Editors' Suggestion

Critical issues in statistical causal inference for observational physics education research

Vidushi Adlakha and Eric Kuo
University of Illinois Urbana-Champaign, Urbana, Illinois 61801, USA

(Received 23 May 2023; accepted 9 October 2023; published 20 November 2023)

Recent critiques of physics education research (PER) studies have revoiced the critical issues when drawing causal inferences from observational data where no intervention is present. In response to a call for a "causal reasoning primer" in PER, this paper discusses some of the fundamental issues in statistical causal inference. In reviewing these issues, we discuss well-established causal inference methods commonly applied in other fields and discuss their application to PER. Using simulated data sets, we illustrate (i) why analysis for causal inference should control for confounders but not control for mediators and colliders and (ii) that multiple proposed causal models can fit a highly correlated dataset. Finally, we discuss how these causal inference methods can be used to represent and explain existing issues in quantitative PER. Throughout, we discuss a central issue in observational studies: A good quantitative model fit for a proposed causal model is not sufficient to support that proposed model over alternative models. To address this issue, we propose an explicit role for observational studies in PER that draw statistical causal inferences: Proposing future intervention studies and predicting their outcomes. Mirroring the way that theory can motivate experiments in physics, observational studies in PER can predict the causal effects of interventions, and future intervention studies can test those predictions directly.

DOI: 10.1103/PhysRevPhysEducRes.19.020160

I. INTRODUCTION

Recent critiques of physics education research (PER) studies [1,2] have revoiced the critical issues when drawing causal inferences from observational data where no intervention is present. In response to a call for a "causal reasoning primer" in PER [1], this paper discusses some of the fundamental issues underlying statistical causal inference. In reviewing these issues, we discuss well-established causal inference methods commonly employed in other fields [3–13] and discuss their application to PER. We suspect that many physics education researchers who engage in quantitative analysis will be familiar with these methods. At the same time, we propose that more widespread knowledge of these causal inference methods can help establish greater consensus in the PER community on how to establish causal relationships from quantitative data. The causal inference methods we present provide a powerful set of conceptual and mathematical tools for analysis and make clear the potential causal misinterpretations and biases that can be introduced during analysis. For readers interested in a more in-depth discussion of the causal methods discussed here, there are both more popular [14] and technical references [15–22] available.

II. CAUSAL VS PREDICTIVE MODELING

In this paper, to discuss causal modeling, we will consider the special case of path analysis using multiple linear regression on standardized variables. We chose linear regression because it is a standard method that we expect many readers to be familiar with. Although many of the quantitative details of our discussion will be particular to multiple linear regression, the causal issues we illustrate extend to other analytic methods as well (e.g., structural equation modeling).

Consider the case where three standardized variables, X, Y, and Z, are measured and multiple regression is performed with Z as the dependent variable and X and Y as independent variables (an analysis denoted as $Z \sim X + Y$). This best-fit linear model produced by this analysis is $Z = \beta_{XZ}X + \beta_{YZ}Y$ (note: there will be a nonzero constant term β_0 if the variables are not all standardized). Conceptually, this analysis is commonly interpreted as finding the variance explained by one independent variable while controlling for another (i.e., finding the regression coefficient of X on Z, β_{XZ}, when controlling for Y). For this regression analysis, β_{XZ} is

$$\beta_{XZ} = \frac{r_{XZ} - r_{XY}r_{YZ}}{1 - r_{XY}^2}, \tag{1}$$

where r_{AB} is the bivariate linear correlation between A and B. A conceptually important limiting case is that when $r_{XY} = 0$, $\beta_{XZ} = r_{XZ}$. This indicates that controlling for the independent variable Y has no effect on the association between independent variable X and dependent variable Z if the two independent variables are not correlated.

The interpretation and appropriateness of the analysis $Z \sim X + Y$ will depend on whether the goal of this regression model is predictive or causal [23]. For a predictive model [24], the goal would be to explain the most variance in Z with other measured variables—that is, to reduce the error in predicting Z. An example from the field of education would be using early pre- and in-semester measures X and Y to predict students' final physics course grade Z [25–27]. Establishing this predictive model using data from previous semesters may allow instructors and researchers to identify which students are at risk of failing a course early enough to provide additional support. In a predictive model, it is sensible to include as many variables as available to improve the R^2 of the model. The βs indicate which variables explain the most variance in Z that is not explained by other variables in the model—the variance explained by one independent variable controlling for all others. In a predictive model, it does not matter if X and Y are causes of, effects of, or noncausally associated with Z; so long as X and Y are associated with Z, they can be used as predictors.

By contrast, if the goal of the analysis $Z \sim X + Y$ were causal modeling [14], this analysis would aim to estimate the causal impact of how intervening on X and Y should affect Z. That is, β_{XZ} would indicate the "direct effect" of X on Z: how Z would change if X were increased by one standard deviation and Y were held constant. However, the accuracy of β_{XZ} as a causal estimate depends critically on whether the analysis $Z \sim X + Y$ reflects the actual causal relations between X, Y, Z, and other unmeasured variables. The analysis $Z \sim X + Y$ is aligned with a causal model where X and Y are causal factors that act on the effect Z. The key conceptual prerequisite for causal inference is to specify the causal model. This must be done from a conceptual understanding of causal mechanisms and cannot be determined solely by the quantitative fits of different models onto observational data. If the causal model is incorrect, then the causal implications of the regression coefficients determined by this analysis will be misinterpreted by the researcher.

To illustrate how the predictive and causal inference goals of statistical modeling can be misaligned, consider the case where X is a student's midsemester score in their calculus course. In this case, although midsemester calculus grades may help predict final physics course grade, interventions to improve midsemester calculus grades may not improve final course grade. For instance, an intervention that increases time spent on calculus study might actually reduce the time available for studying physics, causing no improvement or even decreasing students' physics final grade. In reality, midsemester calculus grades may serve as a proxy indicator of the causal role of students' more general math preparation or general study practices on their physics course grade. Although a predictive model does not necessarily care why X explains variance in Z, the causal details of why X explains variance in Z are critical for making accurate causal estimates and effectively intervening on student outcomes.

The rest of this paper elaborates on causal inference techniques for determining the appropriate analysis for estimating the causal impacts of one variable on another when many variables are correlated together. Central to these methods are diagrams that embody a theoretical model of the cause-effect relationships between variables. After describing three fundamental causal structures, we will use simulated datasets to illustrate two issues related to causal inference. First, the proposed causal structure between variables determines whether or not it is appropriate to control for other variables in causal inference. Second, multiple proposed causal structures can quantitatively fit the same dataset, and quantitative indices of statistical model fit or nonzero path coefficients cannot validate one proposed causal structure over another. Ultimately, the statistical causal inferences made are only as valid as the proposed conceptual causal structure between variables.

III. THREE FUNDAMENTAL CAUSAL STRUCTURES: CHAIN, FORK, COLLIDER

Relations between quantitative variables can be represented through directed acyclic graphs (DAGs) [19,28,29], which represent the variables as nodes connected by directed arrows. DAG-like diagrams are commonly used to represent the results of path analysis or structural equation modeling. When the DAG is constructed to reflect a proposed causal model, the arrows indicate the direction of causality between variables, and coefficients associated with each arrow reflect the direct causal impact of changing one variable on another.

For instance, $X \rightarrow Y$ is a causal model where X has a causal impact on Y—that is, intervening to change X will produce a change in Y, and that intervening to change Y directly (i.e., through a method besides changing X) will *not* change X. The analysis $Y \sim X$ would produce the coefficient of the linear equation $Y = \beta_{XY}X$, and the coefficient β_{XY} would be associated with the path connecting X to Y. This equation (and diagram) also represents a quantitative causal prediction: that changing X by ΔX will change Y by $\beta_{XY}\Delta X$.

There are three fundamental causal structures—chain, fork, and collider [14]—through which more complicated causal models can be constructed. These three structures illustrate the ways in which correlation may or may not reflect causation and also the different rules for whether to control for other variables in statistical causal inference.

A. Chain

A causal chain is represented as $X \rightarrow Y \rightarrow Z$ [Fig. 1(a)]. This chain represents a causal mediation where X causes Z through the mediator Y: X causes Y and Y causes Z, so therefore, X causes Z. An everyday example of a causal chain is $Fire \rightarrow Smoke \rightarrow Alarm$. Here, smoke is the mediator caused by fire and causes the smoke alarm to sound. In principle, any single causal link can be modeled as a chain by explicitly breaking down the causal mechanism into mediators. In practice, mediators are commonly omitted from causal diagrams if they are not measured and/or are not of theoretical interest.

The path coefficients of the chain $X \rightarrow Y \rightarrow Z$ are associated with two linear regressions, $Y = \beta_{XY}X$ (associated with $X \rightarrow Y$) and $Z = \beta_{YZ}Y$ (associated with $Y \rightarrow Z$). Because X, Y, and Z are standardized variables and the regressions only have a single-independent variable, $\beta_{XY} = r_{XY}$ and $\beta_{YZ} = r_{YZ}$. There are two ways to determine the (indirect) causal impact of X on Z. The first is the chain rule: changes in X cause changes in Y, and these changes in Y cause changes in Z, so the total effect of X on Z is $r_{XY}r_{YZ}$. The second is the analysis $Z \sim X$. In this analysis, the coefficient for X will equal the total effect $r_{XY}r_{YZ}$, and, given that there exist no other pathways associating X with Z except mediation through Y, this value is equivalent to r_{XZ}.

In the idealized chain $X \rightarrow Y \rightarrow Z$, controlling for Y will block the causal relationship between X and Z. Analyzing the relationship of X on Z while controlling for Y can be accomplished through the analysis $Z \sim X + Y$ which would yield a coefficient for Y of r_{YZ} and a coefficient for X of zero. This can be intuitively understood through the fire alarm example: controlling for the mediator "smoke" blocks the relationship between fire and alarm. We could physically control smoke by very efficiently removing smoke from a room with a fume hood. In this case, there will be no smoke in the room, whether or not a fire is present, and the alarm will not sound. We could also hold smoke constant by filling the room with smoke using a fog machine. In this case, the alarm will sound whether or not fire is present. By holding the presence or absence of smoke constant, the causal link between fire and the alarm sounding is broken. Therefore, controlling for the mediator Y will screen off information about the actual, indirect causal relationship of X on Z. To determine the causal impact of X on Z, one should not control for the mediator Y.

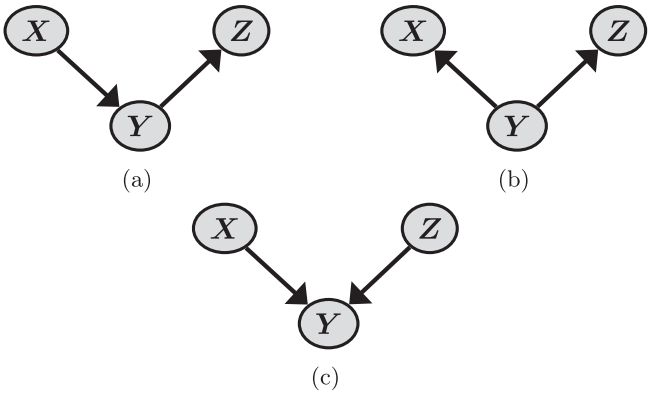


FIG. 1. Fundamental causal structures: (a) chain—where Y acts as a mediator, (b) fork—where Y acts as a confounder, and (c) collider—where Y acts as a collider.

