison, 1994, p. 181; Gelman & Hill, 2007, p. 246) also seem to further confuse many applied researchers.

The purpose of this article is to provide an intuitive but nonetheless formal discussion of the causal identification and estimation of treatment effects using basic FE and RE models in a linear setting of a pretest-posttest design with no treatment exposure prior to the pretest. To achieve this goal, the article translates the conventional algebraic expressions of FE and RE models into *causal graphical models*, also referred to as *directed acyclic graphs* (DAGs; Pearl, 1988; Spirtes, Glymour, & Scheines, 2001). A group of social scientists has recently shown that some key aspects of causal inference can be formalized in an intuitively appealing way using causal graphs: Elwert and Winship (2014) formalized endogenous selection bias, Thoemmes and Mohan (2015) discussed missing data problems, and Steiner, Kim, Hall, and Su (2017)

developed causal graphs for quasi-experimental designs. Similar graphical advances are also possible for FE and RE models.

Using causal graphs, this article visualizes the similarities and differences between FE and RE models in terms of presumed data-generating models and specific bias-removing mechanisms. We argue that both models belong to the same class of methods, aiming at *offsetting* rather than partialling out confounding bias. They offset the confounding association by deliberately creating non-causal associations of opposite sign. This is in sharp contrast to the bias-removing mechanism of covariate adjustment such as regular regression or matching methods[1] that aim at *blocking* the confounding association. However, FE and RE models differ in how they offset the confounding association. We provide causal graphical explanations about how FE estimators like *gain score* and *deviation score estimators*, and the RE or *quasi-deviation score estimator* address confounding bias in their unique ways.

The rest of this article is organized as follows. The next section introduces the conventional algebraic formalization of FE models. This formalization is then translated into causal graphical models in the following section. Augmented graphical models then show the specific bias-removing mechanisms using gain scores and deviation scores. In the following section, graphical representations of RE models, in particular random intercept models, using quasi-deviation scores are presented. Finally, the article compares biases of FE and RE estimators when their causal identification assumptions are violated.

[1] By regular matching, we consider a process to match *different* units that have the same or similar baseline covariates. This does not include what Imai and Kim (2019) referred to as "within-unit matching" that matches repeated measures of the *same* unit.

ALGEBRAIC FORMALIZATION OF FIXED EFFECTS MODELS

For a better understanding of the key features of FE and RE models, we focus on a simple pretest-posttest study, also referred to as before-after study, where the outcome variable is measured at two time points, the first time before treatment implementation and the second time after treatment implementation. The two outcome measurements are referred to as *pretest* and *posttest*, respectively (Shadish, Cook, & Campbell, 2002). As typical for many pretest-posttest studies in psychology and education, we assume that no units have been exposed to the treatment condition prior to the pretest. After the pretest, units self-select or are non-randomly assigned into the treatment condition. Although we focus on fixed effects across time, FE models may also exploit spatial or other variations. For example, Ashenfelter and Krueger (1994) used *twins* to deal with *family*-fixed effects. The principles discussed here extend to such cases as well.

Fixed Effects Data-Generating Models

A *fixed effect* is an unobserved unit-specific effect that affects the pretest and posttest to the same extent. That is, the effect does not change over time. The basic FE model is expressed as

$$Y_{it} = \tau A_{it} + \theta_i + \varepsilon_{it}, \quad (1)$$

where $i = 1, \ldots, N$ denotes the units (e.g., participants) and t denotes the time index, $t \in \{1 = pretest,\ 2 = posttest\}$. The outcome Y_{it}, the treatment variable A_{it}, and the random disturbance ε_{it} are time-varying as indicated by the time index t. In contrast, the unit-specific but unobserved fixed effect θ_i does not depend on time (no time-indexed), and thus affects a unit's pretest and posttest to the same extent. We consider that all variables are continuous here, though

the treatment may also be binary (then, our pretest-posttest design becomes a standard difference-in-differences setup). The constant treatment effect τ is the target causal quantity of interest.

The meaning of Equation (1) can be better understood if one derives separate equations for the pretest and posttest, respectively. At the pretest, $t = 1$, where all units are still in the control condition, $A_{i1} = 0$,[2] Equation (1) can be written as

$$\begin{aligned} Y_{i1} &= \tau A_{i1} + \theta_i + \varepsilon_{i1} \\ &= \theta_i + \varepsilon_{i1}. \end{aligned} \tag{2}$$

Thus, the pretest Y_1 is determined by (i) the unit-specific fixed effect θ and (ii) the time-specific random disturbance ε_1 (hereafter, we drop the unit-subscript *i* unless it causes ambiguities). The random disturbance is assumed to be independent of the unit-fixed effect, $Cov(\varepsilon_{i1}, \theta_i) = 0$.

Similarly, at the posttest, $t = 2$, we obtain

$$Y_{i2} = \tau A_{i2} + \theta_i + \varepsilon_{i2}. \tag{3}$$

The posttest Y_2 is determined by (i) the same unobserved fixed effect θ, (ii) the time-specific random disturbance ε_2, and (iii) the treatment effect τ. Again, the random disturbance is assumed to be uncorrelated with the unit-fixed effect, $Cov(\varepsilon_{i2}, \theta_i) = 0$, also with the treatment, $Cov(\varepsilon_{i2}, A_{i2}) = 0$, but the fixed effect is correlated with the treatment, $Cov(A_{i2}, \theta_i) \neq 0$. This dependence indicates that the fixed effect confounds the relation between treatment selection A_2

[2] Formally, in our setup where no one was exposed to the treatment at the pretest ($t = 1$), $A_{i1} = A_{j1}$, for all $i \neq j$, which results in $Var(A_{i1}) = 0$, that is, A_{i1} is a constant. For ease of exposition, we use $A_{i1} = 0$ but it can be any constant value. But, at the posttest, A_{i2} is not a constant, that is, there should be both treated and control cases at $t = 2$.

and the posttest Y_2. Note that we do not impose any distributional restrictions on the random disturbances. This implies that we allow for $E(\varepsilon_{i1}) \neq E(\varepsilon_{i2})$.[3] The inequality may reflect changes in the outcome over time due to maturation, changes in instrumentation, or other history effects like economic changes (Shadish et al., 2002). We also allow the two random disturbances to be correlated with each other, $Cov(\varepsilon_{i1}, \varepsilon_{i2}) \neq 0$, possibly due to a common source of measurement error (Kim & Steiner, 2019).

Equations (1) to (3) describe the presumed data-generating model of FE models with no treatment exposure at the pretest. Those equations should not be confused with analytic models chosen by *researchers* because they differ from how the data were generated by *nature*. To distinguish data-generating models from analytic models, we refer to Equations (1) to (3) as the *FE data-generating model*.

