Gain Scores, ANCOVA, and Propensity Matching Procedures for Evaluating Treatments in Education

Daniel B. Wright
University of Nevada, Las Vegas

Abstract

Researchers have several options available to analyze data from interventions when participants have not been randomly allocated into conditions. Among these are the gain score, ANCOVA, and propensity matching procedures. Each of these attempts to account for pre-treatment differences among the conditions, but they do so differently. These procedures are reviewed and methods for estimating them in R are shown. The choice of which of these procedures to use can be difficult. Different situations are shown where they perform differently. The primary conclusion of this paper is that models should be hypothesized for how the data may arise, data simulated for these models, and the properties of statistical procedures evaluated. A goal of this paper is to show these procedures without extensive mathematics in order to allow a broad readership to use these methods in their situations.

Keywords: Propensity Matching, ANCOVA, Gain scores, Lord's paradox, Causality, Graphs, Simulation

When education researchers propose a new treatment, it is important to evaluate whether it is beneficial (Mosteller & Boruch, 2002). If students have been randomly allocated into conditions, the researchers could compare the treatment group with a control group on the outcome variable using something like a t-test. While this is likely not the most powerful statistical test depending on several factors, it provides an unbiased test of the treatment effect.

In much education research it is not practical (and sometimes impossible) to allocate students into conditions: for example, to change students' social class; to change their early childhood experiences; to change their pre-study knowledge; to make them have specific psychological conditions, *etc.* This is true in many disciplines: e.g., an astronomer cannot

Thanks to Sarah Wells and the reviewers for helpful comments on an earlier draft.
Department of Educational Psychology & Higher Education, College of Education, University of Nevada, Las Vegas. Daniel Wright is the Dunn Family Endowed Chair and Professor of Educational Assessment.
This research received no funding beyond the endowment.
Email: daniel.wright@unlv.edu or dbrookswr@gmail.com

assign a star to go supernova; a historian cannot randomly decide whether Trotsky or Stalin succeeds Lenin; see also the evaluations at `www.povertyactionlab.org`.

Because of this, researchers rely on statistical procedures, often with the belief that certain statistical procedures somehow "control" the effects of other variables, usually called covariates. These procedures do not physically "control" or in any other way affect these covariates. Some people believe this, what Braun (2013) describes as magical thinking. Unfortunately this misnomer has been shared with generations of students. Three procedures are discussed: gain scores, ANCOVA, and propensity matching. All of these provide accurate solutions in certain circumstances to certain research questions. The difficulty is knowing when each of these procedures, and if any of these, is appropriate. These methods are described in more detail using two examples. Graphical models and simulations are used to illustrate how a researcher might decide which to use. Two examples are used to illustrate steps that can be used to help guide this choice.

The following steps can be taken before conducting statistical analysis of any intervention:

1. describe models for how the data might arise,
2. simulate data according to these models, and
3. evaluate different statistical procedures.

A Mathematics Intervention and Review of Procedures

A mathematics intervention example was chosen to represent evaluations where the researcher wants to know if a treatment works, has a prior score measured before the treatment that is on the same scale as the outcome measure, but the researcher cannot randomly allocate people into a treatment and a control condition.

Suppose you want to assess whether a year-long program of a weekly after-school "math club" improves students' scores at the end of the year. Denote being in the club with $Treatment = 1$ and 0 otherwise, the assessment before treatment as *Pre*, and the end of year assessment with *Post* (subscripts will not be used on variables in text or in figures, but all vary by student). There would be complaints if students were randomly assigned to this intervention. First, let's assume students volunteer for the club. Who volunteers for an after-school math club is not random. Assume there is some latent variable, call it *Propensity*, that predicts is the probability of whether a student (or their guardian) volunteers. This is an unmeasured variable. *Propensity* is likely to be influenced by many things, including how much math knowledge the student has already learned and aspects of their home environment (e.g., diligence doing homework). Call this *Knowledge*. *Knowledge* will also affect the two test scores. *Knowledge* and *Propensity* will be influenced by many other variables, and will also influence other variables. Some of these may be observed, but most will not be. For illustration, only simple models are used in this section.

It is often useful to draw relationships using the mathematics concept of a graph (not everyone agrees, see Imbens & Rubin, 2015, p. 22). A graph, in its mathematical sense, is a set of nodes (called vertices in some texts), some of which are connected by edges. In

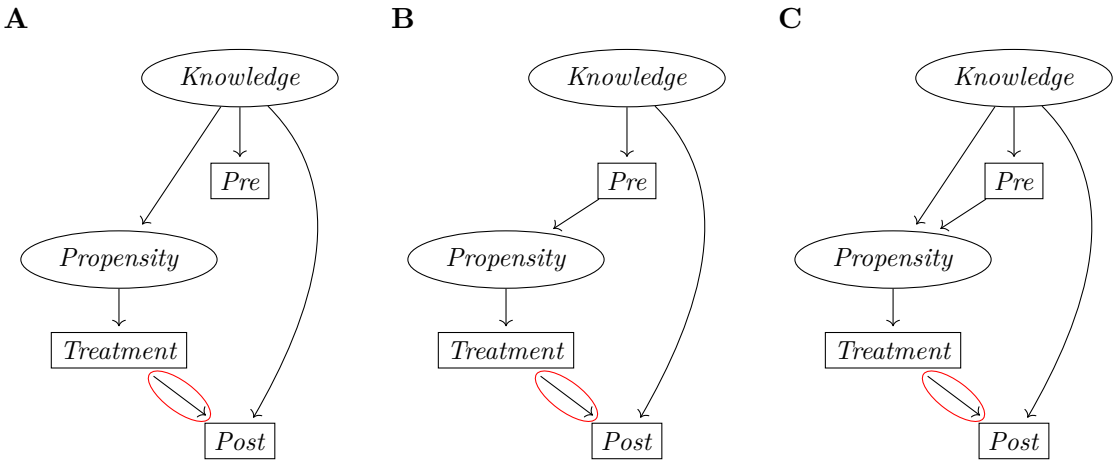

Figure 1. Three data-creation models that vary by how *Knowledge* and *Pre* influence *Propensity*. In Panel A only *Knowledge* influences it, in Panel B only *Pre* influences it, and in Panel C both influence it.

causal models these edges are often shown with an arrow on one end to denote the direction of causality. Following the style used in structural equation modeling, ellipses will be used here for unmeasured constructs and rectangles for observed variables. The seminal reference applying graphs to causal models in science is Pearl (2009), and good introductions include: Elwart (2013), Morgan & Winship (2015), and Pearl et al. (2016). Figure 1A shows the model described above. Each observed variable also includes an error term, which are not shown in order to make the figures less cluttered (they would be small ellipses with an arrow pointed towards the observed variable). Models are simplifications. For example, there will be other variables that influence *Knowledge* and *Propensity*, and are influenced by these. The primary interest for the researcher is estimating the edge *Treatment* → *Post*, enclosed with a red ellipse.

For the model in Figure 1A, the students' scores on *Pre* do not influence whether someone is in the treatment. This might be because treatment allocation is decided before the tests are scored. Figure 1B supposes that the *Pre* scores do influence propensity, and that *Knowledge* does not. This might be if school administrators were deciding who was in the math club based solely on the scores from this assessment, and the other factors influencing *Propensity* were unrelated to *Knowledge* (e.g., whether the student was available when the club occurred). Figure 1C allows both of these to influence *Propensity*. More complex models are considered in Example #2 of this paper.

Review of Statistical Methods using R

Comparing scores on *Post* by *Treatment* using something like a *t*-test is not a good measure of the intervention when the two groups begin systematically different in ways associated with the outcome variable. Differences on the outcome variable could be due either to the intervention or to the pre-existing differences. Three procedures will be considered in this paper: gain scores, ANCOVA, and propensity matching. Each of these is sometimes

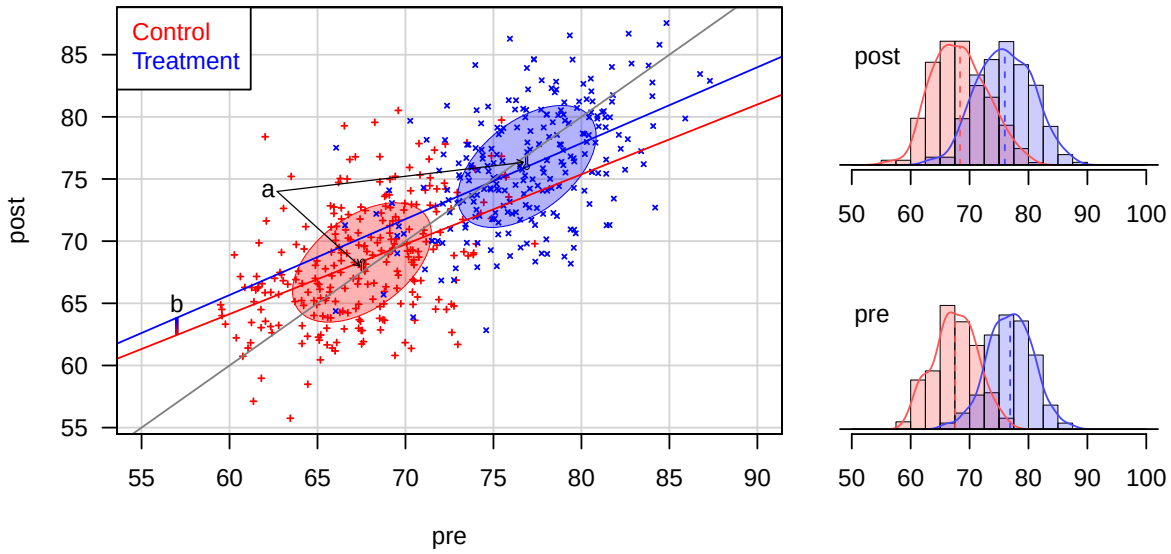

Figure 2. The scatterplot is *post* with *pre* scores ($n = 500$), showing ellipses that are estimated to include 50% of the population data for each group. The lines in the scatterplot are regression lines for each group. The histograms show the distributions of these, with kernel density curves (and means shown with the vertical dashed lines).

described as "controlling for previous performance."