Note that, even in this relatively simple case, the correct causal interpretation depends critically on having the correct causal diagram. If the causal chain were actually $X \leftarrow Y \leftarrow Z$, then the causal coefficient $r_{XY}r_{YZ}$ would actually represent how changing Z would impact X, not how changing X would change Z. Though it is often theoretically clear which factor is the cause and which is the effect, there are systems where determining causes and effects is nontrivial.

B. Fork

A fork is represented as $X \leftarrow Y \rightarrow Z$ [Fig. 1(b)]. Here, Y is a common cause of both X and Z. Therefore, X and Z are correlated because changes in Y will lead to changes in both X and Z, but this correlation does not reflect a causal relationship between X and Z. An everyday example of a causal fork is $Shoe\ Size \leftarrow Age\ of\ Child \rightarrow Reading\ Ability$ [10]. Children with larger shoes tend to read at a higher level because they are older, but the relationship is not one of cause and effect. Giving a child larger shoes will not cause their reading ability to increase nor will improving a child's reading ability cause their shoe size to increase.

Here, the causal diagram indicates that the causal impact of Y on X is represented through the equation $X = \beta_{YX}Y = r_{YX}Y$, and the causal impact of Y on Z is represented through $Z = \beta_{YZ}Y = r_{YZ}Y$. The analysis $Z \sim X$ will produce a coefficient for X of $r_{XY}r_{YZ}$, but this indicates a *noncausal* association between X and Z. $r_{XY}r_{YZ}$ reflects how X and Z are correlated through Y, but directly changing X (through a method that does not change Y) will produce no effect on Z.

To determine the correct causal coefficients, one should control for Y, which is a common cause of X and Z. Here, we can control for Y through analysis by only analyzing subsets of the data with the same value for Y. For our everyday example, when analyzing subsets of same-aged children, the remaining variations in shoe size and reading ability should be uncorrelated, reflecting that there is no association after the common cause is controlled for. For the fork $X \leftarrow Y \rightarrow Z$, controlling for Y through the analysis $Z \sim X + Y$ will produce a coefficient for Y of r_{YZ} and a coefficient for X of zero. These coefficients reflect the causal impact of Y and X, respectively, on Z. In causal analysis, a common cause of two variables is called a *confounder* since, if uncontrolled, it confounds our ability to estimate the causal relationship between those two variables by contributing a noncausal association. In the

causal diagram representation, noncausal pathways, such as the one from X to Z through the fork $X \leftarrow Y \rightarrow Z$, are called backdoor paths and controlling for confounders closes these backdoor paths.

Note that the interpretation of which regression analysis yields an accurate estimate of causal impacts depends on the proposed causal structure. For both the chain and the fork discussed, $Z \sim X$ will yield $Z = (r_{XY}r_{YZ})X$, and $Z \sim X + Y$ will yield $Z = (0)X + (r_{YZ})Y$. Which coefficient is the causal coefficient describing how intervening directly on X can change Z: $r_{XY}r_{YZ}$ or zero? For the chain, the correct causal coefficient is $r_{XY}r_{YZ}$. The appropriate causal analysis does not control for the mediator Y since this will mask the actual causal relationship between X and Z. For the fork, the correct causal coefficient is zero. The appropriate causal analysis does control for the confounder Y since this will block the backdoor path that contributes a noncausal association between X and Z. This highlights the critical importance of constructing the correct causal diagram when estimating the causal impacts of one variable on another.

C. Collider

A collider is represented as $X \rightarrow Y \leftarrow Z$ [Fig. 1(c)]. Here, Y is a common effect of both X and Z. In the idealized case depicted, X and Z are uncorrelated ($r_{XZ} = 0$), because there are no direct or backdoor paths connecting them.

Here, the causal diagram indicates that the causal impact of X on Y and Z on Y is represented through the equation $Y = (\beta_{XY})X + (\beta_{ZY})Z$. Because X and Z are uncorrelated in this idealized diagram, $\beta_{XY} = r_{XY}$ and $\beta_{ZY} = r_{ZY}$ [if X and Z were correlated, the βs could be computed with Eq. (1)]. These βs are the correct causal coefficients and indicate how changing X and Z will affect Y. X and Z do not become correlated through a collider, so the analyses $Z \sim X$ and $X \sim Z$ would both yield coefficients equal to zero, which correctly indicates the lack of causal association between them.

Here, controlling for Y will produce a unique noncausal association: the analysis $Z \sim X + Y$ will produce a noncausal coefficient for X of $\frac{-r_{XY}r_{ZY}}{1 - r_{XY}^2}$. That is, in the case that X and Z have positive causal impacts on Y, controlling for Y will produce a negative noncausal association between X and Z. To see why this would be the case, consider an example of *Academic GPA* $\rightarrow$ *College Scholarship* $\leftarrow$ *Athletic Talent* (Fig. 2). This causal diagram reflects that students can receive college scholarships based on either academic achievement or athletic talent (which, for the purposes of this example, we are imagining is uncorrelated with academic achievement). When we consider the subset of students who have been awarded a college scholarship (controlling for the collider), academic GPA will be anticorrelated with athletic talent. For example, if a student with a lower GPA receives a scholarship, it is more likely that they received a scholarship for playing sports, so they are more likely to have more athletic talent. Similarly, students with less athletic talent who received a scholarship are more likely to have a higher academic GPA. Here, controlling for the outcome opens a backdoor path through the collider, revealing an association between causes that is present when considering a same-outcome subgroup but is not present when considering the entire population. Therefore, one should not control for a collider in causal modeling since doing so can open noncausal associations.

FIG. 2. Causal diagram illustrating the relationship between academic GPA, college scholarship, and athletic talent.

IV. ANALYZING A SIMULATED DATASET I: WHEN YOU SHOULD AND SHOULD NOT CONTROL FOR VARIABLES IN CAUSAL INFERENCE

Although one may be tempted to "control for everything" in quantitative analyses involving multiple measured variables, this approach does not necessarily produce the correct causal coefficients. To summarize the conclusions from the previous section: when seeking to produce accurate causal estimates, one should control for confounders but not mediators or colliders.

Though simply stated, the application of these rules for confounders, mediators, and colliders can become more complex as the causal diagrams become more complex. To demonstrate these applications, we created a simulated dataset based on the causal diagram shown in Fig. 3. The simulations were conducted using RStudio [30]. First, $N = 10,000,000$ counts of variable X were generated randomly using $X = rnorm(N, 0, 1)$, where the function *rnorm* generates a standardized normal variable with mean 0 and standard deviation 1. Then, N counts for a new variable Z were computed using $Z = (0.20) * X + rnorm(N, 0, 0.9800)$, where $rnorm(N, 0, 0.9800)$ is the random error determined

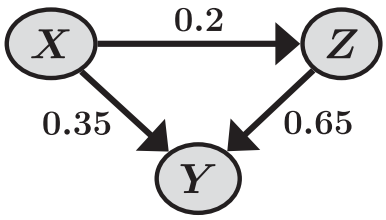


FIG. 3. Causal diagram depicting the relationships among variables X, Y, and Z, where X causes Z, and X and Z jointly cause Y with corresponding causal coefficients indicated on the arrows.

to make Z a standardized variable with standard deviation = 1. Finally, N counts for Y were computed using $Y=(0.35)*X+(0.65)*Z+rnorm(N,0,0.6035)$, where the random error term added was determined to make Y a standardized variable. This stepwise simulation procedure followed the causal pathways in Fig. 3: X determines Z, and then X and Z together determine Y. Simulating the data in this way created a dataset X, Y, Z, where the correct causal diagram and the causal coefficients associated with each directed arrow are known (Fig. 3).

In situations where the data are not simulated (and the path coefficients are not known), the path coefficients in Fig. 3 can be determined by basing the analyses on the causal diagram. In general, the analytic rule is that any node in the causal diagram with incoming arrows is a dependent variable in a regression analysis where the independent variables are the nodes that are the sources of those incoming arrows. The causal diagram in Fig. 3 illustrates two regressions needed to determine all path coefficients, $Y \sim X + Z$ and $Z \sim X$. To illustrate this analysis, we conducted these regressions and present the results in Tables I and II. The regression coefficients match the associated path coefficients in Fig. 3, verifying the accuracy of the simulation procedure.

The standard errors of these regression coefficients are not reported because they are both irrelevant to and a distraction from the larger purpose of the simulation: to illustrate how different regression models produce different regression coefficients. The magnitudes of the errors depend mainly on our chosen value of N. We deliberately chose a very large value for N in the data simulation to shrink the standard errors close to zero. On the computed regression coefficients that follow, all standard errors are less than 0.003.

TABLE I. Standardized coefficients for the model $Y \sim X + Z$ in the regression analysis of dependent variable Y with other variables as shown in Fig. 3.

Regression model: $Y = \beta_0 + \beta_{XY}X + \beta_{ZY}Z$	
Variable	β
Intercept	−0.0003
X	0.35
Z	0.65

TABLE II. Standardized coefficients for the model $Z \sim X$ in the regression analysis of dependent variable Z as shown in Fig. 3.

Regression model: $Z = \beta_0 + \beta_{XZ}X$	
Variable	β
Intercept	−0.0003
X	0.20

Next, we demonstrate correct (and incorrect) analyses for determining the magnitude of causal effects between variables depending on the causal structure. We show how the results determined from analyzing the simulated data relate to the path coefficients depicted in Fig. 3. This will also illustrate how the path coefficients in Fig. 3 can be used to determine the magnitudes of various causal and noncausal associations.

A. Rule: Do not control for mediators

Consider an analysis aiming to determine the causal impact of X on Y. Figure 3 shows that this total causal effect is the sum of the direct path $X \to Y$ and the indirect path $X \to Z \to Y$. Therefore, the total causal effect is 0.48: The sum of the direct effect 0.35 and the indirect effect $(0.2)(0.65) = 0.13$. This indicates that changing X by +1 SD would produce a change in Y of +0.48 SD.

Conceptually, the relationship between total, direct, and indirect causal effects can be understood as analogous to how total and partial derivatives are connected through the chain rule. Consider the function $y = f(x, z(x))$, where $x, y, z \in \mathbb{R}$. This function is analogous to the causal diagram in Fig. 3 since y depends on x and z, while z itself also depends on x. The total derivative $\frac{dy}{dx}$ can be written as

$$\frac{dy}{dx} = \frac{\partial y}{\partial x} + \frac{\partial y}{\partial z}\frac{\partial z}{\partial x}. \tag{2}$$

Analogously, $\frac{dy}{dx}$ represents the total effect of x on y, $\frac{\partial y}{\partial x}$ represents the direct effect of y on x when keeping z constant, and $\frac{\partial y}{\partial z}\frac{\partial z}{\partial x}$ represents the indirect effect of y on x that is mediated through changes in z.

