Careful Selection of Covariates in the Presence of Model Uncertainty for Evaluators Interested in Unbiased Estimation of Causal Effects

Brian Knaeble, Ph.D.; Assistant Professor
Utah Valley University, Orem, Utah, bknaeble@uvu.edu

Libby Smith, M.S.; Evaluation Project Manager, University of Wisconsin-Stout,
Applied Research Center, Menomonie, Wisconsin; smithlib@uwstout.edu

Aric Gregg, M.S.; Program Evaluator, University of Wisconsin-Stout,
Applied Research Center, Menomonie, Wisconsin; greggar@uwstout.edu

Levi Roth, M.S.; Program Evaluator, University of Wisconsin-Stout,
Applied Research Center, Menomonie, Wisconsin; rothle@uwstout.edu

Brenda Krueger, M.S.; Program Evaluator, University of Wisconsin-Stout,
Applied Research Center, Menomonie, Wisconsin; brkrueger@uwstout.edu

Gina Lawton, M.S.; Data Manager, University of Wisconsin-Stout,
Applied Research Center, Menomonie, Wisconsin; lawtong0808@my.uwstout.edu

Phillip Stoeklen, M.S. (ABT); Research Technician, University of Wisconsin-Stout,
Applied Research Center, Menomonie, Wisconsin; stoeklenp@uwstout.edu

Correspondence: Libby Smith, P.O. Box 790, Menomonie, Wisconsin 54751; Telephone: (715) 232-5412; Fax: (715) 232-5406; University of Wisconsin-Stout, Applied Research Center

This workforce product was funded by a grant awarded by the U.S. Department of Labor's Employment and Training Administration. The product was created by the grantee and does not necessarily reflect the official position of the U.S. Department of Labor. The U.S. Department of Labor makes no guarantees, warranties, or assurances of any kind, express or implied, with respect to such information, including any information on linked sites and including, but not limited to, accuracy of the information or its completeness, timeliness, usefulness, adequacy, continued availability, or ownership. This is an equal opportunity program. Assistive technologies are available upon request and include Voice/TTY (771 or 800-947-6644).

Abstract

The ongoing INTERFACE Project uses funds from the third round of the Trade Adjustment Assistance Community College and Career Training (TAACCCT) grant program to develop, improve, and expand educational training within the Wisconsin Technical College System (WTCS). An evaluation team from the Applied Research Center (ARC) at the University of Wisconsin-Stout was tasked with collecting and analyzing data to evaluate the effectiveness of the INTERFACE Project, regarding student outcomes relating to graduation and employment. This paper describes the statistical model that will be used during this evaluation, with an emphasis on theoretical details of general interest to evaluators.

Keywords: propensity score, potential outcomes, ignorable treatment assignment

Careful Selection of Covariates in the Presence of Model Uncertainty for Evaluators Interested in Unbiased Estimation of Causal Effects

Statisticians make assumptions, such as the assumption that a study sample was randomly selected (Berk & Freedman, 2001). If a sample is not random, then assumptions of normality may not be justified (Friedman, 1937) and mistaken assumptions of independence can be problematic (Kruskal, 1988; Kelley, 1999). Also, statistical inference can be compromised when these assumptions are not met (Chatfield, 1995). Diagnostics, or validation techniques (NIST, 2016), can be useful when checking assumptions. Statistical simulation can be used to assess the sensitivity of outcomes or conclusions to departures from assumptions (Burton, Altman, Royston, & Holder, 2006). For more reading on sensitivity analysis see Rosenbaum (2005).