Before doing any inferential statistical modelling, descriptive statistics and visualizations should be done (Wilkinson & the Task Force on Statistical Inference, APA Board of Scientific Affairs, 1999; Wright, 2003). Example plots are shown in Figure 2 for hypothetical data for *post* and *pre* scores for two groups. This shows a scatterplot with ellipses to show where approximately 50% of each group lie. These are made with the `dataEllipse` function of **car** (Fox & Weisberg, 2019). The colored lines are regression lines for each group. The slopes of these are less than one (shown with the grey line, for $Pre = Post$), which is consistent with regression towards the mean (Galton, 1886). The histograms for each variable are shown to the right of the scatterplot. These show the distributions for both groups for both variables, and include a kernel density curve made with R's `density` function. The mean for each group is shown by the vertical dashed lines. Further information about making plots in R can be found in Murrell (2019). The a and b in the plot are discussed below.

The first two procedures are often discussed with reference to Lord's paradox (1967; 1969). Lord described a situation: students' weights, before and after a year of college, where interest was with the gender difference. Lord imagined two statisticians proposing different methods for the analysis. The first statistician proposed subtracting the two weights and comparing means of these gain scores. The second statistician proposed predicting final weight from gender after conditioning on the initial weight using an ANCOVA. The statisticians reached different conclusions. The first statistician found no difference in weight gain, for either females or males, so no gender difference. The second found males weighed more than females after conditioning on pre-weights. Several authors have shown

when and why these approaches can produce different effects (e.g., Hand, 1994; Holland & Rubin, 1983; Kim & Steiner, 2020; Pearl, 2016; Wainer, 1991; Wright, 2006, 2020). Since Lord described this paradox, propensity matching (e.g. Rosenbaum, 2002; Rubin, 2006) has become very popular, though many express concerns that it is sometimes used without due concern (e.g., Pearl, 2009; Sekhon, 2009). Therefore these three procedures will be reviewed. The environment language R (R Core Team, 2019) will be used for simulations, so here the code to estimate these models in R is presented.

Gain scores. The simplest of the procedures considered, computationally, is the gain score method. Let $Gain = Post - Pre$ and it is assumed that these gain scores have approximately the same meaning for each level of *Pre*. It must make sense to equate, for example, Tom's increase from 96 to 99 with Jerry's increase from 47 to 50. Analyses can then be conducted on this variable using *Treatment* with or without the other observed variables. Lord's (1967) first statistician conducted a *t*-test between the two groups on the gain score (i.e., no other covariates). Eqn. 1 shows this as a regression model:

$$Gain_i = Post_i - Pre_i = \beta_0 + \beta_1\, Treatment_i + \cdots + e_i \quad . \tag{1}$$

The procedure estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$, usually by finding these values such that $\sum e_i^2$ is as small as possible (i.e., least squares). If only the single predictor variable *Treatment* is used this is equivalent to a *t*-test. The three dots, $\cdots$, are shown to emphasize that further covariates could be included. Most interest is in $\widehat{\beta}_1$, the value associated with the treatment, and the usual null hypothesis is: H_0: $\beta_1 = 0$. In R, where `Covs` is the set of any covariates beyond the treatment (`Group`) and initial score (`Pre`), this would be:

```
gainmod <- lm(Post - Pre ~ Group + Covs)
```

The `lm` function stands for linear model. The gain scores are calculated on the left of the `~` and the model is on the right. If you have several data sources, you can tell R which of these objects the variable is in. This can be done in a few ways. Suppose those variables are stored in a data frame called `dandata`:

```
gainmod <- lm(Post - Pre ~ Group + Covs, data = dandata)
with(dandata, gainmod <- lm(Post - Pre ~ Group + Covs))
gainmod <- lm(dandata$Post - dandata$Pre ~ dandata$Group + dandata$Covs)
```

The result (`gainmod`) is stored here as an object called `gainmod`. Information from this object can be printed by typing `gainmod`, or by placing it inside functions like `summary(gainmod)` (summary output), `anova(gainmod)` (ANOVA table), `confint(gainmod)` (confidence intervals of the fixed effects), `predict(gainmod)` (predicted values), and `plot(gainmod)` (useful plots for checking assumptions). These functions all find the class of object `gainmod` is (class is `"lm"`) and report these results accordingly. For introductions to using R for statistics see Crawley (2015), Field et al. (2012), Matloff (2020), and Venables & Ripley (2002, R is an implementation of the language S, so books about S are also applicable). For discussions of philosophy behind and history of R see Chambers (2008) for details and Chambers (2009) for an abbreviated discussion. Useful sources

about the programming language include Chambers (1998; 2016), Matloff (2011), Venables & Ripley (2000), and Wickham (2015).

The output from `coef(gainmod)` for the data from Figure 2 is:

```
summary(lm(post-pre ~ group))$coefficients

##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  0.822138  0.2704364  3.040042 2.489938e-03
## group       -1.727181  0.3919780 -4.406321 1.288354e-05

confint(lm(post-pre ~ group))

##                  2.5 %     97.5 %
## (Intercept)  0.2908011  1.3534749
## group       -2.4973151 -0.9570463
```

The intercept is the mean gain for the control group ($\widehat{\beta}_0 = 0.82$). The difference between the two gain scores, shown in the final line, is: $\widehat{\beta}_1 = -1.73$, with $t(498) = -4.41, p < .001$. The coefficient is negative meaning the gain by the treatment group is less than the gain by the control group. The mean for the treatment group is the sum of these two estimates: $0.82 - 1.73 = -0.91$.49. The equivalent t-test provides a little more output. Both show the same t-value and p-value. The group variable

```
t.test(post-pre ~ group, var.equal=TRUE)

##
##  Two Sample t-test
##
## data:  post - pre by group
## t = 4.4063, df = 498, p-value = 1.288e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.9570463 2.4973151
## sample estimates:
## mean in group 0 mean in group 1
##       0.8221380      -0.9050427
```

The gain is positive for the control group and negative for the treatment group. The t-value is negative in the regression output but positive in the t-test output just because of how the t-test function interval represents the group variable. Since the hope that the treatment has a positive effect this should probably be written as a negative t-value. The mean gain score for each group is in Figure 2 by the two lines marked by a with arrows. These lines go from the means of the two variables for each group, shown with a circle in the middle of the ellipse, to the value where the group would have been if there had been no gain (both the horizontal and vertical values being the mean for the *Pre* score). The conclusion from the gain score procedure would likely be that the treatment had a negative effect on achievement.

ANCOVA. Lord's (1967) second statistician conducted an ANCOVA. This is the procedure most often described as *controling* covariates. There have been decades of warnings about the limitations of this procedure (e.g., Kahneman, 1965; Meehl, 1970). The phrase ANCOVA can mean different things to different people, but here it will refer to the following model:

$$Post_i = \beta_0 + \beta_1\, Treatment_i + \beta_2\, Pre_i + \cdots + e_i \quad . \tag{2}$$

If β_2 is fixed at 1, this becomes eqn. 1. This ANCOVA also tests if $\beta_1 = 0$, like eqn. 1, but this β_1 is different. It is the effect after conditioning on *Pre* and the outcome variable is different. More covariates are often added to the regression, further obfuscating the meaning of β_1.