The correct analysis for determining the total causal effect is $Y \sim X$. Because Z is a mediator in the indirect causal path, it should not be controlled for, as doing so will block this causal path. The linear regression analysis $Y \sim X$ on the simulated data results in the following regression coefficient:

$$Y = (0.48) * X, \tag{3}$$

where the intercept $|\beta_0| < 0.001$ is omitted. In this case, the coefficient for X is the total causal effect of X on Y. If one were to (incorrectly) control for Z by performing the analysis $Y \sim X + Z$, the regression coefficients would be

$$Y = (0.35) * X + (0.65) * Z. \tag{4}$$

Controlling for the mediator blocks the indirect causal effect, leaving only the direct causal effect, 0.35, as the coefficient for X.

B. Rule: Control for confounders

How would one determine the causal impact of Z on Y? Now, X is a confounder (common cause) of Z and Y, so it

should be controlled in the analysis. The total causal effect of Z on Y is only a direct effect, 0.65. However, the confounder creates a noncausal association of $(0.2)(0.35) = 0.07$ through the backdoor path $Z \leftarrow X \rightarrow Y$.

The correct analysis for determining the total causal effect of Z on Y is $Y \sim Z + X$. When applied to the simulated data, this analysis yields

$$Y = (0.65) * Z + (0.35) * X. \quad (5)$$

Here, controlling for X in the analysis means that the coefficient for Z will be the causal coefficient, representing the total causal effect of 0.65. However, if one does not control for X by performing the analysis $Y \sim Z$, this yields an incorrect causal coefficient:

$$Y = (0.72) * Z. \quad (6)$$

Note that for linear regression with one standardized independent variable Z and one standardized dependent variable Y, the regression coefficient equals $r_{YZ} = 0.72$. Because X was not controlled for, the backdoor path added the noncausal association 0.07 to the causal effect 0.65 to produce the regression coefficient 0.72. This example illustrates the problem with unmeasured confounders. Because the confounders must be controlled for to produce the correct causal coefficients, the existence of unmeasured confounders makes accurate causal analysis impossible. This is why observational study design should seek to measure all confounders so that they can be controlled for to block the noncausal associations from backdoor paths.

C. Rule: Do not control for colliders

For X and Z, Y is a collider. Because colliders should not be controlled in causal analysis, the analysis for determining the causal impact of X on Z should not control for Y. Controlling for variable Y opens a backdoor path contributing a negative, noncausal association between X and Z. This noncausal association means the coefficient on X will deviate from the correct total causal effect of X on Z. These deviated coefficients are commonly called "biased" coefficients, and a bias arising by controlling for collider Y is commonly called "collider stratification bias." Therefore, to find the causal coefficient of X on Z, one should not control for Y. To demonstrate this, first, we perform the regression analysis without controlling for Y, $Z \sim X$, which yields

$$Z = (0.20) * X. \quad (7)$$

This analysis gives the correct causal coefficient for X, 0.2. On the other hand, controlling for the collider Y through the analysis $Z \sim X + Y$ yields

$$Z = (-0.19) * X + (0.81) * Y. \quad (8)$$

The coefficient for X becomes negative, reflecting the fact that controlling for Y in this analysis has opened an additional negative, noncausal association between X and Z. This extreme example shows how controlling for a collider can even flip the sign of a regression coefficient, and naive interpretation of these analyses could produce different qualitative conclusions about the causal impact of X on Z.

D. Omitted variable bias

Omitted variable bias [31–34] is one term commonly used to describe a change in regression coefficients when the analysis does not control for other variables (i.e., omits the variables from the statistical model) [35]. The general conditions for omitted variable bias are that (i) the omitted variable has a nonzero regression coefficient when predicting the dependent variable and (ii) the omitted variable is correlated with other independent variables used in the regression analysis. This effect was demonstrated by showing how omitting or including mediators, confounders, and colliders from regression models can impact regression coefficients.

Although mathematically accurate, labeling this effect a "bias" may imply that no measured variables should be omitted in analyses where causal inference is the goal. This is incorrect. While mediators, confounders, and colliders all satisfy the two general conditions for omitted variable bias, only confounders should be controlled for in causal inference; mediators and colliders should be omitted in analysis. Controlling for mediators and colliders in analysis biases coefficients *away* from total causal effects.

V. THE IMPORTANCE OF RANDOMIZATION IN CAUSAL INFERENCE

One benefit of using DAGs to create causal diagrams is that the diagrams can concretely represent familiar issues in causal inference. For example, causal diagrams can illustrate why randomized controlled trials (RCTs) [14,36–38], where research participants are randomly assigned to a control or intervention group, are considered a "gold standard" for accurately determining the causal impact of one factor on another. Consider the case in Fig. 4(a), where X has a direct causal impact on Y, $X \rightarrow Y$, and multiple confounds C_1, C_2, and C_3—common causes of X and Y—exist.

If we observe these variables *in situ*, the regression analysis that will produce the correct causal coefficient of X on Y, d, is $Y \sim X + C_1 + C_2 + C_3$. The regression analysis $Y \sim X$ will produce a coefficient for X that is the sum of the direct causal effect of $X \rightarrow Y$, d, and the noncausal associations due to the three backdoor paths $X \leftarrow C_1 \rightarrow Y$, $X \leftarrow C_2 \rightarrow Y$, and $X \leftarrow C_3 \rightarrow Y$. Using the path coefficients in Fig. 4(a), this coefficient for X will be $d + a_1b_1 + a_2b_2 + a_3b_3$. This is another example of why controlling for confounders in causal analysis matters.

What could cause such confounds? One example is a selection effect. For instance, consider a scenario where students self-select into an optional physics study program

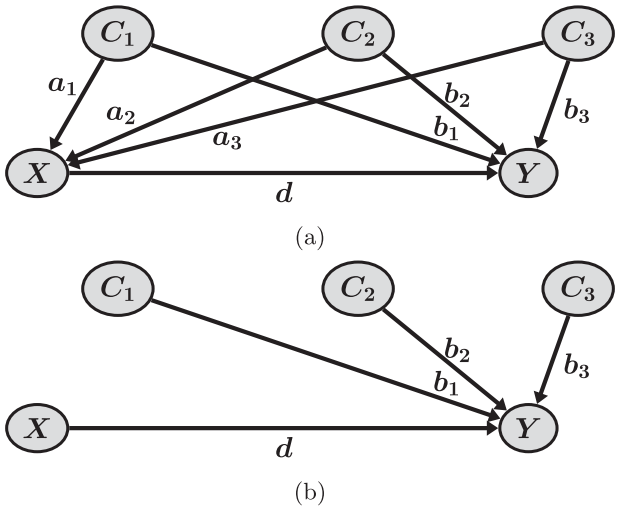


FIG. 4. (a) Causal diagram illustrating the relationship between X and Y, where X causes Y. C_1, C_2, and C_3 are common causes of both X and Y. The labels on the edges represent the causal coefficients. (b) A randomized experiment (or intervention) on X breaks the causal dependence of confounders on X.

based on their preexisting interest in the subject. In this case, the observed outcomes of the program on their later physics course performance might be influenced not only by the study program but also by their initial interest. In terms of the causal diagram, their prior interest (C_1) impacts both their participation in the program (X) and their ultimate performance (Y). To accurately determine the program's effectiveness, the analysis must control for prior interest (the confounder).

Although controlling for confounders is effective, it is important to acknowledge that one may not always be aware of or be able to measure all potential confounders. Randomization is another way to deal with confounders without explicit measurement and control. If experimenters can randomly assign participants to X, this will break the causal dependence on *all* confounders—known or unknown, measured or unmeasured—since the presence of X will no longer depend on C_1, C_2, or C_3 (in our example, randomly assigning students to participate in the optional physics study program means that participation in the optional program would no longer be associated with students' preexisting interest in the topic). In this new diagram [Fig. 4(b)], both the analysis $Y \sim X$ and $Y \sim X + C_1 + C_2 + C_3$ produce the same coefficient for X, d. Therefore, randomization theoretically removes the need to control for confounders to determine correct causal estimates.

VI. ANALYZING A SIMULATED DATASET II: HOW MULTIPLE CAUSAL MODELS CAN FIT THE SAME DATASET

In Sec. IV, because we were privy to the exact causal process through which the data were simulated, we were certain of how the variables were causally related. However, this is rarely, if ever, the case. Although there may be contexts where the causal model is clear, there are also instances where the exact causal model is uncertain or multiple causal models may be theoretically plausible. An important point is that quantitative statistics of model fit, though good at quantifying the predictive value of a statistical model, cannot be used to determine the correct causal model. That is because the model fit indicates a model's quantitative ability to explain the variance, but it does not specify whether that explained variance indicates a causal association, a noncausal association, or a combination of both.

As a clear example of why model fit does not equal causal validity, consider the simulated dataset represented by Fig. 3. As explained previously, in determining the causal impact of X on Y, Z should not be controlled because Z is a mediator of this causal impact. Therefore, the correct causal analysis is $Y \sim X$. However, if we choose whether or not to control for Z based on which regression model produces the highest R^2 fit, we will make the wrong decision. The correct causal analysis $Y \sim X$ has an $R^2 = 0.23$, and the incorrect causal analysis $Y \sim X + Z$ has an $R^2 = 0.63$. The reason is that including the mediator Z in the prediction of Y explains additional variance compared to when only X is used to predict Y, even though controlling for that mediator obscures the causal coefficient of X on Y. Although including mediators and colliders in regression models can provide a greater predictive fit by explaining a greater proportion of variance in the dependent variable, it will also bias regression coefficients away from estimates of the total causal effect.

Likewise, the existence of nonzero coefficients associated with an arrow in a causal diagram does not prove the validity of that causal model. To illustrate this, we simulated a dataset of standardized variables A, B, C that followed the correlations in Table III. Specifically, the data were simulated in RStudio [30] with the function *mvrnorm*, included in the MASS package [39]. We input the 3×3 covariance matrix shown in Table III that defines the three bivariate correlations between A, B, and C, set the mean value of each variable as zero, and set the sample size to $N = 10,000,000$. With these inputs, *mvrnorm* output a multivariate, normally distributed sample where values of A, B, and C were generated for $N = 10,000,000$ counts. Unlike the previous simulated dataset, this method of simulating data matches observational conditions, where correlations are explicitly measured, but the underlying causal structure is unknown.

To illustrate how multiple causal models can produce different, quantitatively reasonable path coefficients for the same dataset, we analyzed the simulated data using six different causal diagrams (Fig. 5). These six models are all the ones allowed when considering models where all pairwise direct links exist and omitting the cyclic models

TABLE III. Covariance matrix for variables A, B, and C.