Standard Regression Estimators

Given data from a pretest-posttest study, researchers can choose among very different analytic models and the corresponding estimators. One of the most popular analytic models for estimating the effect of A_2 on Y_2 is standard regression with or without the pretest included as a control variable. First, consider the naïve regression model that regresses the posttest Y_2 on the treatment A_2 (without the pretest and any other control variables): $\hat{Y}_{i2} = a + bA_{i2}$, where a and b are the intercept and slope, respectively. Then, under the FE data-generating model in Equation (3), the expectation of the effect estimator b is given by

[3] One may explicitly express this mean difference over time in fixed effects data-generating models such as $Y_{it} = \tau A_{it} + \theta_i + \gamma_t + \varepsilon_{it}$, where γ_t denotes the time-fixed effect and $E(\varepsilon_{i1}) = E(\varepsilon_{i2})$. This model is occationally referred to as two-way fixed effects models and is distingushed from unit-fixed effects models where no time-fixed effect is allowed (see Imai & Kim, 2019, 2020).

$$E(b) = \frac{Cov(Y_2, A_2)}{Var(A_2)} = \frac{Cov(\tau A_2 + \theta + \varepsilon_2, A_2)}{Var(A_2)}$$
$$= \tau + \frac{Cov(A_2, \theta)}{Var(A_2)}. \quad (4)$$

The bias term $Cov(A_2, \theta)/Var(A_2)$ does not disappear because we previously assumed that the covariance between the treatment and the fixed effect is not zero (unless the data come from a perfectly implemented randomized controlled trial). Thus, the naïve regression estimator under the FE data-generating model is generally biased.

In an attempt to remove confounding bias, one may control for the pretest Y_1 in the regression $\hat{Y}_{i2} = a + bA_{i2} + cY_{i1}$. This regression model is often called a lagged dependent variable model (Wooldridge, 2010) or ANCOVA approach. The analytic model aims at estimating the treatment effect τ via regression coefficient b after partialling out the effect of the pretest Y_1. However, in general, this does not produce an unbiased effect estimator either. As shown in Appendix A, the expectation of the resulting partial regression coefficient of A_2 from the lagged regression model is given by

$$E(b) = \tau + \frac{Cov(\theta,\ A_2)\{Var(Y_1) - Var(\theta)\}}{Var(A_2)Var(Y_1) - Cov(A_2, \theta)^2} \quad (5)$$

Since $Cov(A_2, \theta) \neq 0$ and $Var(Y_1) \neq Var(\theta)$, the bias term in Equation (5) does not vanish despite the inclusion of the pretest in the regression model. Note that the variance equality, $Var(Y_1) = Var(\theta)$, holds only if $Var(\varepsilon_1) = 0$ because $Var(Y_1) = Var(\theta + \varepsilon_1) = Var(\theta) + Var(\varepsilon_1)$, according to Equation (2). Thus, the variance equality requires the absence of a random disturbance term in the pretest, implying that the pretest is a perfect measure of the unmeasured

(a) (b)

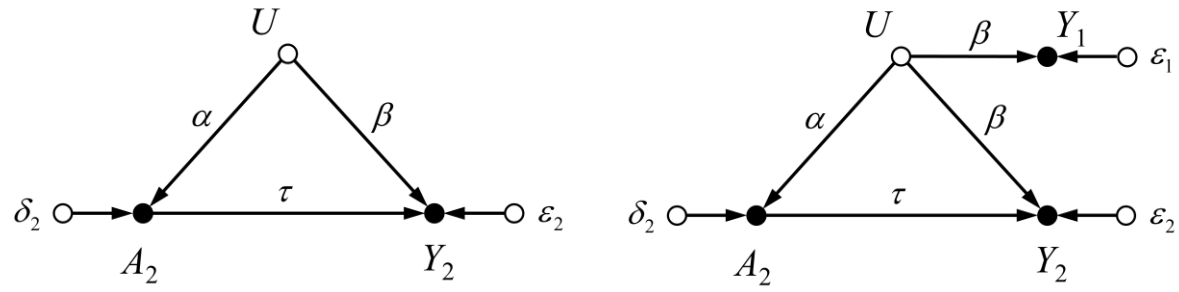

FIGURE 1. *Causal graphs of the FE data-generating models for the posttest only (a) and the pretest and posttest together (b). Vacant nodes indicate unobserved variables.*

fixed effect—a rather unrealistic scenario. In sum, for the FE data-generating model described in Equations (1) to (3), researchers' choice of standard regression estimators will generally result in biased effect estimates.

GRAPHICAL REPRESENTATIONS OF FIXED EFFECTS MODELS

Graphs for Fixed Effects Data-Generating Models

The data-generating process of FE models can also be formalized with causal graphs consisting of arrows and nodes. The causal graph in Figure 1a shows for the posttest at $t = 2$, how the treatment and outcome are generated in an observational setting. The graph encodes that (i) treatment selection A_2 is causally determined by an unmeasured variable *U* and a random disturbance δ_2, and (ii) the posttest Y_2 is causally determined by both *U* and A_2 and its own random disturbance ε_2. Throughout this article, unmeasured variables (e.g., *U*, δ_2, ε_2) are drawn as unfilled nodes while all measured variables (e.g., A_2, Y_2) are drawn as filled nodes. Arrows

represent causal effects, for example, the causal effect of A_2 on Y_2 is τ and the causal effect of *U* on Y_2 is β. The corresponding linear structural causal model generating Y_2 can be written as

$$Y_{i2} = \tau A_{i2} + \beta U_i + \varepsilon_{i2}. \quad (6)$$

Comparing Equations (6) and (3), the fixed effect θ in Equation (3) can be interpreted as a combined entity of an unmeasured confounder *U* and its impact β in Equation (6): $\theta_i = \beta U_i$.[4]

For the pretest model in Equation (2), we obtain a similar linear structural causal model with the same fixed effect ($\theta_i = \beta U_i$):

$$Y_{i1} = \beta U_i + \varepsilon_{i1}. \quad (7)$$

Equation (7) states that the pretest Y_1 is determined by *U* and a random disturbance ε_1. Note that the causal effect of *U* on the pretest Y_1 is identical to the causal effect of *U* on the posttest Y_2 (i.e., both effects are β).

The causal graph in Figure 1b combines the two data-generating processes of the pretest and posttest. Note that we did not draw the node A_1 because $A_{i1} = 0$ (or any other constant) for all units in our setup. The graph clearly shows the implication of the FE data-generating model: the repeated outcome measures Y_1 and Y_2 are affected by *the same unmeasured confounding variable* (*U*) *to the same extent* (β). As the impact of the same unmeasured confounding variable

[4] Together with the linearity and constant effects assumptions, this very specific representation of fixed effects allows for a simple interpretation of the fixed effects assumption. More generally, the fixed effect can be written and graphically represented as a non-parametric function $\theta_i = f_i(\beta_i, U_i)$, where U_i and β_i might be vectors. Then, the fixed effects assumption requires that the associations transmitted along the two paths $A_2 \leftarrow U \rightarrow Y_1$ and $A_2 \leftarrow U \rightarrow Y_2$ are identical (instead of the effects of A_2 on Y_1 and Y_2). The graphical discussions of fixed and random effects estimators presented in the following sections can be extended to this more general case but this is beyond the scope of this article.

is time-invariant, it is referred to as the fixed effect over time. We refer to this equality restriction between the two impacts of *U* as the *fixed effects assumption* (can also be viewed as the *common* or *parallel trend* assumption; see Kim & Steiner, 2019; Lechner, 2011). The causal graph in Figure 1b is a graphical version of the presumed FE data-generating model algebraically described in Equation (1).[5]