Cox & Donnelly (2011, p. 111) suggest using notation that shows all the variables being conditioned upon when reporting an effect makes the complexity more transparent. If the additional covariates were cov_1, cov_2, and cov_3, the treatment effect could be written as: $\beta_{treatment|pre,cov1,cov2,cov3}$. Sometimes further variables are added to regressions/ANCOVAs with the hope that this somehow gets closer to the isolating the causal impact of *Treatment* on *Post*. If the convention were to report estimates with this longer notation it might make clearer that adding more variables is unlikely to simplify the meaning of the effects. Adding or removing a covariate changes the meaning of the parameter being estimated. It is important to think carefully about which covariates to include based on their role in the overall causal model and to evaluate the performance of these statistical models with simulated data sets. In R, where `Covs` is the set of any covariates other than *Treatment* and *Pre* scores, this would be:

```
ANCOVAmod <- lm(Post ~ group + pre + Covs)
```

The results can be found using the same R functions listed above because both are produced with the `lm` function so both produce class `"lm"` objects.

```
summary(lm(post ~ group + pre))$coefficients

##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 28.8064260 3.28627376  8.765681 2.953064e-17
## group        2.1354890 0.58205738  3.668863 2.699493e-04
## pre          0.5857928 0.04849739 12.078851 1.241144e-29

confint(lm(post ~ group + pre))

##                  2.5 %    97.5 %
## (Intercept) 22.3497242 35.263128
## group        0.9918926  3.279085
## pre          0.4905076  0.681078
```

After conditioning on *Pre*, the treatment effect is positive: $\widehat{\beta}_1 = 2.14$, $t(497) = 3.67, p < .001$. Thus, the typical conclusion from this analysis would be that the treatment

had a positive effect on achievement. Because the gain score approach and the ANCOVA approach lead to different conclusions, both cannot be write, which is why Lord (1967) called this a paradox.

Propensity Score Matching. Trying to reach causal conclusions when people are not randomly allocated into groups is difficult (Campbell & Stanley, 1963). Propensity matching attempts to create an accurate model of who chooses (or is chosen) to be in the intervention, and then uses this information to compare people with similar propensities to be treated. Propensity matching was developed in a series of papers by Rosenbaum and Rubin. The seminal textbook is Rosenbaum (2002) and many of their contributions have been re-published in Rubin (2006). The phrase propensity matching now applies to several different approaches that aim to achieve these goals. An excellent introduction to propensity matching, which covers many of these approaches, is Leite (2017). He describes six steps for propensity score analysis (p. 7).

1. Prepare data. This includes choosing which covariates to include. This requires knowledge of how the different variables may relate and therefore knowledge of the research domain. Leite includes dealing with missing values in this step.
2. Propensity score estimation. This involves using the covariates to estimate the probability that each person will be in the treatment group (e.g., with logistic regression).
3. Propensity score method implementation. The analyst decides whether to create "matched" groups or use some other technique (e.g., weighting according to the propensity score in a regression).
4. Covariate balance estimation. Depending on the previous step, the analyst evaluates how successful the implementation is. For example, if matched groups were created, are their scores similar on the covariates?
5. Estimate the effect of a treatment. The method depends on previous steps.
6. Sensitivity analysis. A variety of methods can be used here, including the focus of Example #2 in this paper, seeing if the results vary by whether particular covariates are included or not (#2a), and varying assumptions of the data-creation model (#2b).

Leite (2017) goes through each of these steps. He shows different methods that can be used for each and how these are implemented in R. There are several statistics packages for performing propensity matching and estimating the treatment effects (see `www.biostat.jhsph.edu/~estuart/propensityscoresoftware.html`). Keller and Tipton's (2016) review of R propensity matching packages is particularly relevant. Currently (10 April 2020) the CRAN task view for social science (`https://cran.r-project.org/web/views/SocialSciences.html`) lists the **Matching** (Sekhon, 2011), **MatchIt** (Ho et al., 2011), **optmatch** (Hansen & Klopfer, 2006), and **PSAgraphics** (Helmreich & Pruzek, 2009) packages for propensity matching. Other packages of note include **twang** (Ridgeway et al., 2017) and **CBPS** (Fong et al., 2019).

Here the **Matching** package (Sekhon, 2011) and its `Match` function are used. To use a package you first must install it (e.g., `install.packages("Matching")`). This can be done just once so it is on your computer and whenever you want to make sure that you have the most up-to-date version. Next, you load it (e.g., `library(Matching)`), and this your need to do within each R session in which you want to use the package's functions.

The function can be called just using its name (here `Match`), but if you have several packages open you might have another function also named `Match`. If you are not sure it is safest to use `Matching::Match` to call functions. In this formula `Covs` may or may not include `Pre`, depending on what the analyst believes. The propensity score estimation is done with a logistic regression and the matching and estimation steps are done by the `Match` function. The goal of this paper is not to argue for or against a particular propensity matching approach. There is still much debate about this. Leite's (2017) coverage seems well balanced.

```
propval <- glm(Treatment ~ Covs,family=binomial)$fitted
library(Matching)
PrMatmod <- Match(PostTest,Treatment,propval,
      estimand="ATE",BiasAdjust=TRUE,ties=TRUE)
```

The `Match` function has several options (type `?Match` in R to see these once the package is installed). The `estimand="ATE"` means the estimate is the average treatment effect. This is the average estimated difference due to treatment for both those given the treatment and those not give the treatment. The user also has the option to estimate the treatment effect separately for those in the treatment condition or for those in the control condition. These would be valuable for different applied problems: the former to estimate the effect for those likely to sign up for the treatment and the later the value if you could encourage those who would not normally sign up to sign up. If you type `summary(PrMatmod)` the computer will print the results. Further information can be extracted from this object. Type `PrMatmod$est` to get the estimated treatment effect. To see other information that can be extracted type `str(PrMatmod)`. It is usually used with many variables to predict the group allocation, but for illustration here it is done with just *pre*. Here the results (estimate $= 2.96, z = 2.06, p = .039$) are consistent with the ANCOVA approach, showing a positive effect from the treatment.[1]

```
propval <- glm(group ~ pre,family=binomial)$fitted
suppressPackageStartupMessages(library(Matching))
PrMatmod <- Match(post,group,propval,
      estimand="ATE",BiasAdjust=TRUE,ties=TRUE)
summary(PrMatmod)

##
## Estimate...  2.9567
## AI SE......  1.4326
## T-stat.....  2.0639
## p.val......  0.03903
##
## Original number of observations..............  500
## Original number of treated obs...............  238
## Matched number of observations...............  500
```

[1]The output says t-value, but the author uses `pnorm` in his `summary.Match` function.

```
## Matched number of observations  (unweighted).  572
```

Propensity matching methods can be used in conjunction with the gain score and ANCOVA approaches. For the gain score approach, the gain score would simply be used as the dependent variable. Propensity matching can be combined with ANCOVA; after either matching or weighting for the propensity scores, by using additional covariates (including those used to estimate propensity) to predict the outcome. This approach, sometimes called doubly robust, can increase the power for detecting a treatment effect. These extensions are not explored in the current paper. Further, there are many other methods that can be used for addressing inadequacies of the basic ANCOVA (e.g., non-linearity, measurement error, clustered data).

1. For more information about conducting propensity matching in R see Leite (2017).
2. The gain score and ANCOVA procedures use regression. For much more information about using conducting regressions in R and other aspects of regression see Fox & Weisberg (2019) and Matloff (2017).

Examples using Simulation

Four simulations are presented as examples of how a researcher might go about deciding which approaches to use. The key steps are: a) creating a set of plausible data models, b) creates lots of data sets for these, c) conduct the statistical procedures that you wish to compare on these, and d) compare these findings. These examples are to illustrate this approach, but the topics were chosen because they are important in deciding among procedures: how people are allocated to groups and colliders.

Varying How Students are Allocated into Groups

Simulation methods are used here to explore potential bias for the three statistical procedures for the three data models shown in Figure 1. R (R Core Team, 2019, Version 3.6.1) is used. Other software (e.g., Python, SAS, SPSS, Stata) could have been used, but using R allows use of its propensity matching packages (e.g., Keller & Tipton, 2016) and it is freely available to all readers. Simulation methods allow the data-creation models to be varied to examine further research questions. In this example, the relationships among *Knowledge*, *Pre*, and *Treatment* are varied.

Example #1a: Having the true effect equal zero. Data were created for each of the models depicted in Figure 1 and the three statistical models applied. For the gain score model, *Treatment* was used to predict the difference in the two scores (equivalent to a *t*-test on the gain scores). For the ANCOVA *Treatment* and *Pre* were used to predict *Post*. For the propensity matching, *Pre* was used to predict *Treatment*, propensity values were taken for this, and Sekhon's (2011) `Match` function used to estimate the effect of treatment. Normally propensity matching is used with a larger number of variables. This is done in Example #2. To make these results easier to interpret, in Example #1a the true treatment

Table 1
The mean t *and* z *values for treatment effect for the different statistical procedures for the different data-creation models from Figure 1 (Example #1a). 95% confidence intervals in parentheses. These are based on 10,000 replications.*

	Statistical Procedure		
	Gain Score	ANCOVA	Prop. Matching
Figure 1A	-0.01 (-0.03, 0.01)	-2.41 (-2.43, -2.39)	-2.09 (-2.11, -2.07)
Figure 1B	2.84 (2.82, 2.86)	-0.01 (-0.03, 0.01)	-0.05 (-0.07, -0.02)
Figure 1C	1.92 (1.90, 1.94)	-1.72 (-1.74, -1.70)	-1.50 (-1.52, -1.48)

effect is zero. Thus, the correct answer for all of these models is $\beta_1 = 0$ so the mean of unbiased t and z values should be near zero. In #1b different effect sizes are used.

