© Journal of Higher Education Outreach and Engagement, Volume 27, Number 2, p. 127, (2023)
Copyright © 2023 by the University of Georgia. eISSN 2164-8212

Why Researchers Should Consider Using Propensity Score Matching Methods to Examine Effectiveness of Community Engagement Programming

Geoffrey Maruyama, Isabel Lopez, Anthony Schulzetenberg, and Wei Song

Abstract

This article provides community engagement researchers with an introduction to propensity score matching (PSM) methods. It explains why PSM can serve as a valuable method for evaluating the success of programs when random assignment of individuals to community engagement programs is not possible; it also addresses some of the advantages and challenges in using PSM. It then explains the steps in conducting a PSM study and illustrates them with an example drawn from research our team conducted. That research looked at the success of a community engagement program in which underrepresented college students mentored and tutored middle school students in their community.

Keywords: propensity score matching, community engagement, methods, quantitative

Experimental methods in which participants are randomly assigned to groups provide a highly attractive approach for investigating the impact of educational programs. These methods meet conditions for identifying causes and effects (Maruyama & Ryan, 2014). Such methods are often referred to as *randomized control trials*, or RCTs. When samples are drawn randomly from a larger population, findings from the sample of participants can be generalized to the larger population. Because they are able to establish causality, RCTs are generally considered the "gold standard" for research (e.g., West, 2009).

Unfortunately, when assessing the effectiveness of postsecondary programs, random assignment is often infeasible because students at most colleges and universities self-select their courses, programs, and activities. Further, in educational studies, there may be strong considerations against random assignment that are both ethical (e.g., withholding treatment from one group of subjects who need it) and practical (e.g., treatment noncompliance; e.g., Lanza et al., 2013). Nevertheless, the capacity of researchers to evaluate the effectiveness and impact of programs is critical for continuing and expanding such programming on college campuses. Fortunately, the absence of random assignment does not necessarily preclude one from drawing inferences. Holland (1986), for example, stated that although experimentation is "the simplest such setting" where causal inference can be discussed, it is not the only "proper setting" (p. 946). As we will explain below, however, establishing causality in the absence of RCTs is difficult and not as definitive.

For college students, free choice is more likely when looking at educational activities that engage them in programs working in communities with community partners. Students who are given the option to voluntarily choose whether or not to participate in a community engagement program or a particular service-learning course may differ in a number of ways from those who choose not to participate as well as from those for whom participation is a requirement. To achieve the RCT standard of experimentation, one could randomly assign students to either a course with no service-learning or to a similar course that contains service-learning. However, requiring students to participate in service-learning when they would rather be in the non-service-learning course (or vice versa) is likely to be problematic, given that students' motivation to participate and preferences for particular kinds of service-learning experiences influence the potential for students to achieve positive personal outcomes as well as the intended educational outcomes (e.g., Moely et al., 2008).

Although observed differences between community engagement participants and nonparticipants could be related to the effectiveness of a program or course, they might also be explained by other factors or variables that were not controlled for or considered. When potential differences between the participant and nonparticipant groups are not considered, it is not possible to speak definitively about causal impacts or program effectiveness, given that different outcomes could as reasonably be attributed to group differences as to the program. In such cases, researchers cannot dismiss the possibility that differences in outcomes are due to differences in student groups and not to students' experience in their programs. Program effectiveness may be influenced by students who elect to participate voluntarily, as those students may already be more engaged or receptive to the learning, topic, or activity than students who did not choose to participate voluntarily. Differences between groups caused by variables other than the program or course being evaluated are a major problem in determining the effectiveness of programs that result in positive outcomes for colleges and students.

Consider, for example, students who choose to take a service-learning course rather than a course that does not include a community-based learning approach. Even if those students were similar to (or matched with) other students on their sex, race/ethnicity, and college major, missed domains in which differences occur can create problems for drawing inferences about the program's effectiveness. For example, students who must work while attending college may not have time or flexibility to participate in activities that compete with their paid employment, such as student activities and clubs, volunteering, or unpaid internships. Factors such as financial need may create differences between students related to socioeconomic status, along with other factors such as merit scholarships, prior achievement levels, and access to external resources to support their education. These differences can be illustrated by a conversation we had with a student living in a neighborhood where many of our community engagement experiences occur. This student said she already contributes to the community by holding a job in it, and that she could not afford to do unpaid service when she needs the money from working for paying tuition. Not only is financial need an important concern for this student, but differing attitudes toward what community engagement means may also be a measure important to consider but difficult to apply in matching students in a research study. If a goal is to draw inferences about the effectiveness of a community engagement program on outcomes like retention or graduation, and randomized control trial experiments are not possible, matching students on other observable measures that might be related to the outcomes is important.

If we were able to *exactly* match college students on all variables that potentially provide an alternative explanation for program outcomes, including background and other variables like college of enrollment, major field, and prior achievement levels, we probably would have a good enough match to make the study approximate a true experiment. Such an outcome could occur if we had measured all the background and other variables that could provide alternative explanations for group differences on which to match program participants with nonparticipants and, furthermore, if we had access to a comparison sample sufficient to contain matches. Although this may sound possible, it typically is not feasible, for finding exact matches for all relevant variables for each student who participates in a particular course or program in a pool of students who do not participate in that course or program is an exponential problem. If a program is small, exact matching may be possible at a large university where there are likely many potential matches for each student participating in that program. However, when the size of the program and/or the number of variables to control is large, finding exact matches for program students on all the variables becomes difficult if not impossible, particularly when attempting to exact match on variables that are continuous with many different levels (e.g., high school GPA or ACT/SAT scores). Even when considering the impact of measured variables, other unmeasured or unobserved variables such as engagement and motivation to participate in the program may also impact outcomes. These variables are rarely collected in large scale and not simple to use for matching.

To illustrate the complexity of matching, even matching students only on race/ethnicity requires a large pool from which to secure matches. Race/ethnicity has many possible categories. Assuming that we group into only seven major groups—African American, American Indian, Asian American, Hispanic/Latino, Pacific Islander, White, and Other—we have students who come from multiple backgrounds and would select multiple groups, which increases the number of different categories that have to be considered. If we were using categorical variables based on the different groups to match, we would code race/ethnicity into seven binary race/ethnicity groups (yes/no for each group). If all combinations of one or more racial/ethnic backgrounds were to occur, this would result in 128 (2 × 2 × 2 × 2 × 2 × 2 × 2 = 128) different backgrounds. When adding other variables, such as gender (assuming three levels: male, female, non-gender conforming/nonbinary), Pell eligibility (yes/no), first-generation college student (yes/no), and resident status (naturalized, permanent resident, nonresident), the number of possible combinations increases multiplicatively with each variable. If applied, these variables could produce as many as 4,608 (128 × 3 × 2 × 2 × 3) possible unique background combinations to match, without even considering how to match on the continuous variables high school GPA and ACT/SAT score. As this example demonstrates, exact matching on all variables that potentially could account for finding differences between groups on outcome variables is generally highly impractical as well as rarely likely to be successful.