Assessing Causal Identification Using Graphs

Translating the algebraic formalization of FE data-generating models into causal graphs helps researchers in assessing whether the causal effect of interest is identified and estimable. Given the graph in Figure 1b, one immediately sees why the causal effect of A_2 on Y_2 cannot be estimated without bias by standard regression models. According to the *backdoor criterion* (Pearl, 1988, 2009) or the *adjustment criterion* (Shpitser, VanderWeele, & Robins, 2010), identification of causal effects requires that all non-causal paths are blocked by conditioning on middle-variables on the non-causal paths (i.e., by controlling for or matching on the variables). If it is not possible to block all non-causal paths due to omitted or unreliably measured variables, the treatment and outcome remain connected via non-causal (or backdoor) paths and the remaining confounding association biases the effect estimators. In Figure 1b, since *U* is unmeasured, the non-causal path $A_2 \leftarrow U \rightarrow Y_2$ cannot be blocked, therefore, the causal effect τ is not identified. Conditioning on the observed pretest Y_1 cannot block the non-causal path either because Y_1 is not a variable on this path. Thus, although controlling for Y_1 partially removes

[5] For simplicity, in this article we do not describe the correlation between the two random disturbances, $Cov(\varepsilon_{i1}, \varepsilon_{i2}) \neq 0$, in causal graphs. That is, we do not add a structure like $\varepsilon_1 \leftarrow E \rightarrow \varepsilon_2$, where *E* represent a common source of measurement error. Adding the structure does not affect our findings because all the resulting additional paths will remain blocked. See Kim and Steiner (2019).

confounding bias to the extent of its squared correlation, the standard regression estimators of the treatment effect generally remain biased (Steiner & Kim, 2016).

Notice that the presence of bias in standard regression estimators, algebraically derived in Equations (4) and (5) in the previous section, can be simply assessed by checking for unblocked non-causal paths in the causal graph in Figure 1b. This demonstrates why the causal graphical approach is useful for applied researchers. In the following, we extend such intuitively accessible graphical arguments to other types of estimators. To keep the graphs simple, henceforth we omit the random disturbance terms δ_2, ε_1, and ε_2 (mutually independent of each other) because they do not play a special role in assessing causal identification.

Fixed Effects Estimators: Gain Score, Deviation Score, and Dummy Variable Estimators

Although the open backdoor path via *U* cannot be blocked by standard regression estimators, the econometrics literature suggests three estimators to identify the causal effect τ from the FE data-generating model described in Equation (1) or Figure 1b: i) *first-differencing* estimator, ii) *time-demeaning* estimator, and iii) *dummy variable regression* estimator (Wooldridge, 2010, 2012). In psychology and education, the first-differencing estimator is often referred to as the *gain score* estimator which will be used in this article. We also use the alternative term *deviation score* estimator instead of time-demeaning estimator. The key features of these estimators can become intuitively clear from augmented causal graphs we propose.

Gain Score Estimator

The gain score estimator relies on first differencing between the pretest and posttest. Figure 2a shows the graph of the FE data-generating model with the added gain score variable *G*.

(a) (b)

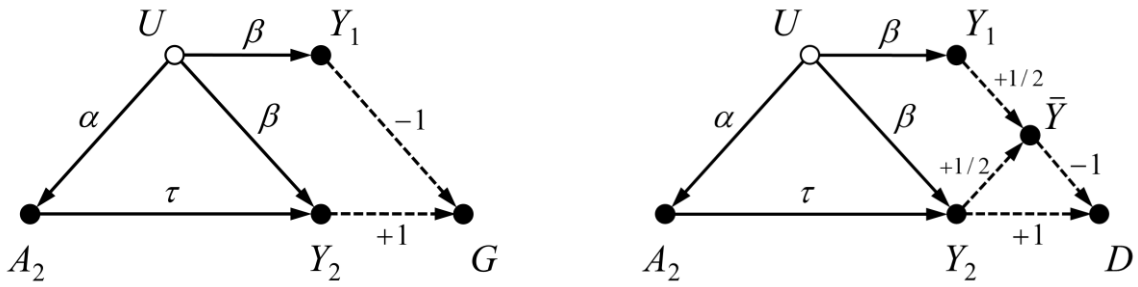

FIGURE 2. *Causal graphs of the gain score model (a) and deviation score model (b). Solid arrows represent the data-generating model and dashed arrows represent the computation of gain and deviation scores.*

We use dashed arrows to visualize the computation of the gain score which is obtained by subtracting the pretest from the posttest:

$$G_i = Y_{i2} - Y_{i1}. \tag{8}$$

The dashed arrows in Figure 2a represent Equation (8) with known coefficients -1 and $+1$ for $Y_1 \dashrightarrow G$ and $Y_2 \dashrightarrow G$, respectively. A standard gain score model then investigates the effect of the treatment on the gain score (instead of the original outcome): $\hat{G}_i = a + bA_{i2}$. The graph in Figure 2a shows that A_2 and G are connected by one *causal* and two *non-causal* paths:

(i) $A_2 \rightarrow Y_2 \dashrightarrow G$ (causal);

(ii) $A_2 \leftarrow U \rightarrow Y_1 \dashrightarrow G$ (non-causal);

(iii) $A_2 \leftarrow U \rightarrow Y_2 \dashrightarrow G$ (non-causal).

An association transmitted along a path can be computed using the following path-tracing rules (Wright, 1921; also see Pearl, 2013): Multiply the variance of the path's *root variable* with the product of path coefficients along the path. A root variable is a variable on a path that does not have any incoming arrows. For example, the root variable in path (i) is A_2, and the root variable in paths (ii) and (iii) is *U*. Thus, the associations transmitted by the three paths are respectively given by

$$\text{(i)} \quad \tau \times (+1) \times Var(A_2) = Var(A_2)\tau;$$

$$\text{(ii)} \quad \alpha \times \beta \times (-1) \times Var(U) = -Var(U)\alpha\beta;$$

$$\text{(iii)} \quad \alpha \times \beta \times (+1) \times Var(U) = Var(U)\alpha\beta.$$

The total association between A_2 and *G*, expressed as $Cov(A_2,\ G)$, is then given by the sum of the three associations. Since the non-causal associations transmitted via path (ii) and (iii) offset each other, $-Var(U)\alpha\beta + Var(U)\alpha\beta = 0$, we obtain $Cov(A_2, G) = Var(A_2)\tau$.[6] That is, in analyzing gain scores, the association transmitted along the newly created non-causal path (ii) via the pretest Y_1 offsets the original confounding association induced by path (iii).[7]

Algebraically, the expectation of the gain score estimator *b* of the regression $\hat{G}_i = a + bA_{i2}$ is given by

[6] With a binary treatment A_2, the meaning of the parameter α becomes more complex but does not affect the bias-removing mechanism of gain (also deviation) score estimators. In this case, the relation between *U* and A_2 is nonlinear, and α now represents the coefficient of the linear projection of *U* onto A_2. However, even with the nonlinear relation, path (ii) and path (iii) offset each other and the resulting covariance is $Cov(A_2, G) = Var(A_2)\tau$, which is only due to path (i).

[7] Thus, the perfect offsetting of the non-causal paths in the gain and deviation score graphs relies on a stable but unfaithful relation between A_2 and *G* (or *D*) in the data-generating process of the population under consideration. Such stable but unfaithful relations in causal graphs are discussed in Greenland and Mansournia (2015).

$$E(b) = \frac{Cov(A_2, G)}{Var(A_2)} = \frac{Var(A_2)\tau}{Var(A_2)} = \tau \quad (9)$$

Thus, the gain score estimator identifies the causal effect τ in the presence of the open backdoor path $A_2 \leftarrow U \rightarrow Y_2 \dashrightarrow G$ in Figure 2a. See Kim and Steiner (2019) for a more detailed discussion of gain score estimators, particularly their insensitivity to measurement error in the pretest, bias amplification, and collider bias (due to the correlation between ε_1 and ε_2), which also apply to deviation score estimators.