Values for the latent variable *Knowledge* were drawn from a unit Normal distribution, i.e., $Knowledge \sim N(\mu = 0, \sigma = 1)$. *Pre* was drawn from $Knowledge + N(\mu = 0, \sigma = 1)$, then standardized so that it has a mean of 0 and standard deviation of 1. *Propensity* was drawn from $Knowledge + N(\mu = 0, \sigma = 1)$, $Pre + N(\mu = 0, \sigma = 1)$, and $Pre + Knowledge + N(\mu = 0, \sigma = 1)$, for panels A, B, and C, respectively, from Figure 1. Each of these was standardized. Treatment was decided by a Bernoulli process with probability equal to the propensity variable; so like the flip of a weighted coin. *Post* is drawn from $Knowledge + N(\mu = 0, \sigma = 1)$, then standardized. It is not affected by *Treatment*, so the true treatment effect is zero. The three statistical models were estimated. This was repeated 10,000 times so that the estimates are quite precise. The code is in the Appendix.

Table 1 shows the mean t (from the gain score and ANCOVA output) and z values (from the propensity matching output) for this simulation. Because this is a simulation it is known that the true effect is 0. The gain score model provides unbiased estimates for the Figure 1A, where *Pre* does not influence the propensity to be in the treatment group and therefore does not influence being in the treatment group. The other two statistical methods provide biased estimates here. They both suggest the treatment had a negative effect. The opposite occurs for data produced according to Figure 1B, where *Knowledge* does not affect propensity. Here ANCOVA and propensity matching produce mean estimates near zero, but the gain score procedure is biased, suggesting a positive effect for the treatment. When both *Pre* and *Knowledge* affect *Propensity*, as in Figure 1C, the gain score procedure estimates a positive treatment affect while the ANCOVA and propensity matching procedures estimate negative treatment effects. The findings for the gain score and ANCOVA procedures are consistent with Wright (2006) and the graphical models of Pearl (2016). See also discussion in Holland & Rubin (1983) and Steiner et al. (2011). The researcher would need to decide which of the data-creation models in Figure 1 is most appropriate in order to decide if any of these three statistical models should be used. Some methods for this are described in Lockwood & McCaffrey (2020).

Example #1b: Varying the True Effect Size. Example #1a shows when each procedure produces unbiased estimates when the true effect is zero. Here this is extended to negative and positive effect sizes. This allows examination of the procedures' power to

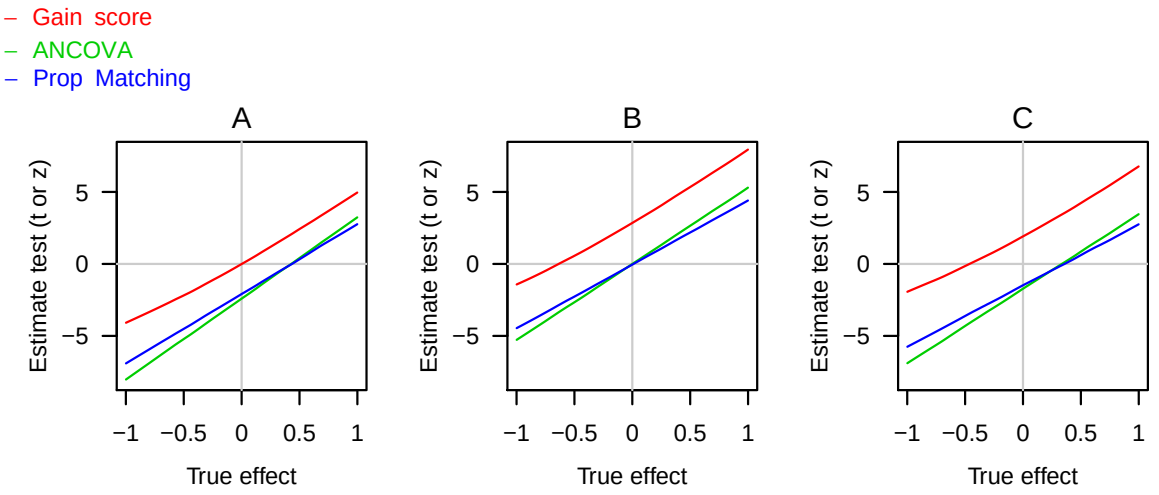

Figure 3. Predicted test effects for all three situations in Figure 1 for each statistical procedure. The predicted values were found using lowess (df=2, span=.25).

detect effects to be examined. The simulation was repeated but when creating the *Post* values a treatment effect was included. The treatment effect was drawn from a uniform distribution from -1 to +1. For those in the treatment condition this was added to the knowledge latent variable (distributed $N(\mu = 0, \sigma = 1)$) and a normally distributions error term ($\mu = 0, \sigma = 1$). The resulting variable was standardized to have a mean of 0 and standard deviation of 1. One hundred thousand replications were done for each situation (A, B, or C from Figure 1 by procedure (gain score, ANCOVA, or propensity matching).

Figure 3 was made predicting each statistical test (t for gain scores and ANCOVA, z for the propensity matching procedure). The loess procedure was used with a span of only 25% and a quadratic curve. This allows non-linearity to be predicted. As can be seen the predicted values are linearly related to the effect size. The left panel shows situation A, where *Pre* scores have no influence on propensity to receive treatment. Here the gain score has a mean test statistic of zero when there is no true effect, and a negative mean with negative true effects and positive means when the true effect is positive. The ANCOVA and propensity matching give downwardly biased estimates of the treatment effect (the effect is downward since those with low knowledge scores had high propensity scores). This means that for small positive effects the procedure was mean test effects in the opposite direction, what Gelman & Carlin (2014) call Type S (for sign) errors. For situation B, where the *Pre* score does influence propensity to receive treatment, the ANCOVA and propensity matching perform well and the gain score more is biased over-estimating the treatment effect (again, it is overestimated since low *Pre* scores had the higher propensity for being in the treatment). For situation C, the gain score procedure and the ANCOVA and propensity matching procedures are biased in opposite directions. This shows the importance of understanding the allocation mechanism when deciding how to analyze data.

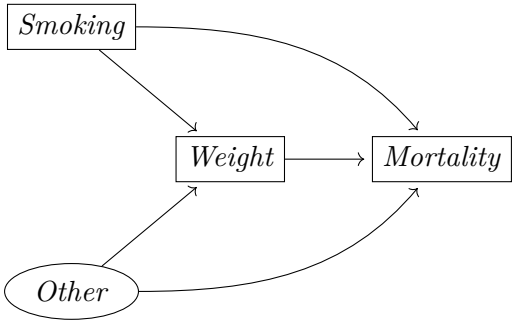

Figure 4. A causal model for the smoking-birth weight paradox.

Causal Models, Colliders, and More Covariates

Causal models, and in particular something called a collider, will be the focus of this second example. A common example used to introduce colliders is the smoking-birth weight paradox (e.g. Pearl et al., 2016). Medical researchers are interested in how a mother smoking affects many things including infant mortality and birth weight. Infants of smokers appear to have higher infant mortality rates, but researchers found that conditioning on birth weight can reverse this effect. Researchers speculated that smoking might somehow protect low-weight infants. However, examining the possible causal models of this situation, which Pearl et al. (2016) do, reveals the likely reason.

Figure 4 shows observed variables for smoking, birth weight, and infant mortality. There are other causes for both low birth weight and infant mortality and these are depicted with an unmeasured variable called *Other*. Because many of these conditions affect infant mortality more than smoking does, it means conditioning on birth weight means you are comparing infants of mothers who smoked with infants who have conditions with higher infant mortality rates.

If the purpose is to measure the direct effect of smoking on infant mortality ($Smoking \rightarrow Mortality$), it is important to consider backdoor (also called indirect) paths between *Smoking* and *Mortality*. A path is any set of edges between two nodes where no node is included more than once (so non-recursive). There are two backdoor paths in Figure 4:

1. $Smoking \rightarrow Weight \rightarrow Mortality$.
2. $Smoking \rightarrow Weight \leftarrow Other \rightarrow Mortality$.

Paths can be either blocked or unblocked. If they are unblocked it means information can flow along them confounding measurement of the direct path. Therefore, often a goal of choosing covariates is to block backdoor paths. How detrimental the effects of an unblocked door path are depends on the product of the path coefficients (Loehlin, 1998). To understand blocking paths it is necessary to consider three ways in which three variables, call them $X = eatcake$, $Y = behappy$, and $Z = smile$, can be causally related:

Chain: X (*eat cake*) $\rightarrow Y$ (*be happy*) $\rightarrow Z$ (*smile*),
Fork: X (*eat cake*) $\leftarrow Y$ (*be happy*) $\rightarrow Z$ (*smile*), and
Collider: X (*eat cake*) $\rightarrow Y$ (*be happy*) $\leftarrow Z$ (*smile*).

Pearl (2009, pp. 16–17) describes two rules to determine if the path between two nodes, X and Z, is blocked and how this is affected by conditioning on the middle variable Y.