Alternatively, researchers could use a regression approach to control for those variables rather than trying to match them. The approach includes all the relevant background variables as covariates (leaving out one background for each variable to avoid collinearlity) and removes their relationship with all outcome variables before looking at the relation of the program (treatment) with the outcomes. Using the same example, if researchers were to consider a regression approach as an alternative to matching and included all the same variables in the analyses as covariates, they would have 12 (6 race/ethnicity + 2 gender + 1 first generation + 1 Pell elgible + 2 citizenship) dummy and two continuous (GPA, ACT/SAT) background variables for which they would control. However, the large number of covariates in the analyses could hinder an accurate interpretation of the findings, for they are likely to be interrelated with one another as well as potentially with the program. Relations with the program could occur if the program were more effective for students from some backgrounds than for students from different backgrounds, but also if a disproportionate number of students in the program were of a particular background.

So, are there other options for researchers who are interested in evaluating the effectiveness of their educational programs when they are not able to randomly assign participation or create exact matches? One method increasingly being used as an alternative is propensity score matching (PSM). PSM is a quasi-experimental approach that matches participants with nonparticipants, matching on the probability that a person would be a participant in the program. Using such an approach when random assignment is not possible can help strengthen the equivalency between a treatment group (e.g., students participating in community engagement) and a comparison group (e.g., students not participating in community engagement), reducing the probability that noted differences in outcomes between groups are due to relations of background characteristics of students with participation differences (i.e., students' self-selection into the community engagement program). When used effectively, it provides two groups made up of individuals with comparable likelihoods of participating in the program, allowing stronger assertions about the impact of the program.

Propensity Score Matching

PSM attempts to capture the strengths of experimental designs in instances when random assignment is not possible; PSM emulates random assignment. As described above, in many situations it is not possible to randomly assign participants to conditions when attempting to evaluate the effectiveness of postsecondary education programs. PSM provides a useful approach for matching individuals across conditions and thereby better determining the effectiveness of treatments. It has been employed widely in medical research, but only more recently has it become regularly used in the social, behavioral, and educational sciences (e.g., Fan & Nowell, 2011).

A main goal of PSM is to establish group equivalency between the treatment and comparison groups. It statistically removes confounds caused by preexisting differences between the treatment group and the non-treated (comparison) group on extraneous, uncontrolled variables, producing similar groups on which to evaluate effects of the treatment. For those infrequent instances where the two groups do not differ—that is, there are no differences on an array of potentially confounding variables between individuals selecting and experiencing a treatment and others not receiving that treatment—approaches like PSM are not needed, given that the two groups are essentially equivalent. In those instances, direct comparisons of the different groups without adjusting for covariates are appropriate.

For those more common instances where differences exist between groups, an approach like PSM can create comparable groups and overcome selection bias. If comparable groups can be created, PSM provides an approach that separates relationships between the controlled variables and the outcome variables from the relationship (effect) of the treatment/program with the outcome variables. PSM eliminates the possibility that the relationship of the treatment/program with outcomes could be due to differences between groups based on other variables that are measured and included in the PSM analyses. Even if the two groups can be made comparable, PSM depends on investigators who collect data on important background variables and who consider a full range of alternative explanations involving background variables when positing relationships between program/treatment and outcome variables.

Understanding Propensity Scores

PSM techniques use information from relevant variables that have been measured previously in related studies of the same participants, in addition to any available pretest scores or other variables pertaining to the participants, to produce a score that represents the likelihood (probability) that any individual will have participated in the program being evaluated. Analyses should include all potentially confounding variables noted through observations and/or previous studies as well as potentially including variables collected that are not necessarily expected to be related to program participation. Their inclusion allows researchers to confirm that these additional variables are not related to program outcomes; erring on the side of inclusion is preferable. The resulting unidimensional score, as described in more detail below, is called a *propensity score* (PS). Rather than using random assignment, matching is performed by pairing individuals from the treatment and comparison groups who have the same propensity score. Matching participants with nonparticipants on that score creates groups that are matched collectively across the set of measured variables.

What Makes Propensity Scores Good for Matching?

Using the language of PSM, a propensity score (PS) is the probability of exposure to a specific treatment or program conditioned on observed variables (e.g., Austin et al., 2007). A propensity score is a single numerical value for each individual, calculated from the covariates (often called *conditioning variables*). Propensity scores range from 0 (no chance of being in the program/treatment condition) to 1 (definitely in program/treatment condition). The score is the likelihood or probability that an individual will/did participate in the treatment/program being assessed. Propensity scores are used to match participants enrolled in a program or treatment with similar individuals about whom the researchers have data but who did not participate in the program. Propensity scores are calculated by regressing the treatment/program participation variable (participates/does not) on a set of potentially confounding variables. In principle, individuals with identical PSs have an equal probability of being in the treatment/participation group. Thus, PSs provide a statistical matching on the set of key background and prior performance characteristics by controlling for the relationship of all those covariates with the treatment or program. After matching, the two groups ideally are matched on all the measured background variables, which means that the relations of those background variables with the treatment or program are removed by controlling them, allowing stronger "apples-to-apples" inferences to be drawn from comparisons between groups (e.g., Rosenbaum & Rubin, 1983). If the two groups are too different on the background variables and cannot be made comparable, PSM is not appropriate.

How Are Propensity Scores Generated?

Propensity scores are generated using the following steps. First, prior to data collection, it is useful to develop a conceptual map tracing how the program ideally would work and the background and demographic variables that would need to be collected in order to eliminate any confounding effects that may account for resulting group differences. Second, during data collection, investigators need to collect the full array of variables in their conceptual map of how the treatment works from a comparison group as well as program participants. Ideally, the comparison group would be larger, providing more opportunities for identifying good matches. Third, mean differences on control variables between the program/treatment group and the comparison group are examined. If no differences between the means of the groups exceed .05 standard deviations, the groups can be judged to be equivalent on the background variables, and simple mean comparisons on the outcome variables can be conducted without using PSM. In the more likely case where mean differences in background variables between groups exist, PSs are created by regressing the binary program variable (assuming a single program) on the full set of background and demographic variables and then using the regression weights for the predictors to calculate predicted scores for each individual. Those predicted scores are the PSs. Fourth, the PSs are used to create individual-level treatment/comparison group matches. Individuals are matched on PSs across the program (treatment) and comparison groups. Before the groups can be compared on the outcome variable(s), additional steps are required to see if any group differences remain; how these are handled will be explained after finishing the discussion of propensity scores.