Deviation Score Estimator

The deviation score estimator relies on a bias-offsetting mechanism that slightly differs from the gain score estimator (Allison, 1994; Wooldridge, 2010, 2012). A deviation score (*D*) is defined as the deviation of the posttest (Y_2) from the unit-specific average ($\bar{Y}$) of the two repeated outcome measures:

$$D_i = Y_{i2} - \bar{Y}_i, \quad (10)$$

where $\bar{Y}_i$ is the average of the pretest and posttest for unit *i*,

$$\bar{Y}_i = \frac{Y_{i1} + Y_{i2}}{2}. \quad (11)$$

In a deviation score analysis, the treatment effect is estimated by regressing the deviation score on the treatment.

Figure 2b shows the augmented graph with the added computation of the deviation score to the FE data-generating model. According to Equation (11), the average node $\bar{Y}$ is depicted as a

variable that is determined by the pretest Y_1 and the posttest Y_2, $Y_1 \dashrightarrow \bar{Y} \dashleftarrow Y_2$, with path coefficients $+1/2$. Equation (10) states that the deviation score *D* is determined by the posttest and the unit-specific mean, $Y_2 \dashrightarrow D \dashleftarrow \bar{Y}$, with the path coefficients of $+1$ and -1 for $Y_2 \dashrightarrow D$ and $\bar{Y} \dashrightarrow D$, respectively. As with the gain score analysis, the graphical representation visualizes how deviation score analysis eliminates unmeasured confounding in the FE data-generating model. Between the treatment A_2 and the deviation score *D*, three open *non-causal* paths transmit the following non-causal associations:

(i) $A_2 \leftarrow U \rightarrow Y_2 \dashrightarrow D$ (non-causal), $Var(U)\alpha\beta$;

(ii) $A_2 \leftarrow U \rightarrow Y_2 \dashrightarrow \bar{Y} \dashrightarrow D$ (non-causal), $-Var(U)\alpha\beta/2$;

(iii) $A_2 \leftarrow U \rightarrow Y_1 \dashrightarrow \bar{Y} \dashrightarrow D$ (non-causal), $-Var(U)\alpha\beta/2$.

Then, the sum of the three non-causal associations is zero. That is, the original confounding bias induced by *U* in path (i) is offset by the two deliberately created non-causal associations transmitted via the non-causal paths (ii) and (iii) traversing through the unit-specific average $\bar{Y}$.

However, despite offsetting all non-causal associations, the effect of A_2 on *D* is not identical to the causal effect of A_2 on Y_2. The graph in Figure 2b shows that the causal effect of A_2 on *D* consists of two causal paths with the following causal associations:

(i) $A_2 \rightarrow Y_2 \dashrightarrow D$ (causal), $Var(A_2)\tau$;

(ii) $A_2 \rightarrow Y_2 \dashrightarrow \bar{Y} \dashrightarrow D$ (causal), $-Var(A_2)\tau/2$.

Thus, the sum of the two causal associations is $Var(A_2)\tau/2$, which is *half* of the causal association between A_2 and *G* in the gain score model. Therefore, the regression of $\widehat{D}_i = a + bA_{i2}$ results in half of the causal effect τ:

$$E(b) = \frac{Cov(A_2, D)}{Var(A_2)} = \frac{\tau}{2}. \quad (12)$$

This is not surprising because time-demeaning the posttest cuts the pretest-posttest differences in half, $D_i = Y_{i2} - \bar{Y}_i = (Y_{i2} - Y_{i1})/2$, which implies that also the treatment effect is cut into half. Thus, a multiplication by *two* is necessary to recover the target causal quantity τ in a two-period panel data study. However, researchers using FE models usually also demean the treatment variable, then this additional consideration is not required. The demeaned treatment *C* is defined as,

$$C_i = A_{i2} - \bar{A}_i, \quad (13)$$

where $\bar{A}_i$ is the average of the repeated treatment measures for unit *i*:

$$\bar{A}_i = \frac{A_{i1} + A_{i2}}{2} = \frac{A_{i2}}{2} \quad (14)$$

because $A_{i1} = 0$. Then, the deviation score estimator obtained from regressing the deviation score *D* on the demeaned treatment C, $\widehat{D}_i = a + bC$, is identical to the causal effect τ:

$$E(b) = \frac{Cov(C, D)}{Var(C)} = \frac{Var(C)\tau/4}{Var(C)/4} = \tau. \quad (15)$$

Although we need to deal with more paths, this causal identification using the demeaned treatment is also well justified using causal graphs (see Appendix B for the full graphical analysis). Our graphical explanation reveals that time-demeaning the posttest and the treatment variable serves two different purposes, an aspect not explicitly addressed in the literature: time-demeaning the posttest offsets the confounding bias while time-demeaning the treatment recovers the causal quantity of interest.

Dummy Variable Regression Estimator

In practice, deviation score estimators are rarely used. Researchers prefer to use *dummy variable regression estimators*, which are numerically identical to deviation score estimators (Allison, 1994; Lockwood & McCaffrey, 2007; Wooldridge, 2010, 2012). Instead of manually computing the deviation score from data in wide format (one row per unit), they restructure the data into long format (two rows per unit, one for the pretest and the other for the posttest in our setup) and add a unit ID variable. Then, the causal effect can be obtained from the restructured data by running a pooled regression that includes a dummy variable for each unit:

$$\hat{Y}_{it} = bA_{it} + cID_i + dT_t \tag{16}$$

where ID_i is the set of N individual dummy variables and T_t is the time indicator (i.e., $T_1 = 0$ for the pretest; $T_2 = 1$ for the posttest).[8] Equation (16) does not include an intercept because N (not $N - 1$) dummies are used. The resulting pooled regression estimator b is then identical to the deviation score estimator regardless of the number of time points (the treatment is automatically time-demeaned in dummy variable regression estimators). The causal meaning of the dummy

[8] The inclusion of the time indicator is necessary because we allow for $E(\varepsilon_{i1}) \neq E(\varepsilon_{i2})$.

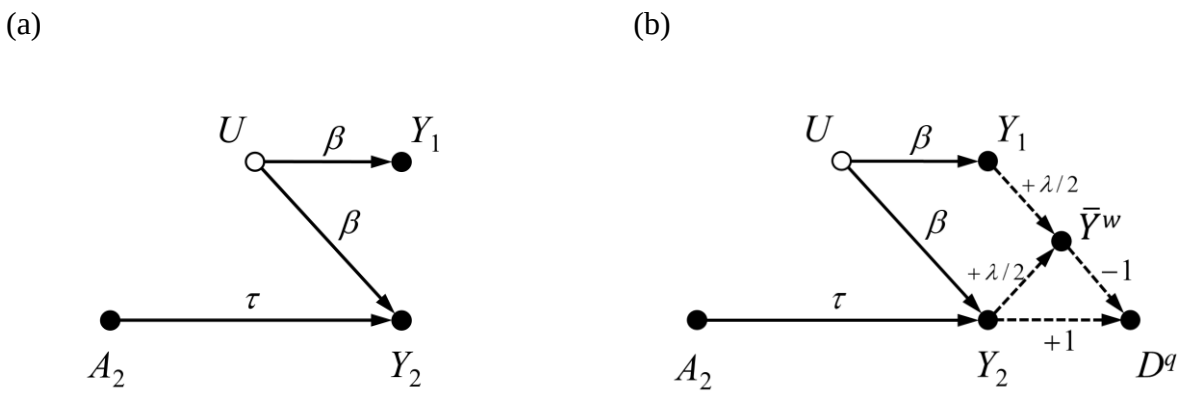

FIGURE 3. *Causal graphs of the RE data-generating model (a) and the quasi-deviation score model (b). Solid arrows represent the data-generating model and dashed arrows represent the computation of gain and deviation scores.*

variable regression estimator can be derived from our graph in Figure 2b because the deviation score estimators and the dummy variable regression estimator are identical. Spurious associations along computationally induced non-causal paths offset the confounding bias in both estimators.