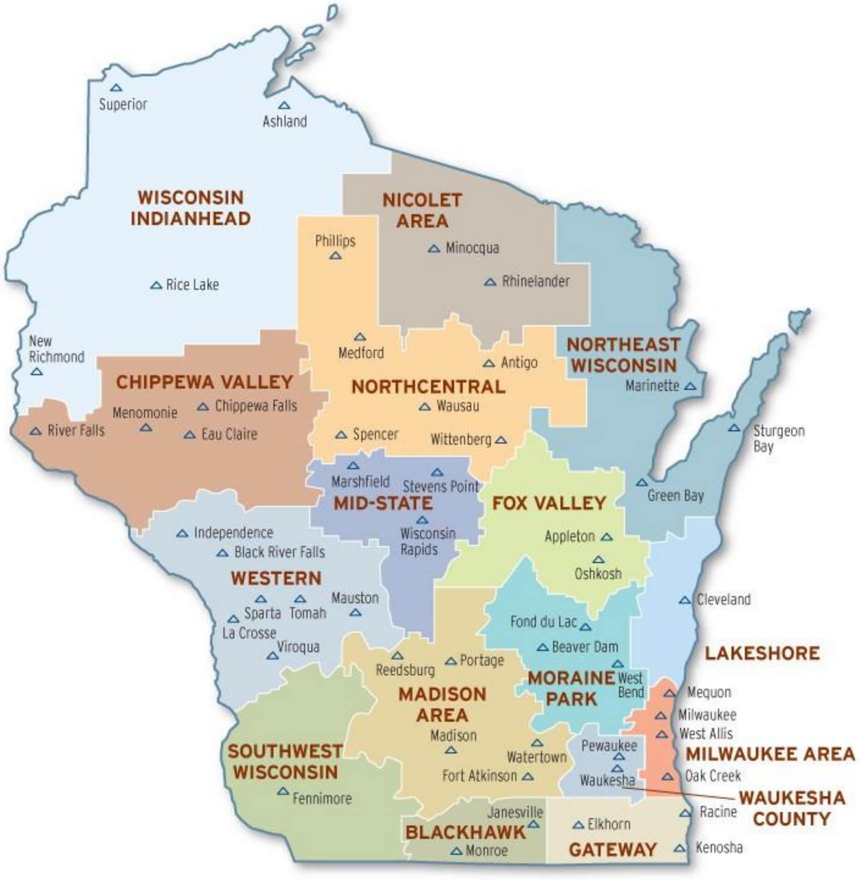
According to Guo & Fraser (2015), strongly ignorable treatment assignment (SITA) is an important assumption of propensity score analysis. This means that conditional on a set of covariates the potential outcomes are independent of treatment assignment (Rosenbaum & Rubin, 1983). This assumption is sometimes referred to as non-confounding (Austin, 2011). It is often difficult to verify SITA (Steiner, Cook, Shadish, & Clark, 2010), and sometimes validation of SITA is neglected altogether (Richars, Smith, Jennings, Bjerregaard, & Fogel, 2014; Choi, Burgard, Elo, & Heisler, 2015). When SITA fails, matching on observed covariates may balance observed covariates, but estimates for causal effects can still be biased due to the presence of unmeasured covariates (Harder, Stuart, & Anthony, 2010; Kretchmann, Vock, & Lüdtke, 2014; Rosenbaum, 2010). Lane, To, Shelley, and Henson (2012) provide an example where propensity scores were used in educational research, and they discussed SITA and sensitivity analysis. Evaluators can find similar discussions of interest in Tipton (2013), McIntire, Nelson, Macy, Seo, & Kolbe (2015), and Guo and Fraser (2015).

The United States Department of Labor (USDOL) has encouraged evaluators to utilize the methodology of propensity score matching (PSM) (Urban Institute, 2013). Rosenbaum and Rubin's (1983) research on the central role of the propensity score for causal effects has been cited over 15,000 times. Yet, Pearl refers to the opacity of SITA as an Achilles' heel, stating "No mortal can apply this condition to judge whether it holds even in simple problems" (Pearl, 2009a, p. 350). Pearl has shown how directed acyclic graphs (DAGs) can be used to select a set of covariates for adjustment (Pearl, 2009b, Section 3), but Rubin (2009) has called such an approach non-scientific. Wasserman (2010) covers both approaches to causal analysis in his book. Herein, the advice from both Pearl and Rubin have been incorporated into a single model. This approach is described within this article using an ongoing evaluation.

Context

In 2009, the American Recovery and Reinvestment Act amended the Trade Act of 1974 to authorize the Trade Adjustment Community College and Career Training (TAACCCT) Grant Program. The TAACCCT grant provided community colleges and other eligible institutions of higher education with funds to expand and improve their ability to deliver education and career training programs (TAACCCT, 2011). In 2013, during the third round of TAACCCT, the Wisconsin Technical College System (WTCS) received $23.1 million in funds meant to support the development and improvement of its Intentional Networks Transforming Effective and Rigorous Facilitation of Assessment, Collaboration, and Education (INTERFACE) Project. Within WTCS, 16 colleges (see Figure 1) participated in the INTERFACE Project, which seeks to develop, improve, and expand adult educational training pathways to information technology-related careers in business, information technology, healthcare, logistics, automation, and manufacturing (INTERFACE Project, 2015).

Figure 1. Map of Wisconsin Technical College System.



Both the consortium of WTCS colleges and the USDOL are interested in the impacts and outcomes of the INTERFACE Project (Advance Wisconsin, 2015). A third-party evaluation team from the Applied Research Center (ARC) at the University of Wisconsin-Stout (UW-Stout) was brought on to assess the project. Over the four-year grant period, this evaluation team collected qualitative data through interviews with project stakeholders and students, collected and analyzed quantitative data being tracked by the colleges, and developed reports to provide formative feedback to project stakeholders. This feedback recommended needed improvements during the implementation phase and allowed the consortium of WTCS colleges and the USDOL to understand the impacts and outcomes of the INTERFACE Project (UW-Stout Evaluation Team, 2015). Student outcomes of interest for the project relate to graduation and employment.

These outcome variables relate to program completion, retention, credits earned, further education, employment, employment retention, and wage.