To paraphrase Rosenbaum and Rubin (1983), the resulting PSs can be used as a unidimensional balancing score where each subject's PS becomes a summary of the pretreatment covariates, such that treated and comparison subjects who have the same PS have a balanced joint distribution of the pretreatment covariates. Two individuals with the same PS can be considered matched, yielding analyses that produce in principle an unbiased estimate of the treatment effect. By controlling other variables, PSM is preferable to simply accepting a nonequivalent comparison group, for it in principle eliminates a number of alternative explanations for differences between groups.

Individuals matched by PSs should approximate random assignment; each student who participated in a program is paired with a student having an equal (or similar if equal is not available) likelihood of participating, but who did not participate in the program. Matching is "approximate" because the effectiveness of the matching is dependent on the particular set of covariates available and selected, and because of the overlap of the two sets of PSs. Identifying and measuring a robust set of covariates helps ensure better matching. Covariates selected for the propensity score model should be conceptually identified as and/or empirically found to be related to both treatment and outcome. Their inclusion as covariates will prevent them from potentially influencing the program's relations with the outcome variables. If there is uncertainty, it is better to err on the side of overinclusion rather than risk excluding potentially important covariates. As noted above, unrelated covariates should not affect the regression analyses, for they will not be related to the program/treatment and will have negligible weights in determining propensity scores. After controlling for appropriate covariates, researchers can claim that treatment assignment is conditionally independent of potential confounding variables that might provide alternative explanations for observed outcomes. The language of PSM describes the effect as "conditioned on the covariates." Propensity score matching rests on the principle that participants in treatment or comparison conditions with identical PSs will have the same probability of being in the treatment condition.

Why Not Just Covary Potentially Confounding Variables?

Earlier we noted that an alternative way of addressing the impact of variables that might provide various explanations for the findings is to include those variables in a regression analysis. By including variables that may be related both to treatment assignment and outcomes, researchers can then statistically judge their impact on the relationship between the treatment and the outcome, and ideally also control for differences due to those variables. This approach is known as "controlling for" potential covariates in multiple regression.

Covariate control is a widely accepted method in statistics. However, matching methods via PSM provide certain practical advantages important to consider. In regression, when multiple variables are involved, the shared variance is attributable to different predictors, which can leave interpretation ambiguous, especially when extraneous variables are highly related to program participation. Propensity scores reduce the array of covariates included to one overall unidimensional score, eliminating the need to include a large number of covariates for regression adjustment (Hong, 2015) and reducing interpretation ambiguity. In addition, PSM allows researchers to assess the covariate distribution between groups before the outcome analysis; regression adjustment during outcome analysis may be unreliable if both groups are far apart (nonequivalent) in covariates (Rubin, 2001). Further, various PSM approaches eliminate individuals who are outliers, which reduces outcomes being unduly influenced by individual extreme cases. Expanding on the prior point, for most PSM approaches, a priori examination of covariates results in the selection of a more balanced subsample by eliminating individuals who cannot be effectively matched. PSM also eliminates the relationship between covariates and the treatment or program variable before looking at the relationship between the treatment/program and outcomes; regression is influenced by interrelations among covariates and the treatment variable. Finally, various authors have pointed out that regression adjustment may increase bias in the treatment effect if the relationship between the covariates and the outcome is even slightly nonlinear (see Stuart, 2010 for review). For these reasons, PSM can provide a better balancing of covariates across treatment and control than covariate adjustment used in regression.

It is important to reiterate that PSM creates propensity scores in a process that occurs *prior to* examining relations of the program/treatment with outcome variables. Similar to other regression family approaches, creating matched groups in the preliminary stages of the analysis may reduce bias and increase the precision of the covariate adjustment in the outcome model (Rubin & Thomas, 1996). Because relations of the treatment with the outcome variable(s) have not been examined during the matching process, PSM allows researchers to try different PSM methods to find the one that does the best job of producing equivalent groups.

What Constitutes Well-Matched Groups?

Thus far, we have assumed that we will be able to create well-matched groups. As noted earlier, however, if we cannot, then PSM is not an appropriate approach, for it works only when groups can be well-matched. To determine the appropriateness of the matching process, after the matched samples are created, all the covariates are related to the treatment variable to examine the magnitude of remaining differences between the program and comparison groups. A set of principles has been adopted to define acceptable differences and to provide options if the groups are not completely matched. As explained above, because the matching process occurs before looking at relations of the program with outcome variables, we recommend trying different PSM matching approaches for generating PSs, and seeing which approach provides the best combination of match and power. Once we select the approach, we impose the decision rules on the chosen PSM approach.

First, if differences in all of the covariates have been reduced to less than .05 standard deviations (*SD*s), simple mean comparisons can be used to assess program effectiveness. If, however, some covariates remain unbalanced with differences greater than .05 *SD*s, we then examine how much greater the remaining differences are. If differences on all covariates are above .05 but less than .25 *SD*s, we can use PSM. We include covariates with differences between groups of greater than .05 *SD*s in the final regression model predicting the dependent variable to be able to control for their remaining relationship to the treatment/program and provide a more accurate estimation of the association between treatment and outcome somewhat independent of the covariates (Zanutto, 2006). If remaining differences between the two groups still exceed .25 *SD*s with the best PSM approach, using PSM is not possible, for in such situations, there is insufficient overlap between the comparison and treated subjects' PSs.

Challenges When Using PSM

As just described, overlap, called *common support*, is necessary to create well-matched groups. Even though weighting to balance program and comparison groups with little or no overlap can be done, PSM is less likely to prove viable because the differences on potentially confounding variables cannot be eliminated. Bai (2015) identified 75% of overlap as the minimum requirement for creating comparable matched groups. Finding no overlap or too little overlap likely indicates that there are too many pretreatment differences between groups, which hinders researchers' ability to draw reliable causal inferences (e.g., Harder et al., 2010). At best, lack of overlap would result in having to discard many participants from the outcome analysis, which would lead at minimum to a reduction in sample size and, consequently, loss of statistical power (e.g., Lane et al., 2012). Even more problematic, it may result in retaining a matched subsample that is not representative of the population from which it is drawn.