GRAPHICAL REPRESENTATIONS OF RANDOM EFFECTS MODELS

Random Effects Data-Generating Models

The FE data-generating model in Equation (1), $Y_{it} = \tau A_{it} + \theta_i + \varepsilon_{it}$, can also be viewed as a RE data-generating model if one restriction is made (Wooldridge, 2010): While FE data-generating models allow for $Cov(A_{it}, \theta_i) \neq 0$, RE data-generating models assume that the random effects θ_i are independent of treatment exposure, $Cov(A_{it}, \theta_i) = 0$, which is referred to as the *random effects assumption* (Bell & Jones, 2015). Again, using $\theta_i = \beta U_i$, the RE data-generating model can be represented by the causal graph in Figure 3a. Note that the RE

assumption $Cov(A_{it}, \theta_i) = 0$ implies $Cov(A_{it}, U_i) = 0$ (because $\theta_i = \beta U_i$), which is graphically encoded as the absent arrow $U \rightarrow A_2$. The graphical model in Figure 3a clearly shows the causal implication of the RE assumption: there is no unmeasured confounding between the treatment and the posttest, which is unlikely in many observational settings. In fact, the graph is identical to the graph for a randomized controlled trial because the treatment variable A_2 is independent of any variables that causally affect the posttest (Steiner et al., 2017). The comparison between Figures 1b and 3a shows that FE and RE models presume completely different data-generating processes regarding the impact of *U* on A_2.

Quasi-Deviation Score Estimator

In practice, RE estimators are implemented using standard statistical packages like R, HLM, or Mplus. Using data in long format, one specifies the same analytic model as for the dummy variable regression in Equation (16) but ID is treated as a random effect. The resulting RE estimators' causal meaning can be intuitively understood with the causal graph in Figure 3b, where the analytic structure of RE estimator is represented with dashed arrows. The computational structure is identical to the structure of the deviation score model in Figure 2b, but the path coefficients on the two arrows, $Y_1 \dashrightarrow \bar{Y}^w$ and $Y_2 \dashrightarrow \bar{Y}^w$, are different. Now we have path coefficients of $+\lambda/2$ instead $+1/2$. The parameter λ is what has become known as *reliability* of the repeated measurements in the multilevel models literature, which results in the so-called *shrinkage estimator* (Raudenbush & Bryk, 2002). Multilevel researchers conceive of the observed average of the repeated measures Y_1 and Y_2 for each unit *i* as consisting of two parts: *i*'s true unknown score and the random error. The reliability for unit *i* quantifies "the ratio of the

true score […] variance, relative to the observed score or total variance of the sample mean" (Raudenbush & Bryk, 2002, p. 46), and is formally given by

$$\lambda_i = \frac{Var(\theta_i)}{Var(\theta_i) + Var(\varepsilon_{it})/n_i}, \tag{17}$$

where n_i denotes the number of repeated measures for unit *i*. Since $n_i = 2$ for every unit in our simple pretest-posttest design (with no missing values), all units have the same reliability λ.

Note that in the graph, the average node is referred to as $\bar{Y}$ instead of $\bar{Y}^w$, and the resulting deviation score is referred to as D^q where the superscript *q* denotes the *quasi-deviation score*. Applying the RE estimator, the reliability λ automatically enters the weights in computing unit-specific averages $\bar{Y}^w$ of repeated outcome measures, $\bar{Y}^w = \lambda(Y_1 + Y_2)/2$. This weighted average is then used to compute the quasi-deviation score $D^q = Y_2 - \bar{Y}^w$. Wooldridge (2010) clarified this computation procedure and referred to it as *quasi-time demeaning* (see also Lockwood & McCaffrey, 2007). We refer to the resulting estimator as the *quasi-deviation score estimator*. The expectation of the quasi-deviation score estimator *b* of the regression $\widehat{D}^q = a + bA_2$ is given by

$$E(b) = \frac{Cov(A_2, D^q)}{Var(A_2)} = \tau\left(\frac{2-\lambda}{2}\right). \tag{18}$$

Thus, to recover the original causal quantity, the quasi-deviation score estimator requires a multiplicative factor of $2/(2-\lambda)$ in the two-period data. But, like deviation score estimators, this computation is not necessary if the treatment variable is also quasi-demeaned.

The augmented graph in Figure 3b highlights two important aspects of RE models. First, according to the RE data-generating model, *U* is not a confounding variable. The absence of

backdoor paths connecting A_2 and D^q, guarantees an unbiased quasi-deviation score estimator (but also unbiased gain score and deviation score estimators). However, as will be discussed in the next section, if the RE assumption, $Cov(A_{it}, U_i) = 0$, is not met, RE estimators, that is, quasi-deviation score estimators will be biased. Second, as for FE estimators, quasi-deviation score estimators do not condition on the pretest measure to block any potential confounding path (when the RE assumption is not met) but use the pretest to create non-causal paths. This offsetting mechanism is similar as in the deviation score model in Figure 2b but the path coefficients are different and this results in different point estimates between FE and RE estimators.

BIAS IN GAIN, DEVIATION, AND QUASI-DEVIATION SCORES ESTIMATORS

Every estimator is biased if its causal identification assumptions are violated. This section discusses how FE and RE estimators—gain score, deviation score, and quasi-deviation score estimators—become biased when the FE and RE assumptions are not met.

Data-Generating Models for an Observational Study

The graph in Figure 4a describes the data-generating model for a general pretest-posttest study with an unmeasued counfounder *U*. Comparing Figure 4a and Figure 1b, the impacts of *U* on Y_1 and Y_2 are no longer assumed to be identical. That is, *U*'s impact on the pretest and posttest is now denoted by separete parameters β_1 and β_2, respectively. This indicates that the FE assumption is violated in Figure 4a unless $\beta_1 = \beta_2$. Further, the comparison of Figure 4a and Figure 3a reveals that the RE assumption is not met either unless $\alpha = 0$ (then the arrow from *U* to