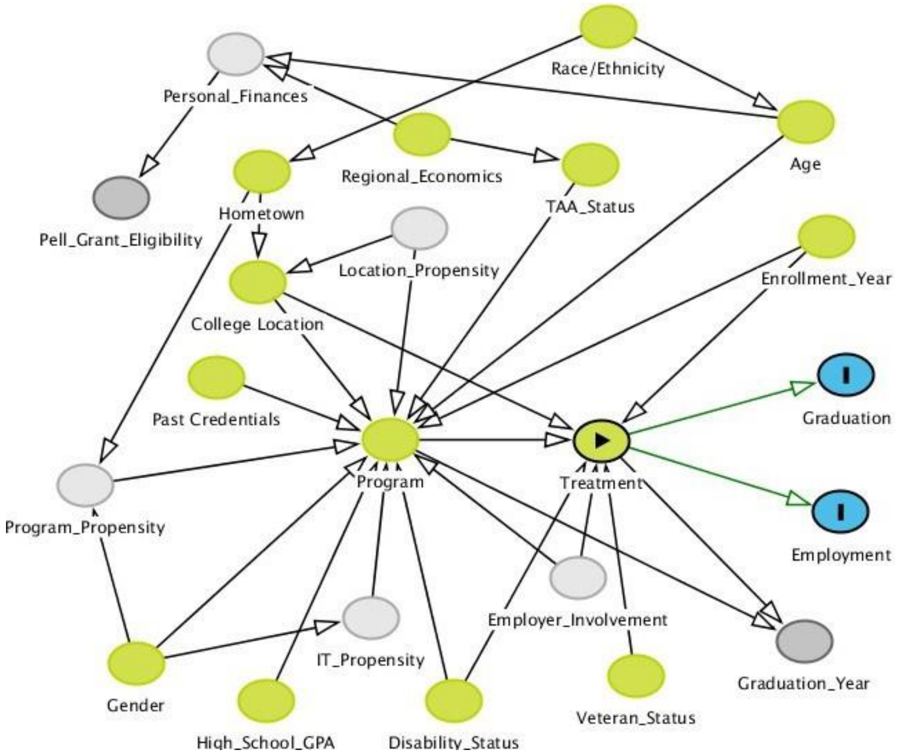
The INTERFACE Project is considered an intervention (a set of treatments) on the population of all students enrolled within WTCS between January 2014 and March 2017. The population is expected to be about 350,000 students. About 4,000 of these students will belong to the treatment group, meaning that they have been impacted by the INTERFACE Project. The causal effect of treatment was defined as the difference between that student's eventual outcome (data that will be obtained) and the outcome that student would have obtained had the intervention not occurred. Since the hypothetical outcome is counterfactual (Wasserman, 2010, Chapter 16), students will be matched based on covariate data where similar untreated students and outcomes will be compared. This difference in outcomes will be averaged over the population of all treated students and the result described as the treatment effect on the treated (Morgan & Winship, 2007). This analysis will be done separately for each outcome (Johnson & Wichern, 2007).

Covariates that the students will be matched on include ethnicity, gender, disability status, veteran status, hometown location, hometown regional economics, college for study, age, program of study, high school grade point average, past credentials, employer involvement in study, TAA status, veteran status, disability status, enrollment year, location propensity, program propensity, IT propensity, Pell Grant eligibility, and graduation year. Note that location propensity, program propensity, and IT propensity will be used within the model of treatment propensity. For example, an individual's treatment propensity is conditional based on their IT propensity. More specifically, it is anticipated that treated students will have a higher IT propensity than non-treated students. Binary educational outcomes include program completion, program retention (based on credits attempted), and further education. Pass rate is an educational outcome defined as the number of credits earned divided by the number of credits attempted. Binary employment variables include employment upon graduation (yes or no) and employment retention for six months. Wage is an employment variable defined as post-graduation wage minus pre-enrollment wage. See Appendix A for additional details regarding these variables.

Modeling

The variables of interest are shown in Figure 2 as nodes of a DAG, which is also known as a Bayesian Network (Pearl, 2009a). The structure of this graph was based on information obtained from qualitative data collected during site visits, careful reading of the USDOL's Solicitation for Grants Application (United States Department of Labor, 2013), and mathematical simulations (see Appendix B). Within the graph, the indicator variable for treatment is represented with a triangular node. For simplicity, outcome variables are classified into two nodes, one for graduation and one for employment. The remaining nodes represent covariates. Classified as ancestors, the yellow nodes represent variables affecting treatment. Latent variables are represented by light gray nodes; these are unobserved. Variables that are observable, but do not affect treatment, are represented by dark gray nodes. All covariates are thought to affect all outcomes, except for ethnicity, gender, disability status, and veteran status, which are assumed to affect employment variables but not graduation variables. For graduation outcomes, these four covariates are thus considered instrumental variables, because they affect treatment but not outcome (Pearl, 2009b). Arrows are not drawn from covariates to outcomes for simplicity.