A second challenge to PSM occurs with any approach that tries to substitute for random assignment by matching on an array of background and other variables to establish group equivalence. Such a strategy may be limited insofar as it can control only those variables that are observable and that have been measured, which may fail to eliminate fully preexisting group differences that are attributable to other relevant confounding and unmeasured variables. Using a Head Start program as an example, even with a number of appropriate controls, children might still differ on other unmeasured but important variables like the kinds of television programs they watch, their grandparents' education levels, the number of books in their homes, the achievement levels of their friends, and so on. If these variables are important but are not considered, and the treatment group is, in actuality, significantly lower on these unmatched variables, then the final results would be biased in favor of the comparison group. If differences are in the opposite direction, bias would favor the treatment group. The number of variables on which groups are not matched is potentially infinite. When remaining unmatched or undermatched, differences for compensatory programs likely favor the comparison group; in such instances, even effective intervention programs may look harmful or ineffective as a result of the failure to equate groups. It is difficult to know when researchers have matched on enough variables to ensure that the two groups are equivalent, and, for a program like service-learning, the direction of differences on unmeasured variables may vary from setting to setting. Fortunately for PSM, there is some evidence (e.g., Rosenbaum & Rubin, 1983) that it can control for bias from covariates, for many are related to measured covariates. Whether that is true for all settings is not clear; to the extent possible, researchers should carefully plan the covariates that are to be in the design.

One point that should be clear from the second challenge is that selecting the set of covariates is critical. Not surprisingly, there are different views about how the set of covariates should be selected (e.g., Austin et al., 2007).

- One view is to include those variables that are related to treatment assignment.
- A second is to include all variables potentially related to the outcome variable.
- A third is to include only variables associated with both treatment and outcome.

Findings from a Monte Carlo study by Austin et al. (2007) suggest that combining the first and second perspectives is best: The most effective approaches include as covariates variables that are theoretically related to treatment assignment as well as variables related to the outcome variable. These findings are consistent with our experiences as producing findings with the least ambiguity. The U.S. Department of Education's What Works Clearinghouse, in its efforts to emphasize trustworthy, science-based evidence, acknowledges the importance of these characteristics by requiring that at least one socioeconomic background variable and one prior achievement measure be measured and used as control variables for PSM when looking at educational outcome variables.

To recap the second criticism, all variables that potentially could provide alternative explanations for differences between groups on the outcomes of interest ideally comprise the set of covariates/conditioning variables. Not fully controlling for such variables allows them to confound the study, possibly reducing a PSM to a nonequivalent comparison group design. With nonequivalent groups, there are alternative explanations for differences between groups in outcomes. Challenges come when one or more of the potentially confounding variables are unobserved or unmeasured. In some instances, a sensitivity analysis could be conducted to assess the extent to which the estimate would change if an unmeasured covariate were included (see Groenwold & Klungel, 2015; Hong, 2015).

A third criticism of PSM is related to misspecification of the logistic model predicting treatment. Misspecification occurs when a key covariate, that is, a covariate that is highly related to the treatment assignment, is omitted from the propensity score model. This omission leads to a misestimation of the PSs, resulting in biased estimators of the treatment effect (Drake, 2017). Researchers need to ensure that the covariates represent the possible confounding variables related to the implementation of the program, ideally including measures of prior outcomes. In the Head Start example, unmeasured variables like TV programs watched, grandparents' education, and books in the home could all provide alternative explanations for group differences and might have had important regression weights. Having a strong conceptual framework as well as drawing from prior research studies related to the topic under investigation helps to guide identification of possible confounding variables. To help address this potential criticism, many authors describe in detail the theoretical bases, the prior literatures, and the statistical methods they used to determine which covariates to include in the PS model in order to minimize possible misspecification (e.g., Harder et al., 2010; Pattanayak, 2015).

Finally, researchers' decisions about different PSM approaches may affect their findings. The selection of different matching methods or the way specific matching methods are used could result in differing results. In our experience, we never found a perfect matching procedure for a given data set, and we typically tried different approaches to see which provided the best sample. In using PSM, researchers have to make decisions about what to prioritize and accept: maximizing sample size, obtaining the highest quality matches, or selecting acceptable matches. We explain these processes in detail later.

At this point, having provided a summary of advantages as well as potential challenges to address in using PSM, we turn to specific steps in conducting PSM. After that, we describe how it was used to investigate the effectiveness of a community engagement program in which college students from underrepresented populations tutored middle school and high school students. In this study, we assumed the typical case for PSM, that the treatment variable (participated/did not participate) was dichotomous.

Steps in Conducting a Propensity Score Matching Analysis

Step 1. Identify the Variables That Could Account for Favorable or Unfavorable Outcomes

Before analysis, and preferably prior to data collection, it is important to consider variables that potentially could affect the relationship between the treatment/program and the outcome variables. The extant literature on the topic being studied should provide some guidance as to which variables should be included. Identifying covariates in the initial model is critical for establishing comparability between groups, as controlling them should allow one to estimate effects of the treatment program independently of those variables. To the extent possible, such variables need to be measured, for only variables that are measured can be controlled statistically.

As discussed earlier, some researchers suggest that variable selection during this stage should identify variables having a theoretical relationship to participation in the treatment as well as to the outcome variables (e.g., Caliendo & Kopeinig, 2008). Other authors, however, employ statistical approaches for selecting covariates. As an example of the latter, Harder et al. (2010) described testing and comparing three different logistic models: (1) a parsimonious model that includes only the covariates, (2) a more complex model that incorporates some interaction terms, and (3) a generalized boosted model that can include the same terms as the former model but in a nonparametric manner. Although combining both theoretical and statistical guidelines for the selection of covariates in the PS model is reasonable, concern remains about using any outcome variable as a consideration within a PS model (Pattanayak, 2015; Rubin, 2001). Specifically, statistical approaches that require researchers to view correlations between potential covariates and treatment before final outcome model specification potentially introduce bias in the final model structure, which is not recommended (e.g., Rubin, 2001).

The following points are guides for thinking about covariates:

- Identify possible control (conditioning) variables and see how many have been or can be measured. Measure as many as possible. An example from a study of a community engagement program is described in detail later in this article. For that study, we included as covariates sex, ethnic/racial background (dummy coded), prior achievement (ACT or SAT), citizenship status (international, U.S. born/naturalized, permanent resident; again, dummy coded), family income (Pell eligible or other), first-generation college student, honors program participation, and college of enrollment. We recognize that in some instances information on citizenship can be sensitive due to immigration policies and the way they currently are enforced, which may preclude obtaining that information, even with deidentified data.
- A criterion for determining which potential covariates to include in the matching process is that of *strong ignorability*. PSM assumes that there are no unobserved differences between the treatment and control groups, conditional on the observed covariates. In other words, the assumption is that after PSM, the resulting matched groups are similar enough that any difference in the outcome is attributable solely to the treatment. If researchers know about missing variables and feel confident that they know the implications of those unmeasured variables, then they could try to model them, even though doing so is challenging and may be open to criticism.
- As was noted earlier, using as an example the United States Department of Education's What Works Clearinghouse (WWC), at least one prior achievement variable and one prior social class/economic variable need to be included in the control variables for a PSM study to qualify for WWC publication.
- If samples include underrepresented groups, such as students of color, low income, first generation, and students with disabilities, those variables should be included as covariates in order to eliminate differences on those characteristics as reasons for the outcomes.
- Variables that may have been affected by the program should not be included in the matching process (e.g., attitudes about community involvement measured during participation in such a program). Including them eliminates or diminishes researchers' ability to determine effects produced by the program.