Pearl's first rule is that if a path contains a chain or a fork, it is unblocked unless the middle variable is conditioned upon. Chains and forks are associated with the phrases mediation and spurious correlation in the education literature. An example of an unblocked chain is that an exercise pamphlet can lead to students planning to exercise and this can lead to more exercise (Hill et al., 2007). If you prevent students from the planning phase then giving participants a pamphlet is not as effective. A common textbook example of a fork is the positive association between ice cream consumption and murder in cities (Peters, 2013, see also, Vigen, 2015). Warm weather causes both of these to increase. If you could condition on the weather by looking at one particular weather (e.g., sunny and 83°F), this would block the path and assuming no other effects are present, the correlation would be near zero. In Figure 4, the path *Smoking* → *Weight* → *Mortality* includes a chain (middle variable *Weight*) and therefore begins unblocked. Much of the justification for using covariates is to block paths like this. One way to block this path would be to condition upon *Weight*. While it may seem useful to block this path, if the goal is to measure the complete causal effect of smoking, this path simply shows how the effects of smoking may be partially mediated by birth weight.

Pearl's second rule is that a path with a collider begins block, but is unblocked if the collider, or any variable influenced by it (called *descendants* in graph theory terminology), is conditioned upon. Colliders are less discussed in the education literature than forks and chains. Wright (2017) uses a river metaphor to describe a collider. Imagine two tributaries arriving from different directions at a deep sink hole. The hole is the collider. The water from each would not reach the other; the path is blocked. If the hole is filled, water could flow between the tributaries and the path would be unblocked. *Weight* is a collider in: *Smoking* → *Weight* ← *Other* → *Mortality* because it is influenced by both smoking and other causes. One method for blocking the first path (conditioning on *Weight*) will unblock the second path. It is unblocking this second path that creates the illusion that smoking decreases infant mortality (illusory if one thinks the ANCOVA measures the effect of smoking on infant mortality). The smoking-birth weight paradox example was shown because it clearly shows the effects of conditioning on a collider.

The causal model in Figure 4 is similar to ones in education. Wright (2017) provides an education example. It involves a collider that is measured before the treatment. Figure 5 shows the data-creation model that could be assumed when educational systems attempt to estimate school effectiveness. Suppose this is for estimating the effectiveness of a 9th grade class. The effectiveness of the school is influence by environmental factors like the economics of the neighborhood. These also influence previous schooling and therefore grades from earlier years (denoted *Pre*). Characteristics of the student and their family also influence grades before the 9th grade and afterwards (denoted *Post*). The critical edge to estimate school effectiveness is the one from *School* → *Post*. The backdoor path is:

$$\textit{School} \leftarrow \textit{Environment} \rightarrow \textit{Pre} \leftarrow \textit{Knowledge} \rightarrow \textit{Post} \quad .$$

The variable *Pre* is a collider so conditioning on this unblocks this backdoor path thereby causing problems for estimating the direct effect. Using methods often used in the US to measure effectiveness, Wright (2017, 2018) showed that the estimated effectiveness can be

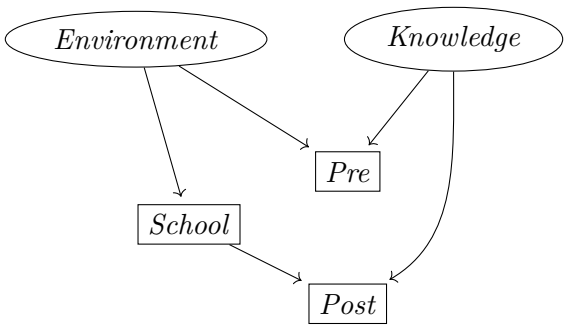

Figure 5. A model for how data could arise that is used to estimate school effectiveness (adapted from Figure 1 of Wright, 2017).

negatively correlated with true effectiveness.

Example #2a: How Manifest and Latent Variables Relate. The example used in this section is more complex than Examples #1a and #1b, to reflect–at least to some extent–the complexity of causal models applicable for many education research situations. It is common to measure several other variables with the hope of blocking (and keeping blocked) all backdoor paths thereby isolating the direct effect.

Figure 6 shows the data-creation model assumed. There are three latent variables (*Environment*, *Knowledge*, and *Grit*) and two key observed variables (whether the person had the *Treatment* and the *Post* score). It is assumed here that *Treatment* is influenced by the *Environment* the student is in (e.g., geographic location). In addition, it is assumed that there are three measured variables that are related to each of the latent variables. For the simulation each of these observed variables have been created to be correlated approximately $r = .50$ with their associated latent variable, but how they are associated with the latent variable differs. One influences the latent variable (e.g., *e1* $\rightarrow$ *Environment*), one is influenced by the latent variable (e.g., *Pre* $\leftarrow$ *Knowledge*), and for the other one, another variable (depicted just with a circle) influences both (e.g., *g3* $\leftarrow \bigcirc \rightarrow$ *Grit*). While this model looks complex, like all models in social science it is still a simplification. Several of the nodes listed likely influence others (e.g., *Environment* $\rightarrow$ *Grit*; *Grit* $\rightarrow$ *Pre*), there are many other constructs that could play a role, and there are many other variables that could be measured. It is worth noting that observed variables are not placed along the paths from *Treatment*, *Knowledge*, and *Grit* to *Post*, and there are no nodes just affecting *Treatment*. These effects could provide additional ways to measure the treatment effect. There are several ways to address measurement of interventions and readers are encouraged to consult textbooks devoted to this (e.g., Imbens & Rubin, 2015; Morgan & Winship, 2007; Pearl et al., 2016).

The variable *Pre* is named in Figure 6 as opposed to just calling it $x2$ because of its central role in the gain score model. It is assumed that it is on the same scale as *Post* (in the simulation both have means of zero and standard deviations of one). It is assumed that $Post - Pre$ is meaningful and that this difference represents the same *gain* throughout the span of *Pre*. This is a vital assumption for the gain score approach. *Pre* is influenced

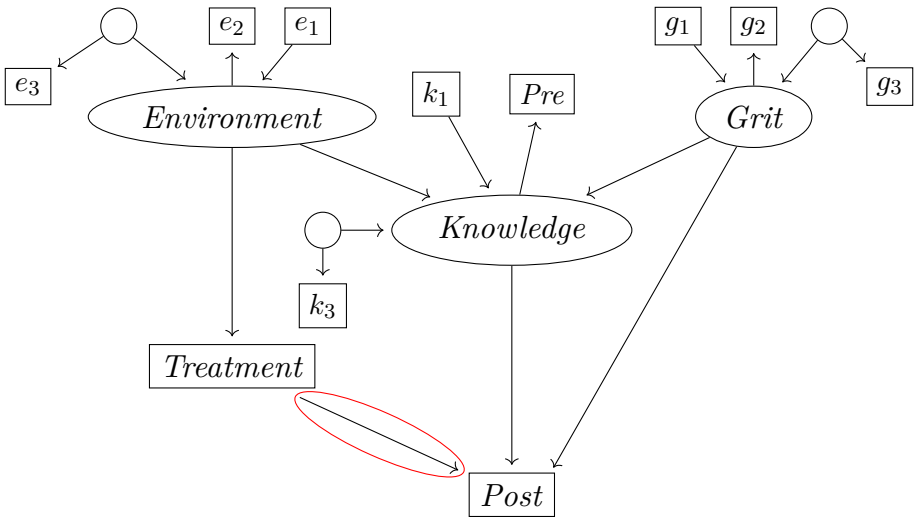

Figure 6. The data model with both latent and observed variables. The direct effect from *Treatment* to *Post*, enclosed in the red ellipse, is the construct the researchers wish to measure accurately. To make the figure less cluttered the individual node error variables are not shown.

by *Knowledge*. In graph theory terminology it is a descendant of *Knowledge*. According to Pearl's second rule, conditioning on *Pre* unblocks the path *Treatment* $\leftarrow$ *Environment* $\rightarrow$ *Knowledge* $\leftarrow$ *Other* $\rightarrow$ *Post* because it is a descendant of a collider (Example #2b explores this further).

As with Example #1a, the data were created so that true treatment effect is 0. *Post* is the sum of *Knowledge*, *Grit*, and a $N(\mu = 0, \sigma = 1)$ error term, and then standardized. To create the observed variables so that they have correlations of approximately $r = .5$ with their associated latent variable required a few steps (this could be done in several ways). These will be shown for k_1, Pre, k_3. `kph` is the unnamed latent variable influencing *Knowledge* (`know` in the code). The R code is:

```
kph <- rnorm(n)
k1 <- rnorm(n)
k3 <- scale(kph + .71*rnorm(n))
know <- scale(.85*kph + .6935*k1 + .6*other + .6*env)
pre <- scale(sqrt(.25)*know + sqrt(.75)*rnorm(n))
```

First, the variables `kph` and `k1` are drawn from unit Normal distributions. `k3` is created by adding `kph` and a unit Normal variable multiplied by slightly more than 1, and this sum standardized. The latent variable `know` is the standardized sum of `kph`, `k1`, `grit`, and a Normal variable, each weighted to produce the desired correlation. `Pre` is the standardized sum of weighted `know` and a Normal variable. The weights were chosen so that the correlation of the variables was .5. As shown in Example #2b, the results are sensitive to these weights.

The propensity to be in the treatment condition was the quantile (the `qrank` function in the appendix) of the *Environment* variable (`env`). The contributions from *Environment* and *Other* to *Knowledge* are equal, and a Normally distributed random error is also added. The code is in the Appendix.