	A	B	C
A	1.00	0.50	0.20
B	0.50	1.00	0.80
C	0.20	0.80	1.00

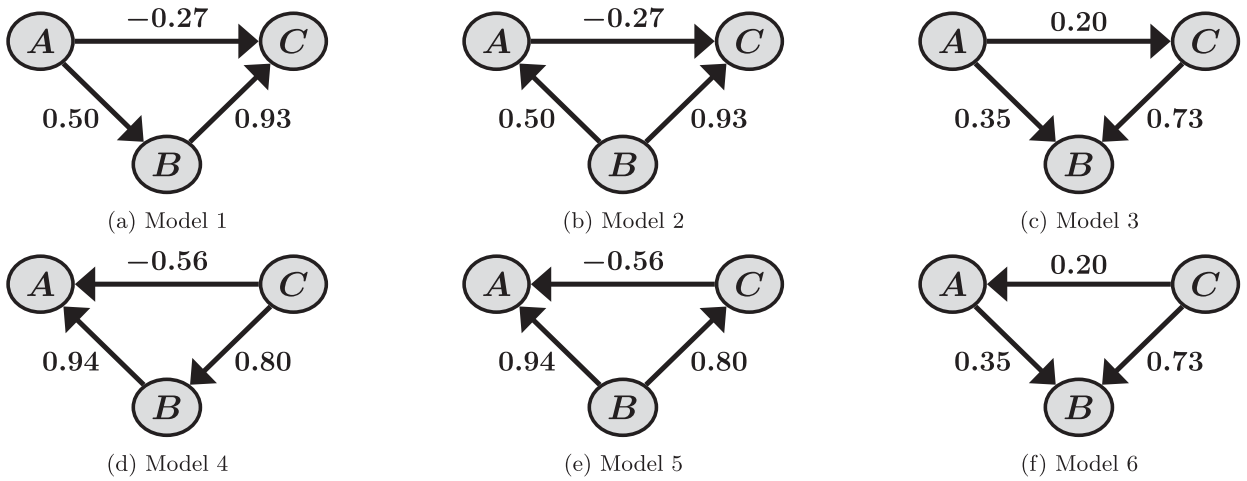
that are disallowed (such as the one with the links $A \to B$, $B \to C$, and $C \to A$). As a reminder, each diagram indicates the analyses required to find the path coefficients. For instance, in model 1, B has one arrow pointing into it coming from A, so the path coefficient for $A \to B$ can be determined through the analysis $B \sim A$, which will yield the regression equation $B = (0.5) * A$. C has direct arrows pointing into it from both A and B, so these path coefficients are determined by the analysis $C \sim A + B$, which will yield the regression line $C = (-0.27) * A + (0.93) * B$. Because A has no incoming arrows, it is not the dependent variable in any analysis.

Although all models find nonzero path coefficients, they make different predictions about how changing one variable will change another. For example, consider the question, "how will intervening on A affect C?" Model 1 indicates a direct effect $A \to C$ of -0.27, an indirect effect $A \to B \to C$ of $(0.50)(0.93) = 0.47$, and a total causal effect of $(-0.27) + (0.47) = 0.20$. Model 2 indicates a total effect of -0.27, which is solely attributed to a direct effect. Model 3 indicates a total effect of 0.20, which is solely attributed to a direct effect. Models 4–6 give an effect of zero since C is the cause and A is the effect, and changing A directly will not change C.

Although models 1 and 3 give the same total causal effect of A on C, this degeneracy is broken when considering how breaking the link $A \to B$ will change this total effect. This break could be accomplished by identifying a mediator M such that $A \to M \to B$ and controlling for M. In model 1, breaking the link $A \to B$ will block the indirect effect of A on C, changing the total effect to -0.27. In model 3, breaking the link $A \to B$ will not affect the causal relationship between A and C, so the total effect will remain 0.20. This is an example of how interventions on the causal system can break the degeneracy between different models.

Another issue is how the different models have different causal implications, even if the path coefficients are numerically identical. For example, consider models 1 and 2, which have the same numerical path coefficients and differ only in how the link between A and B is modeled: either as $A \to B$ or $A \leftarrow B$. Both models give a direct effect for $A \to C$ of -0.27. In model 1, the path through B is causal. B is a partial mediator through the path $A \to B \to C$, which represents an indirect, causal effect of $(0.50)(0.93) = 0.47$. In model 2, the path through B is noncausal. B is a confounder, so the path $A \leftarrow B \to C$ represents a noncausal association of magnitude $(0.50)(0.93) = 0.47$. Therefore, although the choice of $A \to B$ or $A \leftarrow B$ has no impact on the path coefficients computed, it does have an impact on the causal implications of the model.

This simulated example shows how critical choosing the causal model is. Researchers have the freedom to choose any causal model and apply it to the data, and the choice of model changes the conclusions that will be reached. The choice of model can even change the sign of a causal effect, as demonstrated when comparing the total effect of A on C in models 1 and 2. Just like an ansatz, the causal model is a guess—however theoretically or empirically justified—about the causal relationships among a system of variables. However, finding a model that fits the data is not proof that the ansatz was correct in this case. In fact, neither statistical goodness-of-fit nor nonzero path coefficients offer evidence supporting the causal validity of one model over another. The results are only as valid as the researchers' original causal assumptions, which are embodied in the proposed causal diagram. As the number of relevant and



(a) Model 1 (b) Model 2 (c) Model 3

(d) Model 4 (e) Model 5 (f) Model 6

FIG. 5. Six acyclic causal models that fit the same set of correlations between the variables in Table III.

collected variables grows, the number of possible causal models also grows, increasing the possibility that researchers have chosen the wrong model and reached the wrong causal conclusions.

Since the results of causal analysis under a proposed model cannot support the likelihood of that proposed model over others, how can observational research proceed in fields like PER? One way forward is to bridge observational and intervention studies, just as bridging theory and experiment has advanced knowledge in physics. Just as theoretical models in physics can make predictions to be tested experimentally, fitting observational data with candidate causal models can make causal predictions of how changing one factor will affect another, and these causal predictions can be tested through future intervention studies. Like experiments in physics, intervention studies that directly manipulate causes and measure changes to effects can provide empirical data about which associations are causal and which are not, to help support or falsify proposed causal models. In addition, perhaps more important than the theoretical validation of causal models, intervention studies also leverage observational research to design new approaches to improving physics education.

Next, we will apply these causal inference methods to interpret prior work in PER. In doing so, we will provide an example of how observational studies can establish quantitative causal models that can be investigated through future intervention studies.

VII. APPLYING CAUSAL INFERENCE PRINCIPLES TO PRIOR PER STUDIES

Although PER often uses quantitative analysis to draw conclusions about causal impacts, the causal diagrams, assumptions, and analytic techniques discussed in this paper are rarely explicitly employed to justify and structure the analysis. Here, we apply these causal inference methods to make sense of prior work in PER, demonstrating how these methods can provide a unified language for understanding various issues in quantitative PER.

A. Example: Omitted variable bias in PER

Walsh *et al.* [35] explored the effects of omitted variable bias through data from a quasiexperimental study. The study investigated students' attitudes toward experimental physics using Pre and Post E-CLASS survey measurements. Sampled physics students experienced either "transformed" or "highly traditional" physics laboratory instruction and were coded as either intending to major in physics or intending to major in another science or engineering field. Additionally, students' underrepresented minority (URM) status was collected. The focus of their analysis was the magnitude of the omitted variable bias from omitting instruction type from the analysis. They create three regression models using different combinations of pre E-CLASS score, major (physics = 1), instruction (transformed = 1), and URM status (URM = 1) to predict post E-CLASS score. Using the correlation and regression results given in the paper, we propose a causal diagram for these variables (Fig. 6).

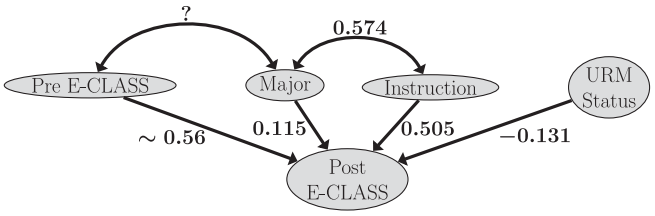


FIG. 6. A proposed causal diagram representing the causal structure of how Pre E-CLASS scores, major, instruction, and URM status predict the Post E-CLASS scores.

When the inclusion of one variable changes the regression coefficient for another variable, then those two variables must be correlated. Comparing regression models 1 and 2 in Ref. [35], adding instruction to the model changes the coefficient for major. Therefore, it is clear that major and instruction are correlated, and this correlation value (0.574) is reported in the paper. In our diagram, we represent this connection as a noncausal association: $Major \leftrightarrow Instruction$.[1] It is conceptually equivalent to the notation: $Major \leftarrow U \rightarrow Instruction$, where U is an unmeasured common cause of major and instruction. For instance, different types of instruction may be randomly assigned to different lab sections, and students may be blind to which sections are associated with each type of instruction. In this case, the association would be purely noncausal since there would be no causal mechanism for students' major to influence which lab instruction they receive. Alternatively, a plausible causal relationship is that students' major may influence the type of lab instruction they receive: $Major \rightarrow Instruction$. For instance, the transformed lab instruction may be officially associated with lab sections for majors such that students are officially advised to enroll in different lab sections by major. The transformed lab may also be messaged as "more advanced" or "for physics majors" in other ways that preferentially attract physics majors. The causal interpretation of this alternative model will be explored later. Because Pre E-CLASS and Major are both included in all three regression analyses, it is unknown whether or not they are correlated and if omitting one of them will change the regression coefficient of the other. We indicate this ambiguity with a "?" on $Pre\,E\text{-}CLASS \leftrightarrow Major$. Table I in Ref. [35] describes the results from three different regression models using Post E-CLASS as the dependent variable. Next, we show how our proposed diagram in Fig. 6 provides a single model that can determine the coefficients in these three

[1]Models in which single-headed and double-headed arrows occur are referred to as ADMGs (acylic directed mixed graphs) [40].

different regression analyses. In all three models, the standardized coefficient of PRE E-CLASS is 0.55–0.56, which is indicated by the path coefficient ∼0.56 for $Pre\,E\text{-}CLASS \rightarrow Post\,E\text{-}CLASS$ in the diagram.

Model 1 performs the analysis $Post\,E\text{-}CLASS \sim Major + Pre\,E\text{-}CLASS$. Using our diagram, we can see that controlling for Pre E-CLASS blocks the potential backdoor path $Major \leftrightarrow Pre\ E\text{-}CLASS \rightarrow Post\ E\text{-}CLASS$, but because this analysis does not control for instruction, the backdoor path from $Major \leftrightarrow Instruction \rightarrow Post\,E\text{-}CLASS$ is open. Using this diagram, we can determine that the regression coefficient for major will not be the correct causal coefficient. This regression coefficient, 0.405, will be the sum of the causal direct effect, 0.115, and the noncausal backdoor association $Major \leftrightarrow Instruction \rightarrow Post\ E\text{-}CLASS$, $(0.574)(0.505) = 0.290$. This is approximately equal to the regression coefficient for Major computed in Ref. [35], given for Model 1 in Table I.