Ideally, one should include in the matching procedure all variables known to be related to both treatment assignment and the outcome (Glazerman et al., 2003; Heckman et al., 1998; Hill et al., 2004; Rubin & Thomas, 1996). There is little downsidsde to including variables that are actually unassociated with treatment assignment, as they will be of little influence in the propensity score (PS) model. Said differently, in computing PSs, collinearity may be relatively unimportant (but see also Zhang et al., 2019), for the goal is to optimize prediction of each individual's likelihood of being in the treatment condition so the matching works well. Only variables that predict participation will have meaningful weights, thus any other variables will not add to the model's prediction. The important point is that when maximizing explained variance, including variables is preferable to not including them. As noted earlier, exact matching on control variables is ideal, but typically not possible when a large number of confounding variables exist, thus warranting PSM.

Step 2. Estimate Propensity Scores

Once the conditioning variables are selected, estimate the PSs for each individual for whom data are available, both those participating in the program of interest and others who are potential comparison group members. Create a logit model using the observed covariates to predict the binary treatment variable (participated/did not participate). The predicted probability of each individual being in the program is their propensity score (PS), generated from the logistic regression, and calculated for each individual based on the selected covariates.

Once the PSs are available for all individuals, assess the degree of overlap (common support) between the PSs in the treatment and comparison groups before choosing a matching technique. This common support can be visualized and assessed by comparing graphs of the density distribution of the PSs for each group (Bai, 2015; Caliendo & Kopeinig, 2008).

Explore Possible Matching Approaches

The next step is to define acceptable "closeness," examining the distances between PSs of matches to determine whether an individual is a good match for another. One way to control matching is to specify the maximum distance allowable between matches. Specifying and using a maximum allowable distance is described as setting up a *caliper*. Rosenbaum and Rubin (1985) recommended using a caliper of a PS distance of 0.25 standard deviations to provide enough of a constraint on matches without sacrificing possible matches. Matches then occur only for scores less than the caliper distance apart. The most typical caliper is a difference in PSs of 0.2.

A second matching decision addresses the individuals whose PS scores fall outside the range of scores found for both the matched groups. Most commonly, those would be larger PSs (higher probabilities of being in the treatment/program) for individuals in the treatment group and lower PSs (lower probabilities of being in the treatment/program) for individuals in the comparison group. Excluding these individuals may benefit the quality of the matched groups, since more extreme cases would be less likely to have effective matches. When deciding who to include and exclude, one approach is to retain individuals whose scores are "close" to the scores of the other group (based, perhaps, on the standard errors of scores to help decide how far beyond the other group would still be a reasonable match—and thinking about calipers) and to exclude those that are beyond the selected range.

Once decisions are made about which individuals are good candidates to include in the matching process, the next step is to select and implement a matching method. Because matching methods are chosen *before* looking at the relationship of any matched groups with the outcome variables of interest, matches are not selected to maximize differences between groups on outcomes, but rather are selected to reduce differences between groups in the matched variables in order to create groups that are as similar as possible to one another. As is discussed in the next section, there are a number of different matching strategies to choose from. Therefore, if the first matching technique selected does not produce a good match, it is appropriate to try other matching approaches to determine which one produces the best possible matching of the groups.

If the treatment/program participation group is small compared to the full population (e.g., a small program within a large college or university) for which data are available, researchers can consider selecting what is called an N to 1 match rather than a 1 to 1 match. An N to 1 match allows multiple individuals from the comparison group to be matched to each individual in the treatment group. In such instances, weighting may be necessary to "balance" the groups. Weighting involves averaging across multiple good matches to provide more stable findings rather than arbitrarily selecting only a single individual for matching when many strong matches are available. If appropriate, one may subclassify or weight (prior to selection) the matches, then select the best matches within subgroups.

Step 3. Select a Matching Method for PSM

As was noted earlier, a number of different approaches are available. Those that use pairwise matching typically include a caliper to establish the maximum allowable distance between matched pairs. Other approaches try to retain as many individuals as possible, but use weighting rather than matching to keep balance across conditions. Among the most common approaches are the following:

1. *Nearest neighbor* (NN) matching. With NN, matching is performed sequentially (stepwise), so the order in which the treated subjects are matched may affect the quality of the matches. Because NN is performed randomly, each instance of NN can produce different matches, for the starting point for individual matching changes. This matching approach is often described as a "greedy" approach, for, because of the sequential nature of the matching, earlier matches may "use up" the best matches for individuals who are matched later. Typically, NN is selected without replacement, making any comparison individual eligible for only one match, which limits later

matches to those comparison individuals remaining unmatched. However, sometimes matched individuals are kept in the matching pool after being matched, allowing a comparison individual to be matched to more than one treatment individual. In such instances where a comparison individual is matched against several treatment/program participants, weighting that comparison individual more heavily to balance the size of the groups is suggested. Note that this matching is opposite to N to 1 matching, which underweights comparison individuals, whereas NN overweights them. In this case the characteristics and outcomes of this comparison subject need to have a heavier influence in the final outcome model when compared to other comparison subjects. Alternatively, N to 1 matching underweights comparison individuals so their outcomes have a lower impact on the final outcome model.

2. *Optimal pair matching* (OM). Like NN, the OM approach matches each individual program participant with an individual from the comparison group. In contrast, however, the OM approach minimizes the squared distances between matches across groups at a sample level. OM provides the best possible full sample matches by finding the smallest possible total squared differences in propensity scores between treatment and comparison groups. This approach is preferable if one wants to optimize well-matched pairs within the matched groups. Like NN, the OM approach matches program participants individually with an individual from the comparison group.

3. *Full matching* (FM). The FM approach finds pairs or groups of treated and control participants that are close based on the distance measure. It ideally keeps the full sample, limited only by eliminating individuals who fall outside the range of scores where there is overlap of the groups (common support). The ratio of matching (1:4, 3:2, etc.) can be selected on an a priori basis or by a caliper to constrain the groups. These groups are then used to create regression weights that are incorporated into the outcome analysis in order to balance the sizes of the groups—most often weighting the comparison sample to the sample size of the treatment group.