The study had $n = 200$ and there were 10,000 replications per condition. This number of replications was used so that the standard error for the z value for each statistical model was approximately 0.01 (and thus the widths of the 95% confidence intervals about .04).

Table 2

The mean t (gain score and ANCOVA) and z values when using all the covariates, excluding each set of three covariates, and excluding each individual covariate for Example #2a. The data were created so that the true treatment effect was zero.

	Gain Score	ANCOVA	Prop. Match
All covariates included	0.00	-0.94	-0.60
Without the three Environment Variables	-0.03	-1.98	-1.31
without just $e1$	-0.03	-1.31	-0.84
without just $e2$	-0.03	-1.13	-0.73
without just $e3$	-0.03	-1.31	-0.84
Without the three Knowledge Variables	-0.01	-0.99	-0.55
without just *k1*	-0.03	-0.89	-0.59
without just *Pre*		-1.13	-0.63
without just $k3$	-0.03	-0.88	-0.59
Without the three Grit Variables	-0.01	-0.70	-0.56
without just $g1$	-0.02	-0.88	-0.61
without just $g2$	-0.03	-0.95	-0.62
without just $g3$	-0.03	-0.89	-0.61

The mean t (for gain score and ANCOVA) and z (for propensity matching) values for the treatment are shown in Table 2. The first line shows the mean values when all nine observed variables are used as covariates. The mean for the gain score model is approximately zero, so provides nearly unbiased estimates. It is important, however, not to conclude that the gain score methods works well for the data-creation model in Figure 6. The analyst should test the sensitivity of these findings varying the strength of the different relationships (Leite's [2017] sixth step). This is shown in Example #2b.

Both the ANCOVA and propensity matching procedures are biased, suggesting the treatment has a negative effect. As with Examples #1a and #1b, the ANCOVA procedure is more biased than propensity matching. Next, sets of covariates were excluded. For the gain score model excluding any variable created a slight downward bias. For the ANCOVA and propensity matching procedures excluding the three observed variables related to *Knowledge* increased the bias further. The increased bias, for both these statistical procedures, was most evident for $e1$ and $e3$, but excluding each of the three increased the bias. This shows all three of these variables are useful to include in this situation for these statistical procedures.

Excluding the three observed variables associated with *Knowledge* slightly increased the bias for ANCOVA and slightly decreased it for propensity matching. However, the effect of excluding the three individual observed variables was different. For both ANCOVA and

propensity matching, excluding $k1$ and $k3$ decreased the bias, but excluding *Pre* increased the bias, substantially for ANCOVA. Excluding all the *Grit* variables decreased the bias, though excluding just $g2$ had only minimal effect for ANCOVA and none of these greatly affected the propensity matching bias.

Example #2b: Improving the Pre score as a measure of Knowledge. Leite's (2017) sixth step for propensity matching–exploring the sensitive of your conclusions to variations–is relevant for all simulations. Example #2a appears to show that the gain score model may be better for the data-creation model in Figure 6. However, it is important to examine how this conclusion is affected by changing the weights used to create the data.

Here only one variation is chosen. Many people create assessments to accurately measure *Knowledge* and may combine many such assessments into one *Pre* score. This aggregate score might correlate with *Knowledge* more highly than the $r = .5$ in Example #2a. Only one change was made to the code. This line:

```
pre <- scale(sqrt(.25)*know + sqrt(.75)*rnorm(n))
```

is changed to this:

```
pre <- scale(sqrt(.75)*know + sqrt(.25)*rnorm(n))
```

The correlation between *Pre* and *Knowledge* goes up to about $r = .87$. The simulation was repeated and the results shown in Table 3. Now the gain score model produces positively biased estimates, and the ANCOVA and propensity matching procedures produce negatively biased estimates, though these are not as extreme as with Example #2a. As with the previous example, the propensity matching methods is less biased than ANCOVA.

Table 3

The mean t (gain score and ANCOVA) and z values when using all the covariates, excluding each set of three covariates, and excluding each individual covariate for Example #2b. The data were created so that the true treatment effect was zero.

	Gain Score	ANCOVA	Prop. Match
All covariates included	0.73	-0.29	-0.23
Without the three Environment Variables	1.56	-0.54	-0.36
without just $e1$	1.02	-0.36	-0.26
without just $e2$	0.87	-0.31	-0.22
without just $e3$	1.01	-0.36	-0.26
Without the three Knowledge Variables	0.70	-0.99	-0.14
without just *k1*	0.73	-0.19	-0.16
without just *Pre*		-1.08	-0.20
without just $k3$	0.74	-0.19	-0.16
Without the three Grit Variables	0.70	0.15	0.05
without just $g1$	0.74	-0.11	-0.12
without just $g2$	0.75	-0.20	-0.17
without just $g3$	0.74	-0.11	-0.12

Summary

Gain score, ANCOVA, and propensity matching procedures are all used with the hope of isolating the causal effect of an intervention on an outcome measure. Brief introductions were provided for these included R code and output. No statistical procedure can guarantee to isolate this effect in all situations, but each can be useful depending on the model that led to the data. When more effects are added to the data-creation model it can become impossible to block all backdoor paths with any statistical procedure. There are situations where no statistical procedure can accurately estimate a treatment effect. The goal is to choose which procedure is most accurate and to warn readers about this limitation. It is important to make readers aware of the assumptions made when trying to reach causal conclusions.

Four simulations were presented to provide examples for how researchers could decide among these procedures. The overall purpose of this paper is to show that it is important to consider how the data may have arisen, and Examples #1a and #1b focus on understanding how people are allocated to groups. If there are several different plausible models that could account for the data, it is worth examining several of these using simulation methods as was done here. Examples #1a and #1b showed that that there are situations where the gain score model does better than the ANCOVA and propensity matching procedures, and situations where it does worse. In particular, if the covariate influences the propensity to be in the treatment condition, the gain score method is biased, but the ANCOVA and propensity matching methods are not. The results show the opposite is true when other variables influence propensity, but the covariate does not. These findings are consistent with Holland & Rubin (1983); Pearl (2016); Wright (2006). The purpose of Examples #2a and #2b were to show that the choice of covariates is important. They showed that it is important to consider that way in which different observed variables may relate to the latent variables.

Unfortunately, the analyst will not know the true causal model underlying the data. Assumptions must be made in order to make causal conclusions (Cartwright, 2014). This is particularly true when random assignment is not used. The analyst can try several causal models to test if the statistical approach is sensitive to these changes (Leite's sixth step). Researchers should be prepared for reviewers and readers to propose their own causal models. Enough information should be provided to allow these peers to simulate data to show if the statistical procedure used would be appropriate for their choice of causal model. When there are differences, it is often necessary to conduct further research designed to evaluate which set of causal models is more appropriate. This can be time-consuming. Statistical procedures for intervention studies without random allocation are not simple. Campbell & Stanley (1963) describe several threats to the validity. Some, but not all, of these are eased by using random assignment, which they encourage when practical.

It is important to conclude by stressing that plotting the data and exploratory/descriptive data analysis is an important step (e.g., Figure 2). However, the the choice of plots and descriptive statistics is also often influenced by the assumptions made. If you are uncertain which procedures to use, you can use multiple approaches, but you should describe these in the write-up (i.e., not use several and just report the one with the p-value most likely to lead to publication). Steegen et al. (2016) take this to an ex-

treme. They advise analysts to run many potential models and then report the distribution of results and weight them for plausibility. Here, the advice is first think about how the data may have arise to try to limit the choice to a small number of plausible models.

References

Braun, H. I. (2013). Value-added modeling and the power of magical thinking. *Ensaio: Evaluation of Public Policies in Education [Brazil]*, *21*, 115–130. doi: 10.1590/S0104-40362013000100007

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago, IL: Rand McNally.

Cartwright, N. (2014). Causal inference. In N. Cartwright & E. Montuschi (Eds.), *Philosophy of social science* (pp. 308–326). Oxford, UK: Oxford University Press.

Chambers, J. M. (1998). *Programming with data: A guide to the S language.* New York, NY: Springer.

Chambers, J. M. (2008). *Software from data analysis: Analysis with R.* New York, NY: Springer.

Chambers, J. M. (2009). Facets of R. *R Journal*, *1*(1), 5–8. Retrieved from `https://journal.r-project.org/archive/2009-1/RJournal_2009-1_Chambers.pdf`

Chambers, J. M. (2016). *Extending R.* Boca Raton, FL: CRC Press.

Cox, D. R., & Donnelly, C. A. (2011). *Principles of applied statistics.* Cambridge, UK: Cambridge University Press.

Crawley, M. J. (2015). *Statistics: An introduction using R* (2nd ed.). Chichester, UK: Wiley.

Elwart, F. (2013). Causal graphical models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 245–273). New York: Sage Publications.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R.* London, UK: Sage Publications.