Figure 2: Bayesian network for the INTERFACE Project.



Comparisons between the average outcomes for the treated to the associated average outcomes for the untreated is not a sound method (Wasserman, 2010, Theorem 16.1). Such an approach is sometimes described as naive because it fails to control for covariates (Morgan & Winship, 2007). Failure to control for a covariate that affects both treatment and outcome can lead to considerable bias, but inappropriate control for covariates that are affected by treatment can also lead to considerable bias (Pearl, 2014). It is possible to utilize data associated with post-treatment variables as part of a multi-step procedure to estimate a causal effect (Pearl, 2009b), yet it has long been recognized that statisticians should not condition on post-treatment variables (Cox, 1958). Gelman (Gelman et al., 2004) and Rubin and Rosenbaum (as cited by Gelman, 2009) recommend adjusting for as many pre-treatment covariates as possible, but Pearl (2011), Woolridge (2009), and Myers et al. (2011) have pointed out that bias amplification is possible when instrumental variables are used within a propensity score analysis. An admissible set

(Pearl, 2009b, p. 113) of covariates can be selected from our Bayesian network using the back-door criterion (Pearl, 2009b). However, there could be hidden relationships with additional covariates not present in the network (Armistead, 2014; Rosenbaum, 2010). Nevertheless, the network will be used to guide a conservative approach based on all the preceding considerations. The evaluation team will ignore instrumental and post-treatment variables and utilize as many of the remaining covariates as possible within a propensity score matching procedure. The covariate in Figure 3 is an instrumental variable. The covariate in Figure 4 is a post-treatment variable.

Figure 3: An instrumental covariate affecting treatment but not outcomes.

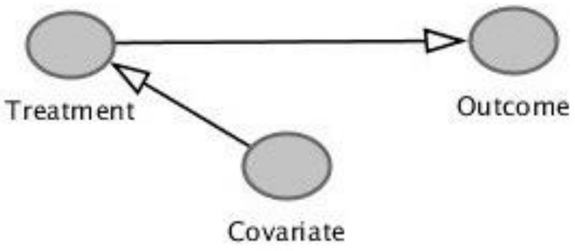
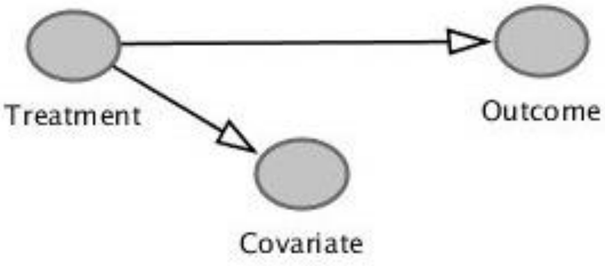


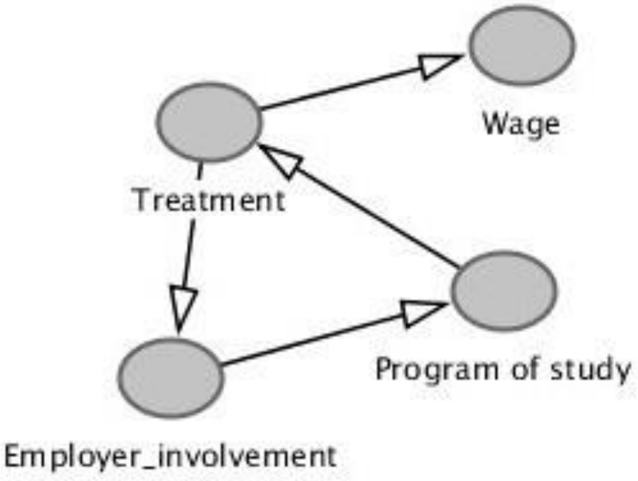
Figure 4: A post-treatment covariate affected by treatment.



After excluding instrumental and post-treatment variables we argue that with more covariates the assumption of SITA is more likely to be satisfied. SITA states that the potential outcomes are independent of treatment assignment conditional on the set of covariates (Wasserman, 2010), leading to the theoretical conclusion that matching produces unbiased estimates for causal effects (Rosenbaum & Rubin, 1983). This also operates under the stable unit treatment value assumption (SUTVA). SUTVA states that the outcome for any given individual is independent of the treatment status of other individuals. SUTVA is not perfectly satisfied in our situation because treatment may improve the chances of employment and graduates may be competing for a finite set of jobs. Mathematical simulations of this approach (see Appendix B) have convinced us to go ahead with our analysis as planned, assuming SITA and SUTVA, but for retrospective sensitivity analysis and reliability analysis we plan to fit related models on subpopulations identified from responses to surveys. When asked about their treatment assignment some respondents may specify that their enrollment was essentially random, and on this subpopulation a less in-depth analysis will be conducted and compared with the overall propensity score analysis.