Model 2 is the regression analysis $Post\ E\text{-}CLASS \sim Major + Instruction + Pre\ E\text{-}CLASS$. This analysis controls for Instruction, blocking the previously open noncausal backdoor path $Major \leftrightarrow Instruction \rightarrow Post\,E\text{-}CLASS$. Now, the regression coefficients in this analysis will match the direct, causal effects in Fig. 6: 0.115 for Major, 0.505 for Instruction, and 0.56 for Pre E-CLASS. These values match those given in Table I of Ref. [35] (though not exactly for Pre E-CLASS since we made the approximation that there is no correlation between Pre E-CLASS and Instruction that is unexplained by Major).

Model 3 in Table I of Ref. [35] describes the regression analysis $Post\ E\text{-}CLASS \sim Major + URM\ status + Pre\ E\text{-}CLASS$. With the backdoor path between Major and Instruction open again, the coefficient for Major will become similar to that in Model 1. Because the Pre E-CLASS and Major coefficients remain similar to the model 1 values, model 3 shows that URM status has a very small correlation with Pre E-CLASS or Major, which we approximate as zero by drawing no direct links from URM status to Pre E-CLASS or from URM status to Major.

This illustrates how causal diagrams can provide a single model that explains the results of different regression analyses while also encoding the causal assumptions of the researchers. What causal inferences can we make from the causal diagram in Fig. 6? Under the theoretical assumption that there is no causal association between major and instruction ($Major \leftrightarrow Instruction$), the type of instruction that students receive should not affect or be affected by their intended major. Therefore, one causal inference represented by this causal model is that experiencing the transformed lab instruction would increase students' average Post E-CLASS by an amount corresponding to a standardized coefficient of 0.505 over the traditional lab instruction. Here, Major is a confounder, so it must be controlled for to close a noncausal pathway between Instruction and Post E-CLASS. If Major was not controlled for, such as in the regression analysis $Post\,E\text{-}CLASS \sim Instruction + Pre\,E\text{-}CLASS$, the Instruction coefficient would be $0.571 = 0.505 + (0.574)(0.115)$, overestimating the causal coefficient by $(0.574)(0.115) = 0.066$ through the noncausal backdoor path $Instruction \leftrightarrow Major \rightarrow Post\,E\text{-}CLASS$.

What causal inference can be drawn about Major? Intended major is a proxy measure for factors that attract students to physics over engineering and other sciences, including academic preparation, interest, etc. These factors are hidden in the diagram as the common causes of Pre E-CLASS and Major, represented by $Pre\,E\text{-}CLASS \leftrightarrow Major$, since they may impact students' beliefs about experimental science as well as their choice of major. These unmeasured factors U could be more explicitly represented by $Pre\,E\text{-}CLASS \leftarrow U \rightarrow Major$. For students with the same Pre E-CLASS score and experiencing the same lab instruction, intending to major in physics will increase students' average Post E-CLASS score by an amount corresponding to a standardized coefficient of 0.115 over those intending to major in other science or engineering fields.

However, when considering potential interventions, it is important to keep in mind which variables are causes and which are proxies for causes. For example, a causally ridiculous conclusion to draw from the link $Major \rightarrow Post\,E\text{-}CLASS$ would be that universities should change all physics students' intended majors to physics in the university registration system to improve their experimental physics attitudes. Because students' major is a proxy for other unmeasured factors that impact their beliefs and learning, intervening on the proxy will not have a causal effect on Post E-CLASS. If the university were to mandate that all students become physics majors, this would not change those unmeasured factors, such as prior interests or experiences. Instead, it would break the association between major and unmeasured factors U, Major would no longer be a proxy measure for U, and, consequently, the coefficient for $Major \rightarrow Post\,E\text{-}CLASS$ would go to zero.

In contrast to this interpretation of Major, we now consider the alternative causal model where instruction partially mediates the causal effect of major: $Major \rightarrow Instruction \rightarrow Post\,E\text{-}CLASS$. If being a physics major increases the chances that one is enrolled in the transformed lab instruction course (such as if a physics majors-centric lab course uses the transformed instruction, while the nonmajors lab course uses the traditional instruction), then instruction should be considered part of the causal mechanism through which intended Major affects Post E-CLASS scores. In this case, the total causal effect of intending to major in physics would be 0.40, which includes the direct effect of Major of 0.115, associated with the unmeasured student factors related to major choice (like academic preparation, interest, etc.) and the indirect effect of physics majors being preferentially guided into the transformed lab

instruction and this lab instruction impacting students' experimental physics attitudes.

Although Walsh *et al.* [35] do not explicitly propose a causal interpretation of the 0.574 correlation between Major and Instruction, the causal diagram and associated rules for causal inference make it clear why this specification is important. While a general focus on omitted variable bias highlights how including or omitting variables from the analysis can affect regression coefficients, these causal techniques highlight additional issues around how those coefficients should be interpreted for accurate causal inference.

B. Example: Collider stratification bias through sampling in PER

The issue of noncausal coefficients arising from controlling for colliders—commonly called collider stratification bias—has been explicitly discussed in many contexts [41–55]. Weissman [56] explicitly discusses this issue in the context of education research, explaining how collider stratification bias can arise when controlling for educational outcomes in analysis. Here, we discuss another way that collider stratification bias can arise: through sampling.

A study in the 1960s investigating the mortality of babies born with a low birth weight counterintuitively found that babies whose mothers were smokers had *better* survival rates than babies of nonsmoking mothers [57]. Collider stratification bias was eventually used to explain why mothers should not be recommended to take up smoking while pregnant. In this example, low birth weight is a collider with multiple alternative causes. Smoking is one, but others also exist (such as birth defects). Since the study only investigated low birth weight babies, the sampling was conditioned on the collider. The result is that smoking and alternative low birth weight causes have a noncausal association in the collected dataset since low birthweight babies are likely to experience at least one of the causes, a smoking mother or an alternative low birthweight cause. Low birth weight babies who do not have a smoking mother are more likely to have alternative low birth weight causes, which may have even greater mortality rates than smoking.

Figure 7 shows a causal model of these variables. Restricting sampling to only low birth weight babies controls for a collider between smoking and alternative low birth weight causes. Therefore, the coefficient for smoking determined from the analysis $Mortality \sim Smoking$ will be the sum of the effects of the causal direct path $Smoking \rightarrow Mortality$ and the noncausal backdoor path $Smoking \rightarrow Low\,Birth\,Weight \leftarrow Alternative\,Low\,Birth\,Weight\,Causes \rightarrow Mortality$ opened by controlling for the collider *Low Birth Weight*. If the noncausal backdoor path has a negative contribution greater in magnitude than the direct path, then the overall regression coefficient will be negative. This explains how having a smoking mother could predict a lower mortality rate than having a nonsmoking mother because nonsmoking becomes associated with other alternative causes of low birth weight with a higher mortality rate.

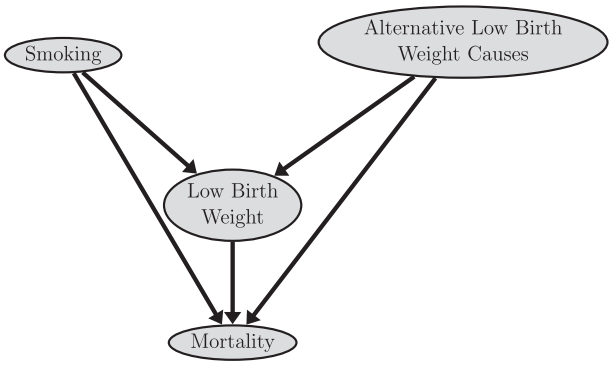


FIG. 7. Causal diagram illustrating the relationships between smoking, low birth weight, alternative low birth weight causes, and mortality as seen in the study [57].

One way to address this collider stratification bias would be to expand sampling to capture a representative distribution of birth weights. This would change the dataset to not condition on the collider, closing this noncausal backdoor path through *Low Birth Weight*. In this case, the analysis $Mortality \sim Smoking$ will produce a regression coefficient that represents the total causal effect (the direct effect of $Smoking \rightarrow Mortality$ plus the indirect effect of $Smoking \rightarrow Low\,Birth\,Weight \rightarrow Mortality$). Another way to address the collider stratification would be to measure and control for alternative causes of low birth weight, like birth defects. Although restricted sampling would still open the backdoor path through the collider (*Low Birth Weight*), controlling for birth defects and other alternative causes (which are common causes of low birth weight and mortality) will close (or reduce, in the case that not all alternative causes of low birth weight also directly affecting mortality can be controlled for) the noncausal confounding paths $Smoking \rightarrow Low\,Birth\,Weight \leftarrow Alternative\,Low\,Birth\,Weight\,Causes \rightarrow Mortality$. In this case, the analysis $Mortality \sim Smoking + Alternative\,Low\,Birth\,Weight\,Causes$, when conditioning on low birth weight through restricted sampling, will produce coefficients that estimate the direct causal effects of $Smoking \rightarrow Mortality$ and $Alternative\,Low\,Birth\,Weight\,Causes \rightarrow Mortality$. Although this removes contributions of noncausal backdoor associations from the regression coefficients, it also does not estimate the total causal impacts on mortality. This is because low birth weight is a partial mediator of the effects of smoking and alternative causes of low birth weight on mortality, and controlling for low birth weight closes these mediation pathways. Another weakness of this second approach is the challenge of precisely assessing the extent to which

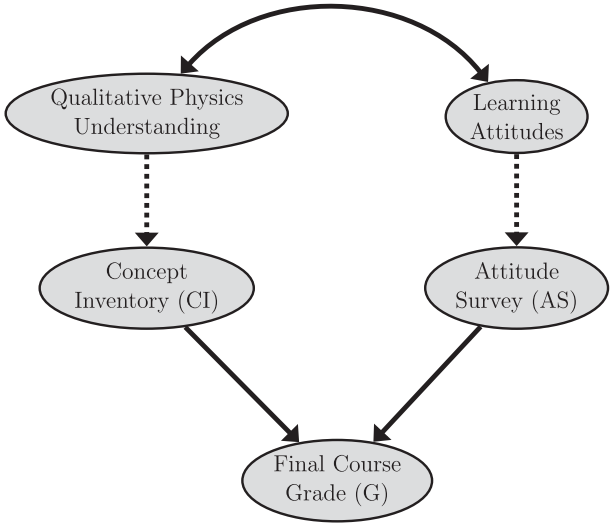


FIG. 8. A proposed causal diagram illustrating the causal structure of how concept inventory (CI) and attitude survey (AS) predict final course grade (G).

confounding is addressed by controlling for a set of covariates. The only statistical way to guarantee that all confounding is removed in this case will be if the measured variables completely account for the variance in low birth weight ($R^2 = 1$ when predicting low birth weight).