Fong, C., Ratkovic, M., & Imai, K. (2019). **CBPS**: Covariate balancing propensity score [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=CBPS` (R package version 0.21)

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Thousand Oaks, CA: Sage.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263. Retrieved from `www.jstor.org/stable/2841583`

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*, 641–651. doi: 10.1177/1745691914551642

Hand, D. J. (1994). Deconstructing statistical questions. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *157*, 317–356. doi: 10.2307/2983526

Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, *15*, 609–627.

Helmreich, J. E., & Pruzek, R. M. (2009). **PSAgraphics**: An R package to support propensity score analysis. *Journal of Statistical Software*, *29*(6), 1–23. Retrieved from `http://www.jstatsoft.org/v29/i06/`

Hill, C., Abraham, C., & Wright, D. B. (2007). Can theory-based messages in combination with cognitive prompts promote exercise in classroom settings? *Social Science & Medicine*, *65*, 1049–1058. doi: https://doi.org/10.1016/j.socscimed.2007.04.024

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). **MatchIt**: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28. Retrieved from `http://www.jstatsoft.org/v42/i08/`

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement* (pp. 3–35). Hillsdale, NJ: Erlbaum.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction.* New York, NY: Cambridge University Press.

Kahneman, D. (1965). Control of spurious association and the reliability of the controlled variable. *Psychological Bulletin*, *64*, 326–329.

Keller, B., & Tipton, E. (2016). Propensity score analysis in R: A software review. *Journal of Educational and Behavioral Statistics*, *41*, 326–348. doi: 10.3102/1076998616631744

Kim, Y., & Steiner, P. M. (2020). Gain scores revisited: A graphical models perspective. *Sociological Methods & Research*, *0*, 0049124119826155. Retrieved from `https://doi.org/10.1177/0049124119826155` doi: 10.1177/0049124119826155

Leite, W. (2017). *Practical propensity score methods using R.* Thousand Oaks, CA: Sage Publications.

Lockwood, J. R., & McCaffrey, D. F. (2020). Using hidden information and performance level boundaries to study student-teacher assignments: Implications for estimating teacher causal effects. *Journal of Royal Statistical Society (A).*

Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path, and structural analysis.* (Third ed.). Mahwah, NJ: Erlbaum.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *68*, 304–305. doi: 10.1037.h0025105

Lord, F. M. (1969). Statistical adjustments when comparing preexisting groups. *Psychological Bulletin*, *72*, 336–337. doi: 10.1037/h0028108

Matloff, N. (2011). *The art of R programming.* San Francisco, CA: No Starch Press, Inc.

Matloff, N. (2017). *Statistical regression and classification: From linear models to machine learning.* Boca Raton, FL: CRC Press.

Matloff, N. (2020). *Probability and statistics for data science: Math + R + data.* Boca Raton, FL: CRC Press.

Meehl, P. E. (1970). Nuisance variales and the ex post facto design. In M. Radner & S. Winokur (Eds.), *Minnesota studies in the philosophy of science: Vol IV. Analysis of theories and methods of physics and psychology* (pp. 373–402). Minneapolis, MN: University of Minnesota Press.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research.* Cambridge, MA: Cambridge University Press.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). Cambridge, MA: Cambridge University Press.

Mosteller, F., & Boruch, R. (Eds.). (2002). *Evidence matters: Randomized trials in education research.* Washington, DC: Brookings Institution Press.

Murrell, P. (2019). *R graphics* (3rd ed.). Boca Raton, FL: CRC Press.

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

Pearl, J. (2016). Lord's paradox revisited (Oh Lord! Kumbaya!). *Journal of Causal Inference*, *4*(2). doi: 10.1515/jci-2016-0021

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

Peters, J. (2013, July). *When ice cream sales rise, so do homicides. coincidence, or will your next cone murder you?* Slate. Retrieved from `slate.com/news-and-politics/2013/07/warm-weather-homicide-rates-when-ice-cream-sales-rise-homicides-rise-coincidence.html`

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., & Burgette, L. (2017). **twang**: Toolkit for weighting and analysis of nonequivalent groups [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=twang` (R package version 1.5)

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer-Verlag. doi: 10.1007/978-1-4757-3692-2

Rubin, D. B. (2006). *Matched sampling for causal effects.* Cambridge, MA: Cambridge University Press. doi: 10.1017/CBO9780511810725

Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science*, *12*, 487–508.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The **Matching** package for R. *Journal of Statistical Software*, *42*(7), 1–52. Retrieved from `http://www.jstatsoft.org/v42/i07/`

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702–712.

Steiner, P., Cook, T., & Shadish, W. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, *36*, 213–236. doi: 10.3102/1076998610375835

Venables, W. N., & Ripley, B. D. (2000). *S programming.* New York, NY: Springer.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer. Retrieved from `http://www.stats.ox.ac.uk/pub/MASS4` (ISBN 0-387-95457-0)

Vigen, T. (2015). *Spurious correlations.* New York, NY: Hatchette Books.

Wainer, H. (1991). Adjusting for differential base rates: Lord's paradox again. *Psychological Bulletin*, *109*, 147–151.

Wickham, H. (2015). *Advanced R.* Boca Raton, FL: CRC Press.

Wilkinson, L., & the Task Force on Statistical Inference, APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604. doi: 10.1037/0003-066X.54.8.594

Wright, D. B. (2003). Making friends with your data: Improving how statistics are conducted and reported. *British Journal of Educational Psychology*, *73*, 123–136. doi:

10.1348/000709903762869950

Wright, D. B. (2006). Comparing groups in a before-after design: When t test and ANCOVA produce different results. *British Journal of Educational Psychology*, *76*, 663–675. doi: 10.1348/000709905X52210

Wright, D. B. (2017). Using graphical models to examine value-added models. *Statistics and Public Policy*, *4*, 1–7. doi: 10.1080/2330443X.2017.1294037

Wright, D. B. (2018). Estimating school effectiveness with student growth percentile and gain score models. *Journal of Applied Statistics*, *45*, 2536–2547. doi: 10.1080/02664763.2018.1426742

Wright, D. B. (2020). Allocation to groups: Examples of Lord's paradox. *British Journal of Educational Psychology*.

Xie, Y. (2013). *Dynamic documents with R and **knitr***. Boca Raton, FL: Chapman and Hall/CRC.

Appendix

R Code for the simulations

This paper was written using LaTeX with **knitr** Xie (2013). This entire document in both `Rnw` and `pdf` formats is at `https://github.com/dbrookswr/GainAncPM`. A LaTeX editor will need to be set-up (`https://yihui.org/knitr/demo/editors/`) and several LaTeXand R packages will need to be installed to compile the `Rnw` file.

The code to create the data and run the simulation used in Example #1a is:

```
qrank <- function(x) 1-(rank(x)+1)/(length(x)+2)

n <- 200
reps <- 10000 #47
set.seed <- 9384
#3*4 is a/b/c by 4 procedures
simpvals <- matrix(ncol=6,nrow=reps*12)
for (i in 1:reps){
  know <- rnorm(n)
  prior <- scale(know + rnorm(n))
  propena <- scale(know + rnorm(n))
  propenb <- scale(prior + rnorm(n))
  propenc <- scale(know + prior + rnorm(n))
  treata <- rbinom(n,1,qrank(propena))
  treatb <- rbinom(n,1,qrank(propenb))
  treatc <- rbinom(n,1,qrank(propenc))
  #posts differ just by rnorm, but two kept in case
  #treatment effect added by someone
  posta <- scale(0*treata + know + rnorm(n))
  postb <- scale(0*treatb + know + rnorm(n))
  postc <- scale(0*treatc + know + rnorm(n))
  simpvals[((i-1)*12+1):(i*12),] <-
    simpallmodels(prior,treata,treatb,treatc,posta,postb,postc)
    }

colnames(simpvals) <-
    c("DAG","Procedure","ATE","seATE","zval","pval")
simpvals <- as.data.frame(simpvals)
simpvals$DAG <-
  factor(recoder(simpvals$DAG,'0:"A";1:"B";2:"C" '),
         levels=c("A","B","C"))
simpvals$Procedure <-
  factor(recoder(simpvals$Procedure,
          '1:"Gain";2:"ANCOVA";3:"PM";4:"t-test" '),
         levels=c("Gain","ANCOVA","PM", "t-test"))
```

The function `simpallmodels` runs all the statistical models and returns a six-column matrix of: the procedure (gain score, ANCOVA, propensity matching), which covariates are included, the effect estimate, its standard error, the t (from `lm`) or z (from `glm` and