Our Bayesian network may not perfectly reflect reality for all students. For example, the network may indicate that employer involvement affects treatment. However, for some students, especially those who anticipated benefits from treatment and made proactive career decisions, it may be the case that treatment affects employer involvement, resulting in the model shown in Figure 5, where the graph is no longer acyclic.

Figure 5: For some students treatment may affect employer involvement.



The evaluation team estimates that 83% of treated students were “blind” to treatment, meaning that these students were unaware that their program was impacted by grant funds. The remaining 17% of students were aware of the grant’s influence on their program of study; therefore, these students will be referred to as un-blind. Separate analyses may be done for blinded and un-blinded populations of students. When asked about their competition for jobs, some respondents may indicate that there was little to no competition. If so, propensity score estimates will be produced on this subpopulation for comparison with the overall estimates. Also, a dose can be assigned to each treatment, and the dose-outcome relationships will be compared with overall estimates for treatment effects.

Discussion

Prospective mathematical simulations helped guide our model construction process. See Appendix B for a sample of the R code that was used along with a sample graph showing agreement between simulation and theory. As an additional precaution, we may test the reliability of our conclusions by comparing the overall results to results obtained on subpopulations of interest. To assess the sensitivity of the conclusions to the particular modeling procedure that avoids conditioning on instrumental variables, avoids conditioning on post-treatment variables, actively conditions on as many pre-treatment variables as possible, and matches on propensity, the procedure may also be modified to see how conclusions are affected. Some modification of the entire analytic framework may be necessary.

There has been some criticism of propensity score methods. Under SITA propensity score matching may produce an unbiased estimate, but unbiasedness is not the only desirable quality of an estimator. To appreciate the historical context of this claim, see Salzburg (2001). In addition to unbiasedness, it is desirable for estimators to be consistent (Wasserman, 2010), efficient (Everitt, 2002, p. 128), and robust (Stigler, 2010). There are additional qualities as well

(Salzburg, 2001, p. 66). Imai, King, and Nall (2009) provide some reasons for preferring fully blocked experiments over completely randomized experiments. King and Nielson (2016) explain how general matching approximates a fully blocked experimental design while PSM approximates a fully randomized experimental design, arguing that this is a weakness of PSM. These researchers argue that PSM can lead to worse imbalance (King and Nielson, 2016, Sections 4 and 5).

The Counterfactual Model (Wasserman, 2004) typically considers two potential outcomes, one for treatment and one for control, with one realized and the other hypothetical for any given individual. The average causal effect can be defined as the average over some population of the difference between the two potential outcomes. For example, for headache relief, an acetaminophen (pain reliever) may be taken (treatment) or not (control), and the average difference in outcome over a whole population represents the causal effect of acetaminophen on headaches, assuming individuals within that population behave identically (excepting acetaminophen usage) under treatment and control. With acetaminophen, this is plausible, but with INTERFACE it may not be. An individual, who was treated with acetaminophen can hypothetically imagine doing everything the same but only without acetaminophen or perhaps with a placebo. An individual who was treated with INTERFACE funds does not have a counterfactual control scenario. For some programs of study, it is impossible to separate treatment from the program itself (i.e. it is impossible to go through the program without being treated, because the programs existence is tied up with the INTERFACE Project). In such cases, it is not clear what counterfactual behavior would occur had INTERFACE not intervened on WTCS. At a minimum, these concerns should be addressed through focus on subpopulations where counterfactual behavior is better defined. A more general framework could be used.

The use of subpopulations to check for sensitivity of results and conclusions to departures from model assumptions has been discussed. The data may also be analyzed to identify subpopulations where treatment affect is higher than average. For example, it may be that treatment is especially effective at preparing female veterans for employment, but less effective at increasing the wage of male incumbent workers generally. Because the overall estimates are for the treatment effect on the treated, these estimates potentially describe the benefits gained due to INTERFACE. There is not a plan to estimate treatment effects on the untreated (i.e. to predict what would happen if INTERFACE were expanded). This is largely because there are approximately 100 untreated students for every treated student. Thus, finding a match for each treated student is easier than the other way around. Likewise, a plan does not exist to estimate the causal treatment effects on the whole population. The focus has exclusively been on describing the benefits accrued to students due to the INTERFACE Project as actually implemented. The quantitative methodology described here is complementary to qualitative assessment and evaluation of the INTERFACE Project.

References

Advance Wisconsin - Information Technology (2015). INTERFACE Project. Retrieved from http://advancewisconsin.org/ advance-wisconsin/it/

Advance Wisconsin. (2015). Site Visits with the UW-Stout Evaluation Team. Retrieved from http://advancewisconsin.org/2015/10/02/ site-visits-with-the-uw-stout-evaluation-team/

Armistead, T. (2014). Resurrecting the third variable: A critique of Pearl's causal analysis of Simpson's paradox. *The American Statistician*, *68*(1), 1-7.