4. *Inverse weighting* (IW). In instances where PSs of the treatment group are much higher than those of the comparison group (e.g., where the treatment PSs are negatively skewed and the comparison group PSs are positively skewed) and the prior matching techniques will not work, it may make sense to upweight individuals on the smaller tails and downweight individuals in the larger part of each distribution for each group separately. Below is a formula for inverse weighting, which keeps all the individuals but weights cases differentially, with larger weights for treatment participants who have low PSs and smaller weights for comparison individuals with high PSs. For inverse weighting, think of two very different distributions of scores that have some overlap, with more of the higher PSs found in the treated individuals, and more of the lower PSs found in comparison individuals. This weighting formula ideally provides a better matched set of scores when groups differ substantially. With the inverse weighting formula,

$$W_i = T_i/e_i + [(1 - T_i)/(1 - e_i)],$$

e_i is the estimated propensity score for individual i, and T_i is treatment condition (treated = 1, control = 0).

As noted earlier about PSs, the likelihoods of being in the treatment condition range from 0 to 1. For inverse weighting, if PSs are close to 0 for treatment group individuals or close to 1 for those in the comparison group, the weights will get large. Having to use very high weights for the outlier cases in large samples with oppositely skewed distributions can amplify the influence of atypical individuals within their group. When using inverse weights, one should look at the distribution of weights to make sure there are not extremely large weights. One option is to discard cases with really big weights. For weights that are large but that seem to be part of the main distribution (i.e., greater than 5), one possibility would be to cap them at a maximum value so they are not too heavily weighted.

One other distance metric, similar to PSs, that we have not described is the *Mahalanobis distance*, which employs a geometric distance to match cases. It provides an alternative scale-invariant, multidimensional measure of the distance between two individuals. For instances where other matching approaches do not produce quality matches, Mahalanobis distance matching may produce better balance on background characteristics since it takes this different approach.

Step 4. Assess the Quality of the Matching

Ideally, at least one of the methods of generating matches in the matching process results in well-matched samples. Well-matched samples occur when mean differences between the groups on the covariates are small in standard deviation differences. As noted, in some instances, one may have to try multiple matching approaches, sampling until well-matched samples result, for conducting PSM analyses requires samples to be well-matched. Pattanayak (2015) suggested estimating the standardized difference in means between treatment and comparison groups on all pretreatment covariates. Rubin (2001) added two more balance measures besides the one suggested in Pattanayak: (1) the standardized difference in means of the PSs and (2) the ratio of the treatment and comparison PS variances. Bai (2015) provided yet another measure, recommending using the percent of bias reduction as a criterion to assess balance. For the standardized difference between means of covariates, Rubin (2001) and Stuart (2010) recommended using 0.25 as the cutoff score to determine balance.

If after matching there still are differences greater than 0.25 standard deviations between the groups on any conditioning variables or other critical variables, then using PSM is not appropriate. It should, however, be noted that if a first matching does not yield well-matched samples, it does not mean that successful matching is not possible. One can rematch to try to make the differences smaller, which might work since starting points for NN matches are random, as are matches for pairwise matches when multiple possible matches are available. (For example, reestimating using NN matches, which are taken sequentially from a random starting point, produces different matches.) If covariates remain unbalanced, additional modeling or other considerations should be explored, such as using exact matching on one or more covariates (Pattanayak, 2015), adding more covariates, or including interaction terms in the original logistic model (Harder et al., 2010). One also can try weighting cases to balance on the problematic variable(s), or creating subgroups within the treatment group and then matching within subgroups to increase similarity of subgroups across treatment conditions. In addition to changing matching, one should inspect distributions visually to see if/how patterns can be understood and controlled. Again, if group differences in conditioning variables cannot be reduced to less than 0.25 standard deviations, then PSM is not appropriate, given that the remaining differences between groups are too great and cannot be eliminated by controlling covariates.

As long as differences remaining on each matching variable are less than the 0.25 standard deviations that preclude the use of PSM, one can proceed with PSM analyses. If, however, some differences less than 0.25 standard deviations still exceed 0.05 standard deviations, those matching variables have not been fully controlled through the matching. They should, therefore, be included as covariates in the analyses to control for their effects more fully.

Step 5. Analyze the Outcome Variable(s) and Estimation of the Treatment Effect

Once it is determined that PSM is appropriate for the sample, analyses comparing the groups should be straightforward. If no differences between groups on the covariates exceed 0.05 standard deviations, no covariates need to be included in the analyses. One can use *t*-tests for continuous variables or chi-square for dichotomous outcome variables to determine whether the groups differ. When differences between comparison and treatment groups on some conditioning variables exceed 0.05 but are less than 0.25, analyses of covariance (ANCOVAs) or logistic regression are appropriate, with analyses controlling for variables with preexisting differences greater than 0.05.

Step 5a. What to Do When Discarding Some Treatment Participants Based on PSs: ATE Versus ATT Findings

In discussions of causal effects, it is common to find estimates described as the Average Treatment Effect (ATE) and the Average Treatment on the Treated (ATT). The ATE, the treatment effect for the entire treatment group, compares all individuals in the sample. If the assignment on treatment is unconfounded (i.e., this assignment is independent of potential outcomes conditional on covariates), one can average the differences between groups to estimate the ATE. In some instances of PSM, however, researchers are unable to estimate the ATE because program participants had to be dropped when creating the matched sample (e.g., subjects who fall outside the common support and are not included in the matching of groups). Studies where some of the treatment participants are dropped become ATT, which only compares subjects successfully matched with a comparison individual with similar probabilities of being in the treatment condition. The loss of treatment participants whose PSs are not similar to any individuals in the comparison group limits the inferences that can be drawn (Dehejia & Wahba, 2002; Stuart, 2010).

Using PSM to Evaluate Community Engagement Outcomes

To illustrate the use of PSM in evaluating community engagement program outcomes, we provide an example in which we applied PSM to a study of a college program involving a YMCA in a Midwestern city in which college students offered mentoring and tutoring to local youth through the YMCA. The program provides extracurricular activities to engage college students as they mentor and tutor local youth. This off-campus program hires primarily underrepresented college students to work with middle-school youth from diverse backgrounds in an after-school program. The program's mission is to facilitate meaningful community engagement by providing college students opportunities to apply their knowledge and skills to help community members while building friendships with other mentors and tutors. This experience is a paid community-engaged employment opportunity designed to address the financial needs of the participating college students while allowing them to apply their skills and knowledge in ways that directly address a community need (Schulzetenberg et al., 2020).