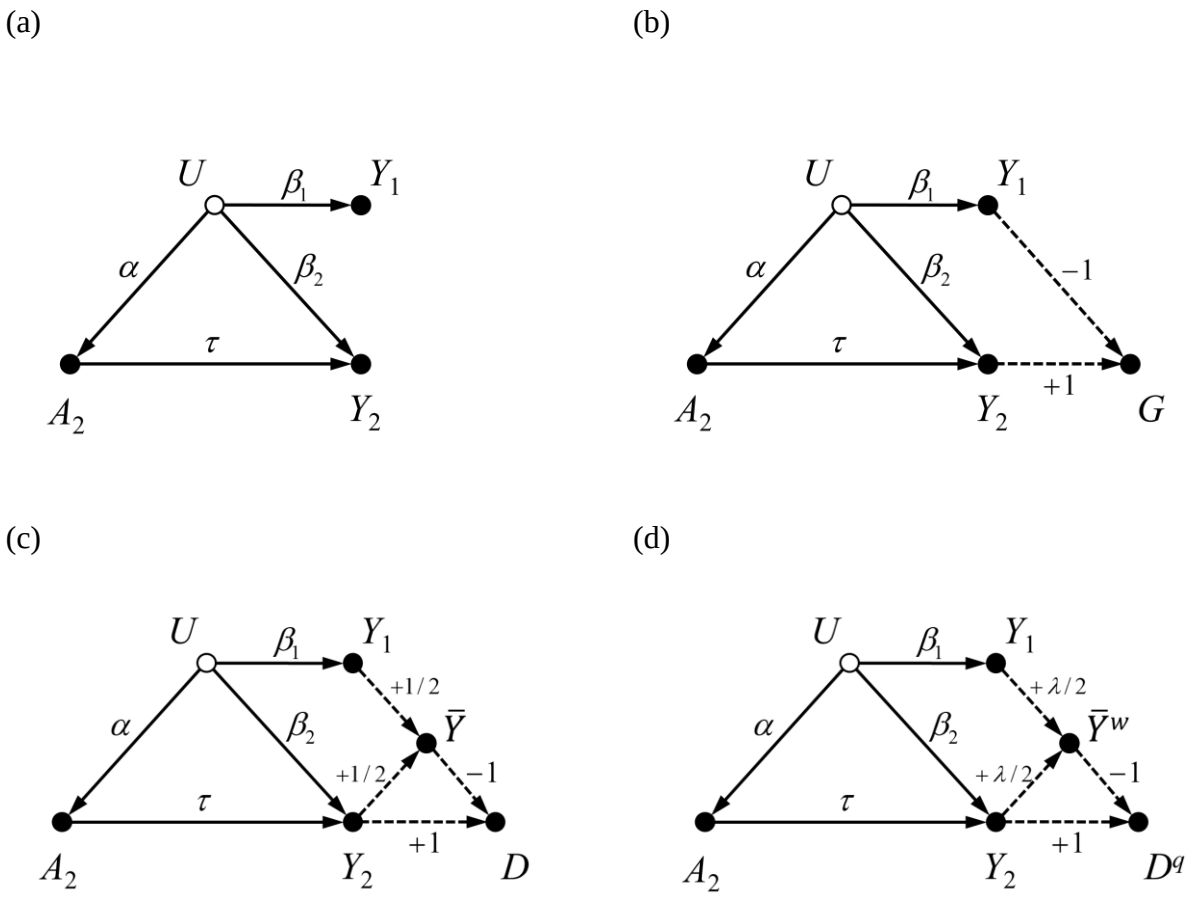

FIGURE 4. *Causal graphs for a general data-generating model where the FE and RE assumptions are violated (a), the gain score model (b), the deviation score model (c), and the quasi-deviation score model (d).*

A_2 would be absent). Such a case where both the FE and RE assumptions are not met is realistic for many observational studies.

Bias in Three Estimators

Given the data-generating model in Figure 4a, the analytic structure of the gain score estimator is graphically represented in Figure 4b. Due to the inequality between β_1 and β_2, that

is, the violation of the FE assumption, the non-causal association transmitted via the path $A_2 \leftarrow U \rightarrow Y_1 \dashrightarrow G$ cannot perfectly offset the non-causal association transmitted via the path $A_2 \leftarrow U \rightarrow Y_2 \dashrightarrow G$, and thus bias remains. The gain score estimator b of the regression of G on A_2, $\hat{G}_i = a + bA_{i2}$, is biased:

$$\begin{aligned} E(b) &= \frac{Cov(A_2, G)}{Var(A_2)} \\ &= \frac{Var(A_2)\tau + Var(U)\alpha(\beta_2 - \beta_1)}{Var(A_2)} \\ &= \tau + v(\beta_2 - \beta_1) \end{aligned} \tag{19}$$

where $v = Var(U)\alpha / Var(A_2)$. Again, the covariance $Cov(A_2, G)$ can be easily computed by the path-tracing rules from the causal graph in Figure 4b. Unless $\beta_1 = \beta_2$, the gain score estimator is biased because the second term of Equation (19) does not vanish. The extent to which the two parameters differ determines the magnitude of the actual bias of the gain score estimator.

A similar imperfect offsetting occurs in the deviation score estimator. According to the corresponding graph in Figure 4c, the deviation score estimator creates two additional non-causal paths, $A_2 \leftarrow U \rightarrow Y_1 \dashrightarrow \bar{Y} \dashrightarrow D$ and $A_2 \leftarrow U \rightarrow Y_2 \dashrightarrow \bar{Y} \dashrightarrow D$, in order to offset the non-causal path $A_2 \leftarrow U \rightarrow Y_2 \dashrightarrow D$. However, since the FE assumption does not hold (i.e., $\beta_1 \neq \beta_2$), the confounding bias is not fully offset. The expectation of the regression coffieicient for A_2 ($\widehat{D}_i = a + bA_{i2}$) is given by:

$$
\begin{aligned}
E(b) &= Cov(A_2, D)/Var(A_2) \\
&= \frac{.5Var(A_2)\tau + Var(U)\alpha\{\beta_2 - (\beta_1 + \beta_2)/2\}}{Var(A_2)} \\
&= \frac{\tau}{2} + v\left(\frac{\beta_2 - \beta_1}{2}\right).
\end{aligned} \tag{20}
$$

Since the treatment variable was not demeaned, it must be multiplied by two to recover the original causal quantity. Then, the expectation of the estimator is identical to Equation (19). This equality between the gain score estimator and the deviation score estimator in a simple pretest-posttest study has been well known in the literature (e.g., Wooldridge, 2010, 2012). It is beyond the scope of this article, but if the number of repeated outcome measures are greater than two, the gain score and the deviation score estimators will generally differ.

Finally, given the same data-generating model, the graphical representation of the quasi-deviation score estimator is described in Figure 4d. The imperfect bias-offsetting occurs for two reasons: First, due to the inequality $\beta_1 \neq \beta_2$ and, second, due to the reliability parameter λ. Regressing D^q on A_2, the expectation of the regression coffieicient for A_2 ($\widehat{D}_i^q = a + bA_{i2}$) is

$$
\begin{aligned}
E(b) &= Cov(A_2, D^q)/Var(A_2) \\
&= \tau\left(\frac{2 - \lambda}{2}\right) + v\left\{\beta_2 - \frac{\lambda(\beta_1 + \beta_2)}{2}\right\},
\end{aligned} \tag{21}
$$

and multiplying Equation (21) by $2/(2 - \lambda)$ (because the treatment was not quasi-demeaned in Figure 4d) results in

$$E(b) \times \frac{2}{2-\lambda} = \tau + v\left\{\beta_2 - \left(\frac{\lambda}{2-\lambda}\right)\beta_1\right\}. \quad (22)$$

If $\lambda = 1$, Equation (22) is identical to Equation (19). Otherwise, even if $\beta_1 = \beta_2$, the bias term does not vanish, therefore, the quasi-deviation score estimator is generally biased under the data-generating model described in Figure 4a.[9] Figure 4d provides the intuition behind RE or quasi-deviation score estimators and reveals the similarity and difference between FE and RE estimators.