`Match`), and the associated two-tailed p-value. It calls three functions, `gain`, `anc`, and `pm`, which run the models for the gain score, ANCOVA, and propensity matching procedures, respectively. The package **Matching** (Sekhon, 2011) needs to be installed and loaded since it is called by the `pm` function. This function and the functions it calls are here:

```
# 0 gain, 1 ancova, 2 pm
simpallmodels <- function(prior,treata,treatb,treatc,
                          posta,postb,postc){
  modvals <- matrix(nrow=12,ncol=6)
  modvals[1,] <- c(0,1,summary(lm(posta - prior ~ treata))$coef[2,])
  modvals[2,] <- c(1,1,summary(lm(postb - prior ~ treatb))$coef[2,])
  modvals[3,] <- c(2,1,summary(lm(postc - prior ~ treatc))$coef[2,])
  modvals[10,] <- c(0,4,summary(lm(posta ~ treata))$coef[2,])
  modvals[11,] <- c(1,4,summary(lm(postb ~ treatb))$coef[2,])
  modvals[12,] <- c(2,4,summary(lm(postc ~ treatc))$coef[2,])
  modvals[4,] <- c(0,2,summary(lm(posta ~ treata + prior))$coef[2,])
  modvals[5,] <- c(1,2,summary(lm(postb ~ treatb + prior))$coef[2,])
  modvals[6,] <- c(2,2,summary(lm(postc ~ treatc + prior))$coef[2,])
  propval <- glm(treata ~ prior,family="binomial")$fitted
  mod <- Match(posta,treata,propval,estimand="ATE",BiasAdjust=TRUE,ties=TRUE)
  modvals[7,] <- c(0,3,mod$est,mod$se,
           zval<-mod$est/mod$se,(1 - pnorm(abs(zval)))*2)
  propval <- glm(treatb ~ prior,family="binomial")$fitted
  mod <- Match(postb,treatb,propval,estimand="ATE",BiasAdjust=TRUE,ties=TRUE)
  modvals[8,] <- c(1,3,mod$est,mod$se,
           zval<-mod$est/mod$se,(1 - pnorm(abs(zval)))*2)
    propval <- glm(treatc ~ prior,family="binomial")$fitted
    mod <- Match(postc,treatc,propval,estimand="ATE",BiasAdjust=TRUE,ties=TRUE)
  modvals[9,] <- c(2,3,mod$est,mod$se,
           zval<-mod$est/mod$se,(1 - pnorm(abs(zval)))*2)
      return(modvals)
       }
qrank <- function(x) 1-(rank(x)+1)/(length(x)+2)
```

The code for #1b is very similar for #1a, but is repeated where for documentation purposes.

```
qrank <- function(x) 1-(rank(x)+1)/(length(x)+2)

n <- 200
reps <- 100000 #47
set.seed <- 28214
#3*4 is a/b/c by 4 procedures
simpvals1b <- matrix(ncol=7,nrow=reps*12)
for (i in 1:reps){
  know <- rnorm(n)
  teff <- runif(1,-1,1)
  prior <- scale(know + rnorm(n))
```

```
  propena <- scale(know + rnorm(n))
  propenb <- scale(prior + rnorm(n))
  propenc <- scale(know + prior + rnorm(n))
  treata <- rbinom(n,1,qrank(propena))
  treatb <- rbinom(n,1,qrank(propenb))
  treatc <- rbinom(n,1,qrank(propenc))
  #posts differ just by rnorm, but two kept in case
  #treatment effect added by someone
  posta <- scale(teff*treata + know + rnorm(n))
  postb <- scale(teff*treatb + know + rnorm(n))
  postc <- scale(teff*treatc + know + rnorm(n))
  simpvals1b[((i-1)*12+1):(i*12),] <-
    cbind(simpallmodels1b(prior,treata,treatb,treatc,posta,postb,postc),rep(teff,12))
    }

colnames(simpvals1b) <-
    c("DAG","Procedure","ATE","seATE","zval","pval","teff")
simpvals1b <- as.data.frame(simpvals1b)
simpvals1b$DAG <-
  factor(recoder(simpvals1b$DAG,'0:"A";1:"B";2:"C" '),
         levels=c("A","B","C"))
simpvals1b$Procedure <-
  factor(recoder(simpvals1b$Procedure,
          '1:"Gain";2:"ANCOVA";3:"PM";4:"t-test" '),
         levels=c("Gain","ANCOVA","PM", "t-test"))
```

and

```
# 0 gain, 1 ancova, 2 pm
simpallmodels1b <- function(prior,treata,treatb,treatc,
                            posta,postb,postc){
  modvals <- matrix(nrow=12,ncol=6)
  modvals[1,] <- c(0,1,summary(lm(posta - prior ~ treata))$coef[2,])
  modvals[2,] <- c(1,1,summary(lm(postb - prior ~ treatb))$coef[2,])
  modvals[3,] <- c(2,1,summary(lm(postc - prior ~ treatc))$coef[2,])
  modvals[10,] <- c(0,4,summary(lm(posta ~ treata))$coef[2,])
  modvals[11,] <- c(1,4,summary(lm(postb ~ treatb))$coef[2,])
  modvals[12,] <- c(2,4,summary(lm(postc ~ treatc))$coef[2,])
  modvals[4,] <- c(0,2,summary(lm(posta ~ treata + prior))$coef[2,])
  modvals[5,] <- c(1,2,summary(lm(postb ~ treatb + prior))$coef[2,])
  modvals[6,] <- c(2,2,summary(lm(postc ~ treatc + prior))$coef[2,])
  propval <- glm(treata ~ prior,family="binomial")$fitted
  mod <- Match(posta,treata,propval,estimand="ATE",BiasAdjust=TRUE,ties=TRUE)
  modvals[7,] <- c(0,3,mod$est,mod$se,
           zval<-mod$est/mod$se,(1 - pnorm(abs(zval)))*2)
  propval <- glm(treatb ~ prior,family="binomial")$fitted
  mod <- Match(postb,treatb,propval,estimand="ATE",BiasAdjust=TRUE,ties=TRUE)
  modvals[8,] <- c(1,3,mod$est,mod$se,
           zval<-mod$est/mod$se,(1 - pnorm(abs(zval)))*2)
    propval <- glm(treatc ~ prior,family="binomial")$fitted
    mod <- Match(postc,treatc,propval,estimand="ATE",BiasAdjust=TRUE,ties=TRUE)
```

```
  modvals[9,] <- c(2,3,mod$est,mod$se,
           zval<-mod$est/mod$se,(1 - pnorm(abs(zval)))*2)
      return(modvals)
       }
qrank <- function(x) 1-(rank(x)+1)/(length(x)+2)
```

The code to create the data and run the simulation for Example #2 when excluding just one covariate is:

```
n <- 200
reps <- 10000 #47 #10000 #47
set.seed <- 9811
vals <- matrix(ncol=6,nrow=reps*3*9)
for (i in 1:reps){
  eph <- rnorm(n)
  e1 <- rnorm(n)
  e3 <- scale(eph + 1.0371*rnorm(n)) #
  env <- scale(.3*eph + .2082*e1  + .2*rnorm(n))
  e2 <- scale(sqrt(.25)*env + sqrt(.75)*rnorm(n))
  oph <- rnorm(n)
  o1 <- rnorm(n)
  o3 <- scale(oph + 1.0371*rnorm(n)) #
  other <- scale(.3*oph + .2082*o1 + .2*rnorm(n))
  o2 <- scale(sqrt(.25)*other + sqrt(.75)*rnorm(n))
  xph <- rnorm(n)
  x1 <- rnorm(n)
  x3 <- scale(xph + .71*rnorm(n)) #
  know <- scale(.85*xph + .6935*x1 + .6*other + .6*env)
  prior <- scale(sqrt(.25)*know + sqrt(.75)*rnorm(n))
  propenval <- qrank(env)
  treat <- rbinom(n,1,propenval)
  post <- scale(other + 0*treat + know + rnorm(n))
  vals[((i-1)*27+1):(i*27),] <- estswo1(e1,e2,e3,x1,prior,x3,
                                  o1,o2,o3,treat,post,know)
    }

colnames(vals) <-
    c("statmodel","MissCOV","ATE","seATE","zval","pval")
ex2vals <- as.data.frame(vals)
```

The function used to estimate the models is:

```
# 0 gain, 1 ancova, 2 pm
estswo2b <- function(e1,e2,e3,x1,prior,x3,o1,o2,o3,treat,post,know){
     covs <- cbind(e1,e2,e3,x1,prior,x3,o1,o2,o3)
     covs2 <- covs
     vals3wo2b <- matrix(nrow=3*3,ncol=6)
     sets <- list(1:3,4:6,7:9)
```

```
     for (j in 1:3){
       ifelse (j == 2, vals3wo2b[(j-1)*3+1,] <-
         c(1,j,summary(lm(post - prior ~ treat + covs[,-sets[[j]]]))$coef[2,]),
                         vals3wo2b[(j-1)*3+1,] <-
         c(1,j,summary(lm(post - prior ~ treat + covs[,-c(5,sets[[j]])]))$coef[2,]))
       vals3wo2b[(j-1)*3+2,] <-
         c(2,j,summary(lm(post ~ treat + covs[,-sets[[j]]]))$coef[2,])
       propval <- glm(treat ~ prior + covs[,-sets[[j]]],family="binomial")$fitted
       mod <- Match(post,treat,propval,estimand="ATE")
      vals3wo2b[(j-1)*3+3,] <- c(3,j,mod$est,mod$se,
                         zval<-mod$est/mod$se,(1 - pnorm(abs(zval)))*2)}
      return(vals3wo2b) }
qrank <- function(x) 1-(rank(x)+1)/(length(x)+2)
```

Similar code was used when excluding the sets of three covariates, and this code is available from the author. The code for Example #2b is not included here. It was just changing a single line in the code, as discussed in the main body of the paper.