Austin, P. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*, 399-424.

Berk, R., & Freedman, D. A. (2001). Statistical assumptions as empirical commitments. *Department of Statistics, UCLA*. UCLA: Department of Statistics, UCLA. Retrieved from: http://escholarship.org/uc/item/0zj8s368

Burton, A., Altman, D., Royston, P., & Holder, R. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, *25*, 4279-4292.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society*, *158*(3), 419-466.

Choi, H., Burgard, S., Elo, I., & Heisler, M. (2015). Are older adults living in more equal counties healthier than older adults living in more unequal counties? A propensity score matching approach. *Social Science and Medicine*, *141*, 82-90.

Cox, D. (1958). *Planning of experiments*. New York, NY: John Wiley & Sons, Inc.

Everitt, B. (2002). *The Cambridge dictionary of statistics*. West Nyack, NY: University Press.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*(200), 675-701.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis* (2nd ed.) Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A. (2009, July 5). Resolving disputes between J. Pearl and D. Rubin on causal inference. Retrieved from http://andrewgelman.com.

Guo, S., & Fraser, M. (2015). *Propensity Score Analysis*. Thousand Oaks, CA: SAGE Publications, Inc.

Harder, V., Stuart, E., & Anthony, J. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, *15*(3), 234-249. doi:10.1037/a0019623

Imai, K., King, G., & Nall, C. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Statistical Science*, *24*(1), 29-53.

Johnson, R., & Wichern, D. (2007). *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ: Pearson Prentice Hall.

Kelley, K. (1999). 95 Million t tests: The empirical findings when the assumption of independence has been violated in the two-sample t test. University of Cincinnati dissertation.

King, G., & Nielson, R. (2016). Why propensity scores should not be used for matching. Retrieved from http://j.mp/PScore.

Kretschmann, J., Vock, M., & Lu¨dtke, O. (2014). Acceleration in elementary school: Using propensity score matching to estimate the effects on academic achievement. *Journal of Educational Psychology*, *106*, 1080-1095.

Kruskal, W. (1988). Miracles and statistics: The casual assumption of independence. *Journal of the American Statistical Association*, *83*(404), 929-940.

Lane, F., To, Y. M., Shelley, K., & Henson, R. (2012). An illustrative example of propensity score matching with education research. *Career and Technical Education Research, 37*(3), 187-212. doi:10.5328/cter37.3.187

McIntire, R., Nelson, A., Macy, J., Seo, D., & Kolbe, L. (2015). Secondhand smoke exposure and other correlates of susceptibility to smoking: A propensity score matching approach. *Addictive Behaviors, 48*, 36-43.

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

Myers, J., Rassen, J., Gagne, J., Huybrechts, K., Schneeweiss, S., Rothman, K., … Glynn, R. (2012). Effects of adjusting for instrumental variables and bias and precision of effect estimates. *Am J Epidemiol*, *174*(11), 1213-1222.

NIST. (2016). *How can I tell if a model fits my data? Engineering Statistics Handbook.* Retrieved from http://www.itl.nist.gov/div898/handbook/ pmd/section4/pmd44.htm

Pearl, J. (2009a). *Causality: models, reasoning, and inference*. New York, NY: Cambridge University Press.

Pearl, J. (2009b). Causal inference in statistics: An overview. *Statistical Surveys*, *3*, 96-146.

Pearl, J. (2011). Invited commentary: Understanding bias amplification. *American Journal of Epidemiology*, *174*(11), 1223-1227.

Pearl, J. (2014). Comment: Understanding simpson’s paradox. *The American Statistician*, *68*(1), 8-13.

Richars, T., Smith, M., Jennings, W., Bjerregaard, B., & Fogel, S. (2014). An examination of defendant sex disparity in capital sentencing: A propensity score matching approach. *American Journal of Criminal Justice*, *39*, 681-697.

Rosenbaum, P.R. (2005). Sensitivity analysis in observational studies. *Encyclopedia of Statistics in Behavioral Science*, *4*, 1809-1814.

Rosenbaum, P.R. (2010). *Design of Observational Studies*. Springer: New York, NY: Springer.

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1)*,* 41-55.

Rubin, D.B. (2009). Authors reply: Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine*, *28*, 1420-1423.

Salzburg, D. (2001). The lady tasting tea.: How statistics revolutionized science in the twentieth century. New York, NY: WH Freeman and Company, USA.Steiner, P. M., Cook, T. D., Shadish, W. R., and Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15*(3), 250-267. doi:10.1037/a0018719

Stigler, S. (2010). The changing history of robustness. *The American Statistician*, *64*(4), 277-281.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassifications: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, *38*, 239-266.