One might believe that quasi-deviation score estimators are always more biased than gain or deviation score estimators because, in addition to $\beta_1 \neq \beta_2$, a reliability λ of less than one results in an imperfect bias-offsetting by the quasi-deviation score estimator. However, depending on the data-generating parameters, FE estimators can be more biased than RE estimators (i.e., quasi-deviation score estimator). Using Equations (19) and (22) with $v = 1$ and $\beta_1 = 1$, Figure 5a compares the absolute bias of FE and RE estimators. The two dark areas in the top part of Figure 5a indicate that the absolute bias of RE estimators is greater than the bias of FE estimators, $|Bias(RE)| > |Bias(FE)|$ (the darker shade indicates $|Bias(RE)| > 2|Bias(FE)|$). In contrast, the two light grey areas in the bottom part of the plot indicate that the absolute bias of FE estimators is greater than the absolute bias of RE estimators, $|Bias(FE)| > |Bias(RE)|$ (the lighter shade indicates $|Bias(FE)| > 2|Bias(RE)|$). The light grey areas demonstrate that RE estimators can be less biased than FE estimators despite the imperfect

[9] There are some solutions for RE models. In our setup, one can include the interaction term between the treatment and the time indicator in the RE model such as $\hat{Y}_{it} = bA_{it} + cID_i + dT_t + k(A_{it} \times T_t)$. Or, more generally, Mundlak's (1978) method or Allison's (2009) hybrid RE estimators will allow researchers to have the same point estimates as FE estimators.

(a) (b)

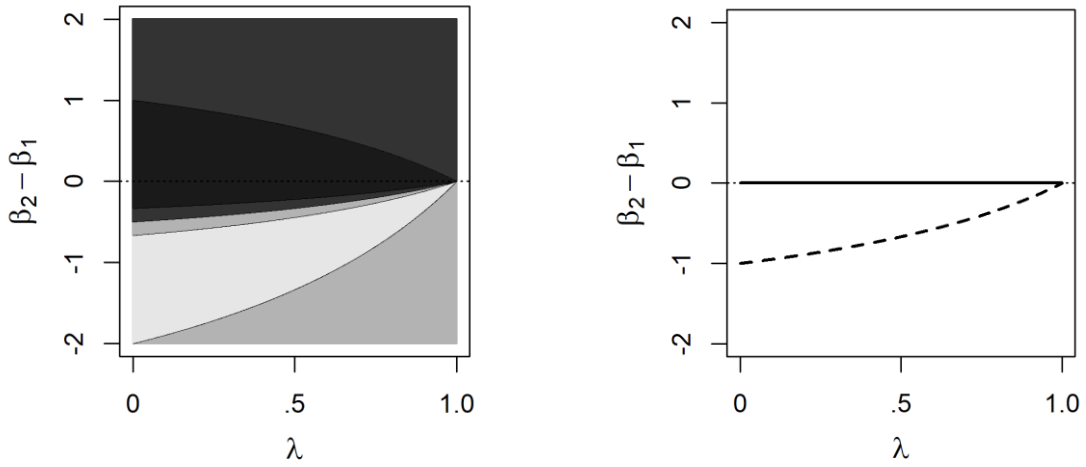

FIGURE 5. *The left plot presents relative bias comparison between FE and RE estimators. The two dark grey areas in the top part of plot (a) indicate that the absolute bias of the RE estimators is (more than two times, for the darker shade) greater than the absolute bias of FE estimators. The two light grey areas in the bottom part indicate that the absolute bias of FE estimators is (more than two times, for the lighter shade) greater than the absolute bias of RE estimators (a). The right plot displays unbiased cases for both FE (solid line) and RE (dashed line) estimators (b).*

offsetting due to the reliability λ. As demonstrasted in Figure 5b, it is even possible that RE estimators are unbiased when both FE and RE assumptions are not met (see the dahsed line in Figure 5b). This is because the imperfect offsetting due to the reliability λ accidentally compensates the imperfect offsetting due to the inequality $\beta_1 \neq \beta_2$. However, without knowledge of the underlying data-generating model, including the magnititude of β_1 and β_2 and the reliabtliy λ, it is impossible to conclude whether FE or RE models offset more confounding bias in practice.

CONCLUSION

This article developed causal graphs to represent the presumed data-generating process and the analytic structure of different estimators for FE and RE models. The proposed graphical representations of FE and RE estimators revealed the difference to standard regression or matching estimators that explicitly condition on the pretest (and other covariates) (Kim & Steiner, 2019). While FE and RE estimators offset confounding bias by creating new non-causal paths and associations, standard regression and matching estimators aim at blocking the confounding paths by conditioning on observed confounders. Thus, with unobserved confounding, standard regression and matching estimators will not be able to remove the entire confounding bias (i.e., backdoor criterion), but FE estimators can be unbiased if the FE assumption ($\beta_1 = \beta_2$) holds because all the biases will be cancelled out. If the FE assumption is not believed to hold, then only strong subject-matter knowledge about the data-generating process can aid researchers in making an informed decision about whether an FE estimator or a standard regression or matching estimator will remove more bias. Indeed, this is a setup that has been known as Lord's (1967) paradox where, given the same data, the standard regression (ANCOVA) estimator and the gain score estimator can result in opposite conclusions.

The causal graphs we presented help researchers to better assess the FE and RE assumptions for a given pretest-posttest study, but also to better understand the similarities and differences between the bias-removing mechanisms of FE and RE estimators. All the estimators—gain score, deviation score, and quasi-deviation score estimators—address confounding by deliberately creating new non-causal paths that induce spurious association of opposite sign to offset the confounding bias. Also, the comparison of the graphs for the FE and RE data-generating models highlighted the restrictiveness of the RE assumption for the quasi-

deviation score estimator of a standard RE model (i.e., no unmeasured confounding) and the inherent imperfect offsetting due to the reliability. Nonetheless, in real observational studies RE estimators may be less biased than FE estimators, depending on the data-generating process. Thus, without strong subject-matter knowledge about how the data were generated, it is generally difficult to compare FE and RE estimators in terms of bias.

To focus on the causal assumptions and basic bias-removing mechanisms of FE and RE estimators, we used a simple linear setup with constant effects and a single unobserved time-invariant confounder. But, our discussion of FE and RE estimators can be generalized to scenarios with multiple unobserved time-invariant confounders. The FE assumption then requires that the overall effect of all confounders on the pretest is the same as the overall effect of the same confounders on the posttest (but the effect of a single confounder on the pre- and posttest no longer needs to be the same). Useful graphical discussions are also possible when other observed time-varying or time-invariant covariates need to be considered, or when the causal effects of flexible treatment regimes with more than two repeated measurements are studied. They are topics for future studies.

REFERENCES

Allison, P. D. (1994). Using panel data to estimate the effects of events. *Sociological Methods & Research, 23*(2), 174-199.

Allison, P. D. (2009). *Fixed effects regression models* (Vol. 160). Thousand Oaks, CA: SAGE publications, Inc.

Ashenfelter, O., & Krueger, A. B. (1994). Estimates of the economic return to schooling from a new sample of twins. *American Economic Review, 84*(5), 1157-1173.

Bell, A., & Jones, K. (2015). Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods, 3*(1), 133-153.

Bollen, K. A., & Brand, J. E. (2010). A general panel model with random and fixed effects: A structural equations approach. *Social Forces, 89*(1), 1-34.

Clark, T. S., & Linzer, D. A. (2015). Should I use fixed or random effects? *Political Science Research and Methods, 3*(2), 399-408.

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology, 40*, 31-53.

Gelman A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Greenland, S., & Mansournia, M. A. (2015). Limitations of individual causal models, causal graphs, and ignorability assumptions, as illustrated by random confounding and design unfaithfulness. *European Journal of Epidemiology, 30*(10), 1101-1110.