We set out to investigate whether participating in community-engaged employment (mentoring and tutoring local youth) at the YMCA was associated with underrepresented college students' persistence in college, their academic performance, and the rate at which they graduated (Schulzetenberg et al., 2020). For this research, we were guided by theory, selecting covariates that previous research on community engagement found to be related to participation in a mentoring and tutoring program, plus other variables that might result in alternative explanations for our findings (Eyler & Giles, 1999). For our covariates, we selected sex, ethnic/racial background (dummy coded), prior achievement (ACT/SAT), citizenship status (international, U.S. born/naturalized, permanent resident; again, dummy coded), family income (Pell eligible or other), first-generation college student, honors program participation, and college of enrollment. To strengthen our match, we also exact-matched on the year each participating college student entered as a freshman.

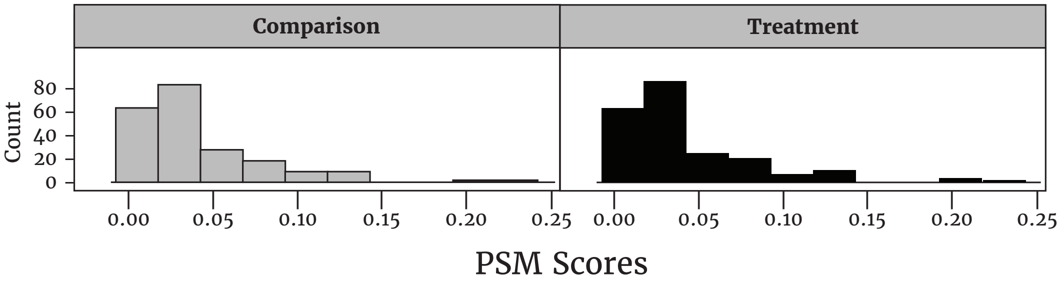
After deciding on covariates, we ran logistic regression to generate propensity scores (PSs). We then tested assumptions by examining the degree of overlap or common support between the PSs in the treatment and the much larger comparison group. As mentioned, this common support can be assessed graphically by comparing the density distribution of the PSs for each group (Bai, 2015; Caliendo & Kopeinig, 2008).

After finding sufficient overlap of the two groups, we worked on matching PSs of students who participated (n = 216) with scores of those of students at the same institution who did not participate (n = 52,693). For pairwise matching approaches, we used caliper matching to set a maximum distance allowed between PSs (in this case, 0.2 *SD*s).

The distribution of scores demonstrated that inverse weighting was not necessary, but we did run analysis using different matching approaches (nearest neighbor, optimal pair, and full matching). As noted earlier, testing different applications and comparing balance between different matches to determine the best possible set is widely accepted as an effective practice, for we at that point had not looked at the outcome variables (e.g., Austin, 2011; Kretschmann et al., 2014; Lanza et al., 2013). From among the different approaches, we found *optimal pair matching* to provide the best matches (see Figure 1).

After creating the matched groups (Figure 1), we assessed the balance across covariates between comparison and treatment groups to ensure that groups were equivalent. In our example, all covariate differences were less than 0.25 standard deviations after matching, indicating that PSM is appropriate. However, several covariates (biological sciences college, science and engineering college, Asian/Pacific Islander race category, and American Indian race category) were greater than 0.05. In order to separate the effects of program participation from these variables, we included each of them

Figure 1. Histogram of Propensity Scores for Treatment and Comparison After Matching



Note. Reprinted from “Improving Outcomes of Underrepresented College Students Through Community-Engaged Employment,” by A. J. Schulzetenberg et al., 2020, I*nternational Journal of Research on Service-Learning and Community Engagement, 8*(1), p. 9 (https://doi.org/10.37333/001c.18719). Copyright 2020 by the International Association for Research on Service-Learning and Community Engagement. Used with permission from the publisher.

as covariates in our analyses to assess the effectiveness of the community-engaged employment program (e.g., Song & Herman, 2010).

Given that differences remained that were not fully eliminated, the outcome model included treatment as the independent variable, the variables listed above as covariates, and continued enrollment (persistence), credits completed, GPA, and graduation status as the dependent variables. The results of these analyses found strong effects of program participation for each of the four dependent variables. Table 1 is included to illustrate outcomes.

In this example, PSM allowed us to build group equivalence between students who participated as mentors and tutors and students who did not, and to measure the differences between groups across key outcome variables pertaining to educational success (continued enrollment, credits completed, GPA, and graduation status). Given the number of variables whose relationships with program participation were controlled, we are able to speak more confidently about the effectiveness of program participation than we would if we did not control for such variables or if we did not have comparable groups. As we previously noted about PSM, we cannot speak definitively about causality. Nevertheless, the findings are encouraging for the program being evaluated, for we can conclude that what resulted was not due to selection differences in the array of variables that we were able to control.

Concluding Discussion

To summarize, this article has described and argued for using a quasi-experimental approach called propensity score matching for situations in which possible comparison individuals exist corresponding to individu-

Table 1. Partial Regression Coefficients for the Relationship Between Community-Engaged Employment and Academic Outcomes for Underrepresented Students (*N* = 432)

	GPA		Credits Earned		Retention		Graduation	
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Participation	0.24*	0.06	19.6*	3.5	.12*	.03	.16*	.05

Note. Analyses controlled for initial enrollment in biological sciences college and engineering college, and for Asian/Pacific Islander and American Indian backgrounds. Reprinted from “Improving Outcomes of Underrepresented College Students Through Community-Engaged Employment,” by A. J. Schulzetenberg et al., 2020, *International Journal of Research on Service-Learning and Community Engagement, 8*(1), p. 9 (https://doi.org/10.37333/001c.18719). Copyright 2020 by the International Association for Research on Service-Learning and Community Engagement. Used with permission from the publisher.
* *p* < .001.

als participating in a particular program, but where the individuals participating have not been (or cannot be) randomly assigned to the program. For researchers of community engagement programs, potential problems of nonequivalence of program participants with nonparticipants are widespread, for randomizing students into community engagement programs is often infeasible and at times unethical. In such instances, it is not possible simply to assume that the groups being compared are equivalent, for students frequently select participation in particular programs. By providing matching approaches, PSM provides a useful approach for studying effectiveness of community engagement programs on an array of student outcomes, including academic success.