In PER, one data collection procedure where sampling creates collider bias is the completion of low-stakes, research-based surveys. Completion of these surveys, such as concept inventories or attitude surveys, during a physics course is associated with final course grade: specifically, students with higher grades are more likely to complete these surveys [58]. For this reason, complete-case analysis, which removes participants with missing data from the analysis, will partially control for final course grade. Consider the proposed causal model where a concept inventory (CI) and an attitude survey (AS) each serve as proxies for the qualitative physics understanding and learning attitudes that improve physics learning and performance as measured by final course grade (G) (Fig. 8). The partial control for the final grade partially opens the noncausal backdoor path through the collider $CI \rightarrow G \leftarrow AS$. Since we expect all causal coefficients to be positive, this backdoor path adds a noncausal negative contribution to the correlation between CI and AS. The expected impact is that measured correlations between CI and AS that do not address this collider stratification bias underestimate the strength of this correlation. Biases associated with missing data have led to increased attention on data imputation techniques, like multiple imputation, for estimating the contributions of missing data in PER [59]. Yet, just as with these causal inference methods, the accuracy of these methods depends critically on often unverifiable assumptions, in this case, about the nature of the missingness of the data and whether observed variables can adequately model the missing data.

C. Dealing with the bidirectional nature of motivation and beliefs with linear models

Although DAGs are easily applied to model unidirectional relationships between causes and effects, cases can exist where causality is bidirectional between two factors, such as in a feedback loop. One example from education where the causal directions are plausibly bidirectional is the relation between academic performance and motivation or beliefs.

To illustrate this, consider research on self-efficacy and academic performance. Self-efficacy and academic performance are correlated with each other, but which is the cause and which is the effect? Although many researchers focus on one causal pathway over another ($SE \rightarrow performance$ or $performance \rightarrow SE$), from its conception, self-efficacy has been theorized to affect and be affected by behavior and performance [60–69]. Self-efficacy influences behaviors, such as whether or not people engage and persist in challenging tasks, which creates opportunities to increase learning and performance. Reciprocally, experiencing mastery and success in performance is a strong predictor of future self-efficacy [70–74]. A sensible causal model between self-efficacy and academic performance would be cyclic [75,76], representing the bidirectional relationship between the two factors. However, these cyclic causal diagrams are disallowed by the formalism because the graphs must be acyclic. In our own work on the relations between self-efficacy and performance, we have grappled with how to causally understand the quantitative relations between self-efficacy and physics performance in the absence of causal diagram methods [77].

One way to conceptualize such reciprocal relationships is through longitudinal measurement and cross-lagged panel analysis. As an example, Talsma *et al.* (2017) [78] conducted a meta-analysis of longitudinal self-efficacy studies with a cross-lagged model where self-efficacy and performance, correlated with each other at time 1, are both allowed to affect self-efficacy and performance at time 2 (Fig. 9). The longitudinal repeated measurements of performance and self-efficacy open up alternatives to cyclic diagrams. The cross-lagged panel model also disentangles the effects of prior self-efficacy and prior performance, which are themselves correlated. This causal diagram also clarifies the risks of simply associating self-efficacy at time 1 with academic performance at time 2. Talsma *et al.* [78] report that this correlation is $r_{\text{SE1-P2}} = 0.248$. However, the analysis associated with this diagram shows that the causal effect $SE1 \rightarrow P2$ is only 0.071 and that the rest of this correlation reflects a noncausal backdoor association $SE1 \leftrightarrow P1 \rightarrow P2$ of $(0.316)(0.560) = 0.177$. That is, the majority of this correlation reflects the fact that self-efficacy and performance at time 1 are correlated with each other and that the direct effect of $P1 \rightarrow P2$ is relatively large. Neglecting this backdoor association in analysis overestimates the causal impact of self-efficacy on

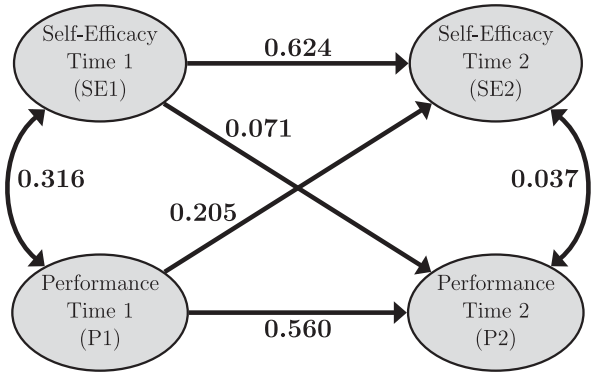


FIG. 9. This is a reproduction of Fig. 2 of Ref. [78]. Causal diagram of the cross-lagged path model between self-efficacy and academic performance at time 1 and time 2.

performance. The cross-lagged diagram also clarifies that the correlation between self-efficacy and performance is mostly explained by the mechanism of their coevolution over time. Although $r_{\text{SE2-P2}} = 0.312$, the direct noncausal association $SE2 \leftrightarrow P2$ only has a coefficient of 0.037. This indicates that an association of 0.275 is explained through the backdoor paths including SE1 and P1. That is, most of the correlation between self-efficacy and performance is due to the fact that they both codevelop out of prior self-efficacy and performance.

This example of self-efficacy shows how these causal methods can potentially clarify the muddy, reciprocal relations commonly theorized when considering relationships between academic performance and behavior with motivation, self-concept, and attitudes. Although this cross-lagged panel analysis illustrates the conceptual issues regarding reciprocal influences between variables, this is, in some sense, a toy model. More modern methods have since been suggested that capture the same conceptual issues while relaxing some of the underlying assumptions required to produce accurate causal estimates [79–81].

D. Proposing an explicit role for causal modeling of observational data in PER: Motivating future intervention studies

The validity of the causal inferences drawn from quantitative analysis depends critically on the validity of the underlying causal model guiding analysis and interpretation. This causal model, which can be represented explicitly with a DAG, is based on researchers' (explicit and/or implicit) theoretical understanding of the causal system. The critical issue is that a researcher's underlying causal model cannot be "verified" by the quantitative results of fitting observational data to that model. Intuitively, it may be appealing to interpret nonzero regression coefficients or extremizing quantitative metrics of model fit as evidence that a proposed causal model is correct. However, this is not a valid inference. Even a *noncausally* correlated set of variables can produce nonzero regression coefficients and provide a good model fit for predicting outcomes.

In this journal, Weissman has called for explicit consideration of multiple plausible causal models for observational studies drawing causal inferences [1], which is especially relevant in cases with a large number of variables and plausible causal relationships between them. We agree with Weissman that this is a sensible call for considering alternative explanations in quantitative research. As demonstrated previously, changing one's assumptions about whether a variable is a mediator, confounder, or collider can change quantitative causal estimates, as well as how the quantitative causal analysis should be conducted and interpreted. Therefore, different causal models can produce different theoretical interpretations of the quantitative data.

Although there may be theoretical reasons to favor one model over another, a strong empirical test of a proposed model is to design experimental interventions based on that proposed model. Therefore, in addition to the consideration of alternative causal models, we propose an explicit goal for observational studies drawing causal inferences: proposing future intervention studies and predicting their outcomes. Just as physics theories motivate future experiments, any proposed causal model of observational data embodies a set of causal explanations for observed correlations, and the causal estimates produced by applying those theoretical models to observational data are predictions of the effects of future interventions. Framing causal inferences from observational studies as theory clarifies that these inferences are one possible set of proposed theoretical explanations of the data. A secondary benefit of this explicit framing of "causal inference from observational data as theory" is that it highlights and promotes the scientific value of experimental intervention. When interventions act on causes, they can break associations with confounders, eliminating noncausal backdoor paths and providing strong tests of proposed causal models.

To propose a concrete example of using observational data to predict the results of future interventions, we consider a recent example from Li and Singh [82], who used observational data to investigate the relations between gender and four motivational constructs: self-efficacy, interest, perceived recognition, and identity. This paper provides a good case study of these issues for the following reasons: It analyzes correlated motivational factors in a nonintervention setting; the motivational variables are highly correlated ($r > 0.6$ for all six pairwise correlations between the four motivational factors); and it deals with motivational variables which can plausibly be modeled as reciprocally co-evolving (i.e., there can be theoretical debate about which factors are directly linked and about the directions of these links). This paper also does the rare work of explicitly comparing alternative models, considering four models where self-efficacy, interest, and perceived recognition mediate the relationship between gender and identity. Although they do not explicitly

state that their goal is causal modeling, Li and Singh end up drawing causal conclusions from the model about how intervening on one factor should change another. Therefore, we consider their four models as causal models.

Model 1 [Fig. 10(a)] considers no causal association between self-efficacy, interest, and perceived recognition. Models 2–4 [Figs. 10(b) to 10(d)] make one of these mediators a cause of the other two mediators. For instance, model 4 describes the total effect of perceived recognition on identity as the sum of a direct effect $Perceived\,Recognition \rightarrow Identity$ and indirect effects mediated through interest and self-efficacy. Because of the highly connected nature of these motivational constructs, all of these models could be viewed as theoretically reasonable to some degree. For example, because it is reasonable for people to be more interested in topics they believe they can successfully learn, $SE \rightarrow Interest$ is plausible. However, because interest also likely increases engagement and learning in an area, $Interest \rightarrow SE$ is also plausible. Similarly, although identity is an effect caused by the three other factors, identity may also be a cause that can impact one's self-efficacy, interest, or perceived recognition. Although this paper uses structural equation modeling rather than path analysis with linear regression to find the path coefficients, the conceptual issues underlying the causal modeling remain the same.