Halaby, C. (2004). Panel models in sociological research: Theory into practice. *Annual Review of Sociology, 30*, 507–544.

Imai, K., & Kim, I. S. (2019). When should we use unit fixed effects regression models for for causal inference with longitudinal data? *American Journal of Political Science*. Advance online publication. doi:10.1111/ajps.12417

Imai, K., & Kim, I. S. (2020, April). *On the use of two-way fixed effects regression models for causal inference with panel data*. Retrieved from https://imai.fas.harvard.edu/research/twoway.html

Kim, Y., & Steiner, P. M. (2019). Gain scores revisited: A graphical models perspective. *Sociological Methods & Research.* Advance online publication. doi:10.1177/0049124119826155

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics, 4*(3), 165-224.

Lockwood, J. R., & McCaffrey, D. F. (2007). Controlling for individual heterogeneity in longitudinal models, with applications to student achievement. *Electronic Journal of Statistics, 1*, 223-252.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 68, 304-305.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica, 46*(1), 69-85.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo: Morgan Kaufman.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). New York, NY: Cambridge University Press.

Pearl, J. (2013). Linear models: A useful "microscope" for causal analysis. *Journal of Causal Inference, 1*, 155–170.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. New York, NY: Wiley.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: SAGE publications, Inc.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shpitser, I., VanderWeele, T. J., & Robins, J. M. (2010). On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the 26th conference on Uncertainty and Artificial Intelligence* (pp. 527–536). Corvallis: AUAI Press.

Spirtes, P., Glymour, C. N., & Scheines, R. (2001). *Causation, prediction, and search* (2nd ed.). Cambridge, MA: MIT Press.

Steiner, P. M., & Kim, Y. (2016). The mechanics of omitted variable bias: bias amplification and cancellation of offsetting biases. *Journal of Causal Inference, 4*(2). doi: 10.1515/jci-2016-0009

Steiner, P. M., Kim, Y., Hall, C. E., & Su, D. (2017). Graphical models for quasi-experimental designs. *Sociological Methods & Research, 46*(2), 155-188.

Thoemmes, F., & Mohan, K. (2015). Graphical representation of missing data problems. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(4), 631-642.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: The MIT Press.

Wooldridge, J. M. (2012). *Introductory econometrics: A modern approach* (5th ed.). South-Western Cengage Learning, Mason.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20*(7), 557-585.

APPENDIX A: REGRESSION ESTIMATOR FORMULA

The population partial regression coefficient for A_2 of the regression model of Y_2 on A_2 and Y_1 can be expressed with bivariate correlations:

$$b_{Y_2A_2|Y_1} = \frac{\rho_{Y_2A_2} - \rho_{Y_2Y_1}\rho_{A_2Y_1}}{1 - \rho_{AY_1}^2} \times \frac{SD(Y_2)}{SD(A_2)}.$$

Given $Y_1 = \theta + \varepsilon_1$ and $Y_2 = \tau A_2 + \theta + \varepsilon_2$, the bivariate correlations are given by

$$\rho_{Y_2A_2} = \{Var(A_2)\tau + Cov(A_2, \theta)\}/\{SD(Y_2)SD(A_2)\},$$

$$\rho_{A_2Y_1} = Cov(A_2, \theta)/\{SD(A_2)SD(Y_1)\},$$

$$\rho_{Y_2Y_1} = \{\tau Cov(A_2, \theta) + Var(\theta)\}/\{SD(Y_2)SD(Y_1)\}.$$

Plugging the population correlations into the formula of $b_{Y_2A_2|Y_1}$ above, we obtain Equation (5).

APPENDIX B: GRAPHICAL REPRESENTATION OF DEVIATION SCORE MODELS WITH THE TIME-DEMEANED TREATMENT

For deviation score models, one may also time-demean the treatment variable. As all units are in the control condition at $t = 1$ ($A_1 = 0$), the average of the two repreated treatments is half of the treatment at $t = 2$: $\bar{A} = A_2/2$. Then, the demeaned treatment is $C = A_2 - \bar{A}$. This computation is added in Figure A using dahsed arrows. If one regresses the time-demeaned outcome (i.e., deviation score) *D* on the time-demeaned treatment C, the total association is transmitted via the folloiwng ten paths:

(1) $C \dashleftarrow \bar{A} \dashleftarrow A_2 \leftarrow U \rightarrow Y_1 \dashrightarrow \bar{Y} \dashrightarrow D,\ Var(U)\alpha\beta_1/4;$

(2) $C \dashleftarrow \bar{A} \dashleftarrow A_2 \leftarrow U \rightarrow Y_2 \dashrightarrow \bar{Y} \dashrightarrow D,\ Var(U)\alpha\beta_2/4;$

(3) $C \dashleftarrow \bar{A} \dashleftarrow A_2 \leftarrow U \rightarrow Y_2 \dashrightarrow D,\ -Var(U)\alpha\beta_2/2;$

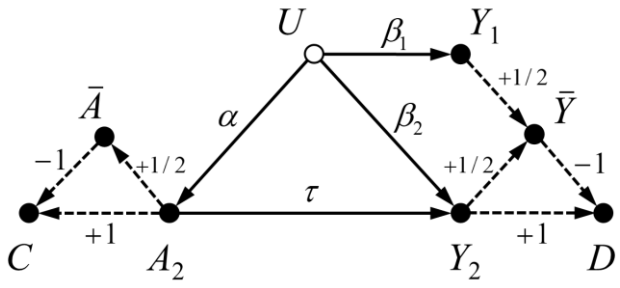

FIGURE A. *Causal graph, combined the data-generating model (solid arrows) and the analytical model (dashed arrows), for the deviation score model with the time-demeaned treatment C.*

(4) $C \dashleftarrow A_2 \leftarrow U \rightarrow Y_1 \dashrightarrow D,\ -Var(U)\alpha\beta_1/2;$

(5) $C \dashleftarrow A_2 \leftarrow U \rightarrow Y_2 \dashrightarrow \bar{Y} \dashrightarrow D,\ -Var(U)\alpha\beta_2/2;$

(6) $C \dashleftarrow A_2 \leftarrow U \rightarrow Y_2 \dashrightarrow D,\ Var(U)\alpha\beta_2;$

(7) $C \dashleftarrow \bar{A} \dashleftarrow A_2 \rightarrow Y_2 \dashrightarrow \bar{Y} \dashrightarrow D,\ Var(A_2)\tau/4;$

(8) $C \dashleftarrow \bar{A} \dashleftarrow A_2 \rightarrow Y_2 \dashrightarrow D,\ -Var(A_2)\tau/2;$

(9) $C \dashleftarrow A_2 \rightarrow Y_2 \dashrightarrow \bar{Y} \dashrightarrow D,\ -Var(A_2)\tau/2;$

(10) $C \dashleftarrow A_2 \rightarrow Y_2 \dashrightarrow D,\ Var(A_2)\tau.$

The sum of the path-specific associations via paths (1) to (6) is $Var(U)\alpha(\beta_2 - \beta_1)/4$, therefore, becomes zero under $\beta_1 = \beta_2$. The sum of the path-specific associations via paths (7) to (10) is $Var(A_2)\tau/4$, which becomes the covariance between C and D. As the regression coefficient of C from regressing D on C is $Cov(C, D)/Var(C)$ and $Var(C) = Var(A_2)/4$, the coefficient equals τ.