United States Department of Labor. (2013). Notice of availability of funds and solicitation for grant applications for trade adjustment assistance community college and career training grants. Retrieved from http://webapps.dol.gov/FederalRegister/HtmlDisplay.aspx?DocId=26789&AgencyId=15&DocumentType=3

United States Department of Labor. (2011). TAACCCT, Trade Adjustment Assistance Community College and Career Training Grant Program. Retrieved from http://doleta.gov/taaccct/

Urban Institute. (2013). *National TAACCCT Evaluation: Comparison Groups [PowerPoint slides]*. Retrieved from https://www.taacccteval.org/wp-content/uploads/2015/04/Comparison-Groups.pdf

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference*. New York, NY: Springer.

Wisc-Online. (2015). INTERFACE Project. Retrieved from https://www.wisc-online.com/interface

Wooldridge, J. (2016). Should instrumental variables be used as matching variables? *Research in Economics*, *70*, 232-237.

Appendix A

Detailed Description of Variables and Techniques

Treatment is a dichotomous variable except during specific subpopulation analysis when treatment dose will be considered as an ordered polytomous variable or a continuous variable. It should be noted that Gender was treated as a dichotomous variable due to the data available from WTCS. For each region, the variable Regional Economic Strength is a weighted average of county-level median income over all counties within that region, with each county's weight equal to the proportion of the regional population living in that county. Each covariate is listed in the table below.

Table A1: List of Covariates.

Continuous Variables	Polytomous Variables	Dichotomous Variables
Regional Economic Strength	Ethnicity (polytomous categorical)	TAA (Trade Adjustment Assistance) Status
Age (in years)	Credentials (ordered polytomous)	Gender
Enrollment Year	Program of Study	Veteran Status
Enrollment Time	Hometown (16 regions shown in Figure 1)	Disability Status
High School GPA	College (16 regions shown in Figure 1)	

The participants will be matched on propensity scores, with the propensity scores estimated from a model of treatment assignment in terms of the covariates just described (excepting ethnicity, gender, veteran status, and disability status for graduation related outcomes). If necessary, these or other variables may be utilized in several ways. For example, a hidden variable may be used to adjust wage data before propensity score analysis if the variable is responsible for large wage value discrepancies. Logistic functions may be utilized for continuous variables, perhaps with interaction, as part of the model of propensity. To ensure sufficient counts within categories for the purposes of modeling propensity, categories may be combined or variables eliminated. This will be done in an objective manner. A separate model of propensity will be fit for each of the seven different outcomes under study. Stratification or multiple regression may be utilized in place of matching when appropriate, especially in situations where simulations indicate the bias can be reduced (see Appendix C). General matching may be used in place of PSM (see Section 5.2). A sensitivity analysis will be conducted as described in Rosenbaum (2010, Section 3.4).

Program completion is a dichotomous outcome variable recording whether the student completed their program or not (their first program of study). Completion may mean being awarded a certification, diploma, or associate degree, depending on the program. Program retention is a dichotomous outcome variable measuring whether a student remained a full-time student throughout their first program of study (allowing for not more than a one semester break). Retention considers only credits attempted. Pass rate is a continuous educational outcome variables measuring the proportion of credits earned divided by credits attempted. Further education is a dichotomous outcome variable indicating whether a student went on to further study after their first program of study (affirmative only if further education begins within one semester of graduation). Employment is a dichotomous outcome variable recording whether recent graduates obtained full-time employment (within 6 months of graduation). Subpopulation analysis excluding employment in fields unrelated to study may be performed. Employment retention is a dichotomous outcome variable recording whether an employed individual (employed in the sense of the previous employment outcome) retains full-time employment for at least six months. Wage is the final outcome variable defined as the difference between post-graduation income for one business quarter and pre-enrollment income for one business quarter. Only individuals with full-time employment (before and after) are eligible for wage analysis. It is acceptable for the career or employment specialty prior to treatment to differ from the career or employment specialty post treatment.

Appendix B

R Programs for Simulations

The following function plots density curves.

```
plot.multi.dens <- function(s) { junk.x = NULL

junk.y  = NULL

for(i in 1:length(s)) {

junk.x   =   c(junk.x, density(s[[i]])$x)

junk.y = c(junk.y,  density(s[[i]])$y)  }  xr <-

range(junk.x)

yr <- range(junk.y)


plot(density(s[[1]]), xlim = xr, ylim  =  yr,  main  =

"Collider",xlab="Bias")

for(i in 1:length(s)) {

lines(density(s[[i]]), xlim =  xr, ylim =  yr, col =  i) }   }
```

The following program plots bias for matching, stratification, and regression.

```
library(nonrandom)  k=100    ### must be even

vb=numeric(k)

vr=numeric(k)

vs=numeric(k)

vm=numeric(k)

for (i in 1:k)

{
```