The quantitative analysis of these four models does not prove or disprove any one model over the others. Li and Singh [82] recognize this and provide alternative reasons to favor model 4 over the others. First, the authors argue that a model making perceived recognition of the parent cause of self-efficacy, interest, and identity provides the best motivation for instructor change. They explain that self-efficacy, interest, and identity may be seen as student properties outside of an instructor's locus of control, whereas perceived recognition is something that a teacher is more likely to believe they can intervene upon. Whether or not this is true, this argument does not support the causal validity of model 4 over the others. Having desirable implications for action does not make a model more causally accurate than other models. The second reason provided to favor model 4 over the others is that interviewed students self-report perceived recognition as the causal antecedent of their later self-efficacy, interest, and identity. Although this is one piece of evidence supporting model 4, it is also true that a person's retrospective self-reported perceptions of a phenomenon, even one regarding their own motivation and beliefs, may not accurately reflect the causal mechanisms

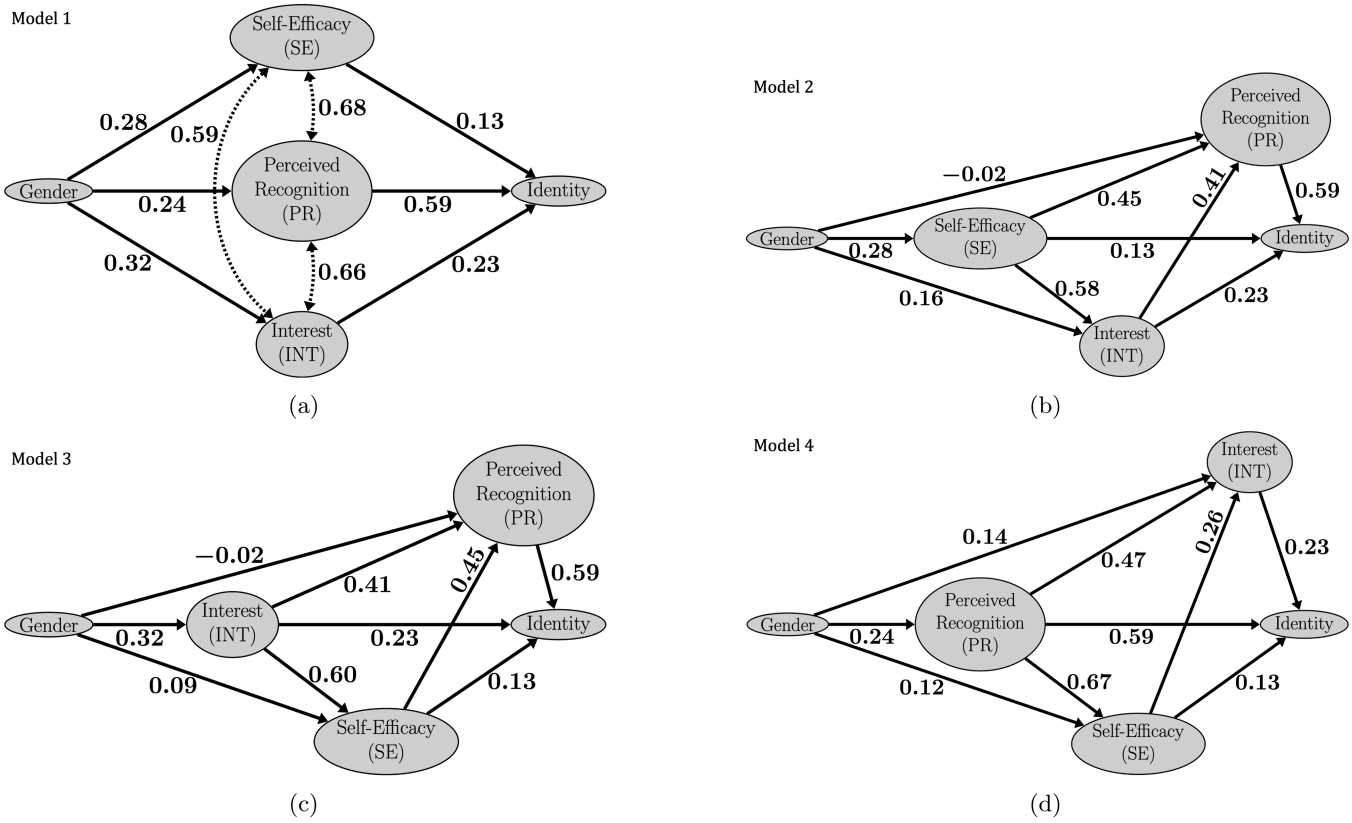


FIG. 10. Models, as shown in Ref. [82], represent theorized relationships between gender and physics identity mediated through three motivational variables: self-efficacy (SE), interest (INT), and perceived recognition (PR). (a) In model 1, the three mediating variables are noncausally associated with one another. (b–d) In models 2–4, there are causal associations between the three motivational variables: one of the three motivational variables is a common cause of the others, and another variable is a common effect of the others.

of that phenomenon. Therefore, there is still value in differentiating the proposed models through experimental intervention.

We propose that the best method to differentiate between these models is to compare how well they predict future interventions. For instance, consider an intervention that aims to increase students' perceived recognition. Models 1–3 predict that a +1 SD increase in perceived recognition should produce a +0.59 SD increase in identity and that SE and interest should not change because they are not causally associated with perceived recognition (model 1) or they are causes of perceived recognition that should not be affected when perceived recognition is directly intervened upon (models 2 and 3). However, model 4 predicts that a +1 SD change to perceived recognition should cause a total effect of $(0.59) + (0.47)(0.23) + (0.67)[0.13 + (0.26 \times 0.23)] = +0.83$ SD on identity, a +0.67 SD change on SE, and a $(0.47) + (0.67)(0.26) = +0.64$ SD change on interest. Therefore, collecting data on the effects of intervening on perceived recognition and comparing the results against the predictions made by each theoretical model can potentially support or downweight model 4. Incorrect causal models conflate causal effects with noncausal associations and can misestimate how interventions on one factor will cascade through the causal system.

In addition, the intervention should have upstream consequences as well. Intervening on perceived recognition should also weaken or break associations with its causes since direct intervention will change perceived recognition so that it is less tied to its original causes. Importantly, these models suggest that perceived recognition is a mediator of this effect of gender on identity, so weakening the direct and indirect paths between gender and perceived recognition should weaken the total causal effect of gender on identity (the ultimate educational goal of this modeling exercise). However, the intervention could have no impact on the total effect of gender on identity, which could indicate that the proposed causal model is incorrect and that perceived recognition, interest, and/or self-efficacy do not mediate the causal impact of gender on identity. In this way, the DAGs provide an explicit, quantitative model for making quantitative predictions about the cascading effects of hypothesized interventions.

This use of observational results to motivate and predict the results of intervention studies is aligned with the ultimate goal of improving educational experiences and outcomes for students. Debates about which theoretical model correctly describes the underlying causal relationships are only useful as far as they inspire and inform the design of future interventions. Proposing (and conducting) future intervention studies motivated by these theories moves us closer to the goal of designing, testing, and disseminating instructional improvements.

VIII. SUMMARY

"Correlation does not imply causation" is a commonly stated aphorism that reminds researchers to err on the side of caution. However, it is equally true that "correlations sometimes indicate causation" and that "correlations can contain information about causation." Causal diagrams and associated rules for statistical causal inference provide a framework for extracting causal information from correlational data when appropriate (and for cautioning researchers from doing so when it is not appropriate). While we expect that many physics education researchers engaged in making statistical causal inferences will be familiar with these methods, we hope that this paper helps knowledge of these techniques become more widespread in PER. This paper describes some of the well-known fundamental principles of statistical causal inference, illustrates some connections to existing PER studies, and proposes a new explicit epistemological role for observational studies as theoretical proposals for future intervention studies. We hope that this paper provides a starting point for researchers to learn more about the causal inference methods and analysis techniques well established outside of PER.

We close by summarizing four main takeaway points from our discussion of causal inference methods:

(1) Researchers should be explicit and consistent about whether their goal is causal modeling or (noncausal) predictive modeling. Causal and predictive models have different goals and different criteria for evaluation. Unclear and inconsistent language around whether or not a model is meant to be causal muddles decisions about how these models should be constructed, evaluated, and interpreted.

(2) A primary rule of causal inference is that analysis should control for confounders and not control for mediators and colliders. Making a researcher's causal assumptions about a system explicit through a DAG provides a diagrammatic method for differentiating confounders, mediators, and colliders. These causal inference techniques are especially important in observational studies, where observed correlations can represent both causal and noncausal associations. The benefit of intervention studies is that direct intervention on causal factors can break or weaken the noncausal backdoor associations opened by confounders.

(3) The biggest weakness of these causal inference techniques is that the validity of the causal inferences depends entirely on the accuracy of the proposed causal diagram. Quantitative metrics, such as path coefficient values or goodness-of-fit statistics, cannot support the causal validity of one proposed model over another. Therefore, even seemingly reasonable regression coefficients produced by quantitative analysis can be causally

incorrect. This highlights the importance of avoiding (explicit and implicit) claims that a satisfactory quantitative fit of a causal model onto observational data "proves," "shows," or "demonstrates" evidence for any causal claim. The most plausible alternative causal models should be explicitly considered. This is especially true in fields like education, where cause-and-effect stories are often complex, and there can be many plausible causal models that explain a highly correlated dataset.

(4) We propose an explicit role for studies applying path analysis, structural equation modeling, or other analyses commonly used to draw causal inferences from observational data: motivating future intervention studies. This role embraces the strengths of observational studies while making explicit the theoretical nature of the causal inferences drawn. It also promotes greater coordination between observation and intervention studies to forward the science of effective instructional interventions.

[1] M. B. Weissman, Policy recommendations from causal inference in physics education research, Phys. Rev. Phys. Educ. Res. **17**, 020118 (2021).

[2] M. B. Weissman, Invalid methods and false answers: Physics education research and the use of GREs, Econ. J. Watch **19**, 4 (2022).

[3] H. R. Varian, Causal inference in economics and marketing, Proc. Natl. Acad. Sci. U.S.A. **113**, 7310 (2016).

[4] J. M. Rohrer, Thinking clearly about correlations and causation: Graphical causal models for observational data, Adv. Methods Pract. Psychol. Sci. **1**, 27 (2018).

[5] E. M. Foster, Causal inference and developmental psychology, Dev. Psychol. **46**, 1454 (2010).

[6] T. A. Glass, S. N. Goodman, M. A. Hernán, and J. M. Samet, Causal inference in public health, Annu. Rev. Public health **34**, 61 (2013).

[7] M. Gangl, Causal inference in sociological research, Annu. Rev. Sociol. **36**, 21 (2010).

[8] L. Keele, The statistics of causal inference: A view from political methodology, Political Anal. **23**, 313 (2015).

[9] R. J. Murnane and J. B. Willett, *Methods Matter: Improving Causal Inference in Educational and Social Science Research* (Oxford University Press, New York, 2010).

[10] D. Freedman, R. Pisani, and R. Purves, *Statistics* (W.W. Norton & Company, New York, 2007).

[11] G. W. Imbens, Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics, J. Econ. Lit. **58**, 1129 (2020).

[12] T. C. Williams, C. C. Bach, N. B. Matthiesen, T. B. Henriksen, and L. Gagliardi, Directed acyclic graphs: A tool for causal studies in paediatrics, Pediatr. Res. **84**, 487 (2018).

[13] G. W. Imbens and D. B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press, Cambridge, England, 2015).

[14] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect* (Basic Books, New York, NY, 2018).

[15] S. Greenland, J. Pearl, and J. M. Robins, Causal diagrams for epidemiologic research, Epidemiology, **10**, 37 (1999).

[16] M. Hernán and J. Robins, *Causal Inference: What If* (Chapman & Hall/CRC, Boca Raton, FL, 2020).

[17] M. Glymour, J. Pearl, and N. P. Jewell, *Causal Inference in Statistics: A Primer* (John Wiley & Sons, New York, 2016).

[18] C. Glymour, K. Zhang, and P. Spirtes, Review of causal discovery methods based on graphical models, Front. Genet. **10**, 524 (2019).

[19] S. L. Morgan, *Handbook of Causal Analysis for Social Research* (Springer, New York, 2013).

[20] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms* (The MIT Press, Cambridge, MA, 2017).

[21] S. L. Morgan and C. Winship, *Counterfactuals and Causal Inference* (Cambridge University Press, Cambridge, England, 2015).