PSM examines equivalence of the groups being compared first by collecting information on a number of background and other variables that are thought to be related to the outcomes of interest, and then by examining differences between the groups on all those variables. If there are differences between groups, PSM attempts to control for those differences to uncover relationships between program participation and outcomes that are independent of those other variables. It accomplishes that goal by matching individuals across groups who have the same likelihood of participating in the program of interest, emulating the process of random assignment where each individual has the same likelihood of being in the treatment group. Once groups are successfully matched, analyses comparing groups can be conducted using the traditional methods, such as *t*-tests, chi-square tests, or regression analysis.

Establishing an argument for causal impacts of community engagement or any other program is integral for building programming on college campuses. Findings from PSM studies like the one summarized in this article illustrate how the method can enrich the evidence base for effectiveness of community engagement programming and promote its use and continued support in higher education programming.

As is true of any analytic approach, PSM has limitations. In deciding whether or not to use PSM techniques, researchers should consider their sample and the variables available, for PSM is not always going to be useful or provide accurate findings. Small sample sizes, particularly in the potential pool of comparison individuals, and a limited availability of variables to be used as covariates both can greatly hinder the quality of the matches and the accuracy of the estimates.

In closing, researchers should consider adding PSM to their toolbox of methods for examining effectiveness of community engagement programming. We hope this overview of PSM has increased awareness of PSM's potential usefulness and has provided researchers with some basics of applying PSM approaches to help understand the impacts of community engagement programs on student outcomes.

Acknowledgment

The contents of this article were developed in part under grant #P116140033 from Fund for the Improvement of Postsecondary Education, First in the World program, the U.S. Department of Education. However, the contents do not necessarily represent the policy of the U.S. Department of Education, and endorsement by the Federal Government should not be assumed.

About the Authors

Geoffrey Maruyama *is a professor in the Department of Educational Psychology at the University of Minnesota.*

Isabel Lopez, Ph.D., *is a postdoctoral researcher at Center for Scientific Research and Higher Education at Ensenada (CICESE) in Baja California, Mexico.*

Anthony Schulzetenberg, *Ph.D., is UX research lead at LexisNexis.*

Wei Song, Ph.D., *is a research scientist at A. J. Drexel Autism Institute, Drexel University, Philadelphia, PA.*

References

Austin, P. C. (2011). A tutorial and case study in propensity score analysis: An application to estimating the effect of in-hospital smoking cessation counseling on mortality. *Multivariate Behavioral Research, 46*(1), 119–151. https://doi.org/10.1080/00273171.2011.54048

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine, 26*(4), 734–753. https://doi.org/10.1002/sim.2580

Bai, H. (2015). Methodological considerations in implementing propensity score matching. In W. Pan & H. Bai (Eds.), *Propensity score analysis* (pp. 75–88). Guilford.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics, 84*(1), 151–161. https://doi.org/10.1162/003465302317331982

Drake, C. (2017). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics, 49*(4), 1231–1236. https://doi.org/10.2307/2532266

Eyler, J., & Giles, D. E., Jr. (1999). *Where's the learning in service-learning?* (Jossey-Bass Higher and Adult Education Series). Jossey-Bass.

Fan, X., & Nowell, D. L. (2011). Using propensity score matching in educational research. *Gifted Child Quarterly, 55*(1), 74–79. https://doi.org/10.1177/0016986210390635

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science, 589*(1), 63–93. https://doi.org/10.1177/0002716203254879

Groenwold, R. H. H., & Klungel, O. H. (2015). Unobserved confounding in propensity score analysis. In W. Pan & H. Bai (Eds.), *Propensity score analysis* (pp. 296–319). Guilford.

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

Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies, 65*(2), 261–294.

Hill, J. L., Reiter, J. P., & Zanutto, E. L. (2004). A comparison of experimental and observational data analyses. In A. Gelman & X.-L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives: An essential journey with Donald Rubin's statistical family* (pp. 49–60). John Wiley & Sons.

Holland, P. W. (1986). Statistic and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960. https://doi.org/10.2307/2289064

Hong, G. (2015). *Causality in a social world: Moderation, mediation and spill-over*. Wiley.

Kretschmann, J., Vock, M., & Lüdtke, O. (2014). Acceleration in elementary school: Using propensity score matching to estimate the effects on academic achievement. *Journal of Educational Psychology, 106*(4), 1080–1095. https://doi.org/10.1037/a0036631

Lanza, S. T., Moore, J. E., & Butera, N. M. (2013). Drawing causal inferences using propensity scores: A practical guide for community psychologists. *American Journal of Community Psychology, 52*(4), 380–392. https://doi.org/10.1007/s10464-013-9604-4

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

Maruyama, G., & Ryan, C. S. (2014). *Research methods in social relations*. John Wiley & Sons.

Moely, B. E., Furco, A., & Reed, J. (2008). Charity and social change: The impact of individual preferences on service-learning outcomes. *Michigan Journal of Community Service Learning, 15*(1), 37–48. http://hdl.handle.net/2027/spo.3239521.0015.103

Pattanayak, C. W. (2015). Evaluating covariate balance. In W. Pan & H. Bai (Eds.), *Propensity score analysis* (pp. 89–112). Guilford.

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. https://doi.org/10.2307/2335942

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33–38. https://doi.org/10.2307/2683903

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology, 2*, 169–188. https://doi.org/10.1023/A:1020363010465

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics, 52*(1), 249–264. https://doi.org/10.2307/2533160

Schulzetenberg, A. J., Wang, Y. C., Hufnagle, A. S., Soria, K. M., Maruyama, G., & Johnson, J. (2020). Improving outcomes of underrepresented college students through community-engaged employment. *International Journal of Research on Service-Learning and Community Engagement, 8*(1), Article 18719. https://doi.org/10.37333/001c.18719

Song, M., & Herman, R. (2010). Critical issues and common pitfalls in designing and conducting impact studies in education: Lessons learned from the What Works Clearinghouse (Phase I). *Educational Evaluation and Policy Analysis, 32*(3), 351–371. https://doi.org/10.3102/0162373710373389

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1–21. https://doi.org/10.1214/09-STS313

West, S. G. (2009). Alternatives to randomized experiments. *Current Directions in Psychological Science, 18*(5), 299–304. https://doi.org/10.1111/j.1467-8721.2009.01656.x

Zanutto, E. L. (2006). A comparison of propensity score and linear regression analysis of complex survey data. *Journal of Data Science, 4*, 67–91. https://doi.org/10.6339/JDS.2006.04(1).233

Zhang, Z., Kim, H. J., Lonjon, G., & Zhu, Y.; written on behalf of AME Big-Data Clinical Trial Collaborative Group. (2019). Balance diagnostics after propensity score matching. *Annals of Translational Medicine, 7*(1), 16. https://doi.org/10.21037/atm.2018.12.10