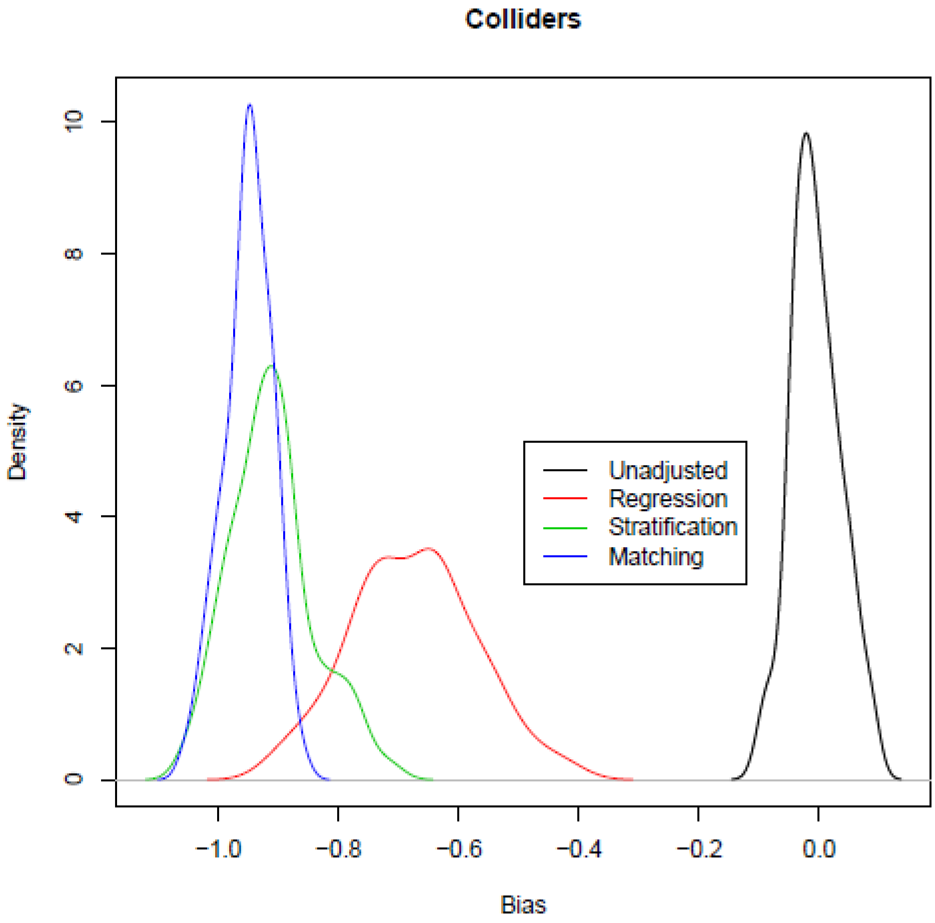
```
###begin.collider

t=c(rep(1,k/2),rep(0,k/2)) y1=rnorm(k,1,.3)

y0=rnorm(k,0,.3) y=c(y1[1:(k/2)],y0[(k/2+1):k])

w=rnorm(k,t,.3)+(y+rnorm(k,0,.3))

###end.collider

vb[i]=summary(lm(y~t))$coefficients[2,1]-mean(y1-y0)

vr[i]=summary(lm(y~t+w))$coefficients[2,1]-mean(y1-y0) M=data.frame(w,t,y)

Ns=ps.makestrata(M,stratified.by="w",breaks=5,name.stratum.index="stratum")

Ps=ps.estimate(Ns$data,treat="t",resp="y",stratum.index="stratum")

vs[i]=as.numeric(Ps$ps.estimation$unadj[2])-mean(y1-y0)

Nm=ps.match(M,matched.by="w",treat="t",name.match.index="match")

Pm=ps.estimate(Nm$data,treat="t",resp="y",match.index="match")

vm[i]=as.numeric(Pm$ps.estimation$unadj[2])-mean(y1-y0)

}

plot.multi.dens(list(vb,vr,vs,vm)) library(Hmisc)

le <- largest.empty(vb,vr,.1,.1)

legend(le,legend=c("Unadjusted","Regression","Stratification","Matching"), col=(1:4),  lwd=2,

lty = 1)
```

Sample output from this program is shown below. The particular program shown above simulates a situation where a single covariate is affected by both treatment and outcome. Such a covariate is called a collider. Modifications to the program between begin.collider and end.collider produce a wide variety of simulations reflecting different data generating processes. The program can be modified to assess susceptibility to bias from:

- multiple covariates
- mistakenly assuming SITA
- mistakenly assuming SUTVA
- misspecification of the propensity function form
- ignoring economic cycles
- using propensity rather than all covariates
- misspecification of the causal graph

Figure 6 shows the bias that can result when conditioning on a collider.

Figure 3: Bias resulting from inappropriate adjustment for a collider.


Colliders
Density
Unadjusted
Regression
Stratification
Matching
Bias