[22] J. Pearl, Linear models: A useful "microscope" for causal analysis, J. Causal Infer. **1**, 155 (2013).

[23] G. Shmueli, To explain or to predict?, Stat. Sci. **25**, 289 (2010).

[24] M. Kuhn, K. Johnson *et al.*, *Applied Predictive Modeling* (Springer, New York, 2013), Vol. 26.

[25] E. W. Burkholder, G. Murillo-Gonzalez, and C. Wieman, Importance of math prerequisites for performance in introductory physics, Phys. Rev. Phys. Educ. Res. **17**, 010108 (2021).

[26] S. McCammon, J. Golden, and K. L. Wuensch, Predicting course performance in freshman and sophomore physics courses: Women are more predictable than men, J. Res. Sci. Teach. **25**, 501 (1988).

[27] M. Verostek, C. W. Miller, and B. Zwickl, Analyzing admissions metrics as predictors of graduate GPS and whether graduate GPA mediates Ph. D. completion, Phys. Rev. Phys. Educ. Res. **17**, 020115 (2021).

[28] J. Pearl, Causal diagrams for empirical research, Biometrika **82**, 669 (1995).

[29] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, Prediction, and Search* (MIT Press, Cambridge, MA, 2000).

[30] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria 2022).

[31] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge University Press, Cambridge, England, 2006).

[32] K. A. Clarke, The phantom menace: Omitted variable bias in econometric research, Confl. Manag. Peace Sci. **22**, 341 (2005).
[33] K. A. Clarke, Return of the phantom menace: Omitted variable bias in political research, Confl. Manag. Peace Sci. **26**, 46 (2009).
[34] S. K. Riegg, Causal inference and omitted variable bias in financial aid research: Assessing solutions, Rev. High. Educ. **31**, 329 (2008).
[35] C. Walsh, M. M. Stein, R. Tapping, E. M. Smith, and N. G. Holmes, Exploring the effects of omitted variable bias in physics education research, Phys. Rev. Phys. Educ. Res. **17**, 010119 (2021).
[36] D. Hutchison and B. Styles, *A Guide to Running Randomised Controlled Trials for Educational Researchers* (NFER, Slough, 2010).
[37] C. Torgerson, A. Wiggins, D. Torgerson, H. Ainsworth, and C. Hewitt, Every child counts: Testing policy effectiveness using a randomised controlled trial, designed, conducted and reported to consort standards, Res. Math. Educ. **15**, 141 (2013).
[38] L. V. Hedges and J. Schauer, Randomised trials in education in the USA, Educ. Res. **60**, 265 (2018).
[39] B. Ripley, B. Venables, D. M. Bates, K. Hornik, A. Gebhardt, D. Firth, and M. B. Ripley, Package 'mass', Cran r **538**, 113 (2013).
[40] T. Richardson, Markov properties for acyclic directed mixed graphs, Scand. J. Stat. Theory Appl. **30**, 145 (2003).
[41] S. Greenland, Quantifying biases in causal models: Classical confounding vs collider-stratification bias, Epidemiology **14**, 300 (2003).
[42] B. W. Whitcomb, E. F. Schisterman, N. J. Perkins, and R. W. Platt, Quantification of collider-stratification bias and the birthweight paradox, Paediatr. Perinat. Epidemiol. **23**, 394 (2009).
[43] F. Elwert and C. Winship, Endogenous selection bias: The problem of conditioning on a collider variable, Annu. Rev. Sociol. **40**, 31 (2014).
[44] H. R. Banack and J. S. Kaufman, From bad to worse: Collider stratification amplifies confounding bias in the "obesity paradox", Eur. J. Epidemiol. **30**, 1111 (2015).
[45] M. Sperrin, J. Candlish, E. Badrick, A. Renehan, and I. Buchan, Collider bias is only a partial explanation for the obesity paradox, Epidemiology **27**, 525 (2016).
[46] C. Coscia, D. Gill, R. Benítez, T. Pérez, N. Malats, and S. Burgess, Avoiding collider bias in Mendelian randomization when performing stratified analyses, Eur. J. Epidemiol. **37**, 671 (2022).
[47] D. J. Del Junco, E. M. Bulger, E. E. Fox, J. B. Holcomb, K. J. Brasel, D. B. Hoyt, J. J. Grady, S. Duran, P. Klotz, M. A. Dubick *et al.*, Collider bias in trauma comparative effectiveness research: The stratification blues for systematic reviews, Injury **46**, 775 (2015).
[48] F. R. Leite, G. G. Nascimento, K. G. Peres, F. F. Demarco, B. L. Horta, and M. A. Peres, Collider bias in the association of periodontitis and carotid intima-media thickness, Community Dent. Oral Epidemiol. **48**, 264 (2020).
[49] M. Sanni Ali, R. H. Groenwold, W. R. Pestman, S. V. Belitser, A. W. Hoes, A. De Boer, and O. H. Klungel, Time-dependent propensity score and collider-stratification bias: An example of beta 2-agonist use and the risk of coronary heart disease, Eur. J. Epidemiol. **28**, 291 (2013).
[50] T. Tönnies, S. Kahl, and O. Kuss, Collider bias in observational studies, Dtsch. Aerztebl. Int. **119**, 107 (2022).
[51] M. J. Holmberg and L. W. Andersen, Collider bias, JAMA, J. Am. Med. Assoc. **327**, 1282 (2022).
[52] G. J. Griffith, T. T. Morris, M. J. Tudball, A. Herbert, G. Mancano, L. Pike, G. C. Sharp, J. Sterne, T. M. Palmer, G. Davey Smith *et al.*, Collider bias undermines our understanding of COVID-19 disease risk and severity, Nat. Commun. **11**, 5749 (2020).
[53] M. A. Hernán, S. Hernández-Díaz, and J. M. Robins, A structural approach to selection bias, Epidemiology **15**, 615 (2004).
[54] T. J. VanderWeele and J. M. Robins, Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect, Am. J. Epidemiol. **166**, 1096 (2007).
[55] S. R. Cole, R. W. Platt, E. F. Schisterman, H. Chu, D. Westreich, D. Richardson, and C. Poole, Illustrating bias due to conditioning on a collider, Int. J. Epidemiol. **39**, 417 (2010).
[56] M. B. Weissman, Do GRE scores help predict getting a physics Ph.D.? A comment on a paper by Miller et al., Sci. Adv. **6**, eaax3787 (2020).
[57] J. Yerushalmy, The relationship of parents' cigarette smoking to outcome of pregnancy—implications as to the problem of inferring causation from observed associations, Am. J. Epidemiol. **93**, 443 (1971).
[58] J. M. Nissen, M. Jariwala, E. W. Close, and B. V. Dusen, Participation and performance on paper-and computer-based low-stakes assessments, Int. J. STEM Educ. **5**, 21 (2018).
[59] J. Nissen, R. Donatello, and B. Van Dusen, Missing data and bias in physics education research: A case for using multiple imputation, Phys. Rev. Phys. Educ. Res. **15**, 020106 (2019).
[60] A. Bandura, *Social Foundations of Thought and Action* (Englewood Cliffs, NJ, 1986), p. 23.
[61] E. A. Locke, Self-efficacy: The exercise of control, Pers. Psychol. **50**, 801 (1997).
[62] A. D. Liem, S. Lau, and Y. Nie, The role of self-efficacy, task value, and achievement goals in predicting learning strategies, task disengagement, peer relationship, and achievement outcome, Contemp. Educ. Psychol. **33**, 486 (2008).
[63] C. M. Vogt, Faculty as a critical juncture in student retention and performance in engineering programs, J. Eng. Educ. **97**, 27 (2008).
[64] B. D. Jones, M. C. Paretti, S. F. Hein, and T. W. Knott, An analysis of motivation constructs with first-year engineering students: Relationships among expectancies, values, achievement, and career plans, J. Eng. Educ. **99**, 319 (2010).
[65] T. Honicke and J. Broadbent, The influence of academic self-efficacy on academic performance: A systematic review, Educ. Res. Rev. **17**, 63 (2016).
[66] J. B. Vancouver, C. M. Thompson, and A. A. Williams, The changing signs in the relationships among self-efficacy, personal goals, and performance, J. Appl. Psychol. **86**, 605 (2001).

[67] J. Hattie and E. M. Anderman, *International Guide to Student Achievement* (Routledge, London, 2013).
[68] F. Pajares, Current directions in self-efficacy research, Adv. Motiv. Achiev. **10**, 1 (1997).
[69] B. J. Zimmerman, Self-efficacy: An essential motive to learn, Contemp. Educ. Psychol. **25**, 82 (2000).
[70] S. L. Britner and F. Pajares, Sources of science self-efficacy beliefs of middle school students, J. Res. Sci. Teach. **43**, 485 (2006).
[71] E. L. Usher and F. Pajares, Sources of self-efficacy in school: Critical review of the literature and future directions, Rev. Educ. Res. **78**, 751 (2008).
[72] R. W. Lent, F. G. Lopez, and K. J. Bieschke, Mathematics self-efficacy: Sources and relation to science-based career choice, J. Counsel. Psychol. **38**, 424 (1991).
[73] R. M. Klassen, A cross-cultural investigation of the efficacy beliefs of South Asian Immigrant and Anglo Canadian nonimmigrant early adolescents, J. Educ. Psychol. **96**, 731 (2004).
[74] T. Matsui, K. Matsui, and R. Ohnishi, Mechanisms underlying math self-efficacy learning of college students, J. Vocat. Behav. **37**, 225 (1990).
[75] K. D. Multon, S. D. Brown, and R. W. Lent, Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation, J. Counsel. Psychol. **38**, 30 (1991).
[76] H. P. Phan, Informational sources, self-efficacy and achievement: A temporally displaced approach, Educ. Psychol. **32**, 699 (2012).
[77] K. Boden, E. Kuo, T. Nokes-Malach, T. Wallace, and M. Menekse, What is the role of motivation in procedural and conceptual physics learning? An examination of self-efficacy and achievement goals, presented at PER Conf. 2017, Cincinnati, OH, 10.1119/perc.2017.pr.010.
[78] K. Talsma, B. Schüz, R. Schwarzer, and K. Norris, I believe, therefore i achieve (and vice versa): A meta-analytic cross-lagged panel analysis of self-efficacy and academic performance, Learn. Individ. Diff. **61**, 136 (2018).
[79] M. Mund and S. Nestler, Beyond the cross-lagged panel model: Next-generation statistical tools for analyzing interdependencies across the life course, Adv. Life Course Res. **41**, 100249 (2019).
[80] E. L. Hamaker, R. M. Kuiper, and R. P. Grasman, A critique of the cross-lagged panel model, Psychol. Methods **20**, 102 (2015).
[81] S. Usami, N. Todo, and K. Murayama, Modeling reciprocal effects in medical research: Critical discussion on the current practices and potential alternative models, PLoS One **14**, e0209133 (2019).
[82] Y. Li and C. Singh, How to select suitable models from many statistically equivalent models: An example from physics identity, arXiv:2303.13786.