

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Copyright is retained by the first or sole author, who grants right of first publication to *Practical Assessment, Research & Evaluation*. Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. PARE has the right to authorize third party reproduction of this article in print, electronic and database forms.

Volume 21, Number 4, February 2016

ISSN 1531-7714

A Brief Guide to Decisions at Each Step of the Propensity Score Matching Process

Heather Harris, *James Madison University*
S. Jeanne Horst, *James Madison University*

Propensity score matching techniques are becoming increasingly common as they afford applied practitioners the ability to account for systematic bias related to self-selection. However, “best practices” for implementing these techniques in applied settings is scattered throughout the literature. The current article aims to provide a brief overview of important considerations at each step of the propensity score matching process. Our hope is that this article will serve as a resource to assessment practitioners and augment previously published papers.

Attempts at drawing appropriate causal inferences are frequently hindered by the fact that, in educational settings, participants are rarely randomly assigned to interventions. By controlling for variables related to self-selection into interventions, propensity score matching techniques afford educational researchers the ability to render a more precise estimate of the effects of an intervention (Rosenbaum & Rubin, 1983, 1984). That is, if factors related to participants’ self-selection into an intervention are known, the bias associated with self-selection can be accounted for using propensity score matching methods (Austin, 2011; Ho, Imai, King, & Stuart, 2007; Rosenbaum & Rubin, 1983, 1984; Steyer, Gabler, von Davier, & Nachtigall, 2000; Stuart, 2010; Stuart & Rubin, 2008a). In order to promote the use of propensity score matching techniques by educational researchers, a step-by-step guide published in *Practical Assessment, Research, & Evaluation* walked readers through the process of creating matches using a common propensity score matching package (Randolph, Falbe, Manuel, & Balloun, 2014). However, in order to implement this technique, a researcher is required to make several decisions at each step of the propensity score matching process. Therefore, the purpose of the current paper is to supplement previous literature (e.g.,

Randolph et al., 2014; Rudner & Peyton, 2006) by providing a summary of the considerations researchers should keep in mind at the various steps of the propensity score matching process.

The process of conducting propensity score matching involves a series of six steps. At each step, decisions must be made regarding the choice of covariates, models for creating propensity scores, matching distances and algorithms, the estimation of treatment effects, and diagnosing the quality of matches (e.g., Caliendo & Kopeinig, 2008; Gu & Rosenbaum, 1993; Ho, King, & Stuart, 2007; Steiner, Shadish, Cook, & Clark, 2010; Stuart, 2010; Stuart & Rubin, 2008a). Figure 1 illustrates the typical steps in the propensity score matching process. Recommendations in the literature are numerous and come from a diverse assembly of disciplines, such as economics (Czajka, Hirabayashi, Little, & Rubin, 1992), medicine (D’Agostino, 1998; Rubin, 2004), statistics (Rosenbaum, 2002; Rubin, 2006; Stuart, 2010), and marketing (Lu, Zanutto, Hornik, & Rosenbaum, 2001). Therefore, this paper synthesizes across the literature, briefly highlighting common “best practices” when facing decisions at each of the six stages illustrated in Figure 1.

Moreover, because the emphasis of the current paper is on practices that are particularly relevant to the applied educational research and assessment context, an applied example of a university honors program will be used throughout.

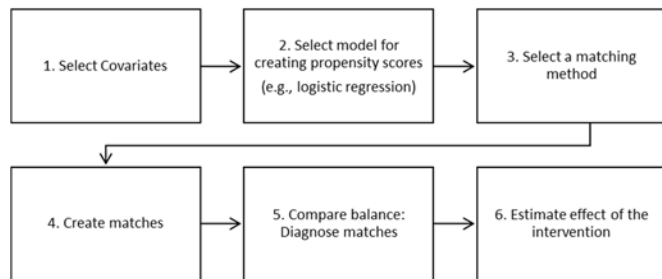


Figure 1. Typical steps involved in the propensity score matching process

Step 1: Select Covariates

The first step of using propensity score matching is to select the variables (aka “covariates”) to be used in the model. Ideally, propensity scores are created from covariates related to participants’ self-selection into an intervention. When propensity scores are created via logistic regression, the covariates serve as the predictors of participation in the intervention (0/1). The probability of treatment (i.e., propensity score) allows the researcher to balance the intervention and comparison group, conditional upon the multivariate distribution of the covariates (Stuart & Rubin, 2008a). The inclusion or exclusion of key covariates affects the accuracy of inferences a researcher can make about the effects of an intervention (Brookhart et al., 2006; Steiner et al., 2010).

Careful consideration should therefore be given to the selection of covariates, as propensity score matches will only be made based on the specific covariates included in the model. Key covariates include variables that are related to self-selection into the intervention and to the outcome of interest (Stuart, 2010). For example, if self-selected (or assigned) entry into a university honors program is related to students’ gender (more women than men join the program), standardized aptitude test scores (e.g., SAT or ACT), high school GPA, and the number of AP courses a student completed in high school, these factors are likely effective covariates. On the other hand, variables not related to self-selection or the outcome of interest are likely not effective covariates, unless they serve as

proxies for related covariates. Therefore, using a large set of covariates is recommended, even if some of the covariates are only related to self-selection and other covariates, and not necessarily to the outcome of interest (Stuart & Rubin, 2008a). Simulation studies in the medical literature suggested that including covariates related to both the intervention and the outcome resulted in the least bias; however, *omitting* important covariates related to both intervention and treatment resulted in bias (Austin, Grootendorst, & Anderson, 2007). Findings such as these underscore how crucial it is for the researcher to carefully consider which covariates to include. However, more research is needed on how the relationship among covariates and *only* selection into the program *or* the outcome affects estimates of a program’s impact on students.

Other considerations include the nature of covariates and theoretical explanations for self-selection into the intervention. There is a distinction between covariates that are observable traits (e.g., personality traits via a personality inventory) and covert, unknown traits (e.g., unreported events; Dehejia & Wahba, 1999). For example, all of the covariates mentioned in the honors program example – gender, standardized scores, high school GPA and AP courses – are observable traits. However, if researchers fail to measure and account for other factors related to students’ incoming predispositions for academic success (e.g., academic motivation), the comparison group created using propensity score matching techniques may remain qualitatively different from the treatment group on the unmeasured variables.

When deciding upon covariates, it is also important to include variables that are theoretically related to self-selection (Brookhart et al., 2006; Steiner et al., 2010). Revisiting the honors program example, standardized aptitude test scores may be important to include as covariates if an aim of the program is to foster academic success. Moreover, if standardized scores are a determinant of admission into an honors program, then without accounting for standardized scores, it is difficult to disentangle the impact of the program from students’ incoming abilities. Characteristics present prior to the intervention are also important to consider, as well as the length of time covariates were present prior to the intervention. For example, there may be notable differences between honors students who have felt academically efficacious their entire lives and students who only recently increased to the same level of self-

reported academic self-efficacy. Despite the same level of recent self-efficacy, time-related factors may also play a role in the degree to which the program impacts certain students.

Another important consideration includes the reliability of covariate measurement (Steiner, Cook, & Shadish, 2011). If covariates lack reliability, the model may be unstable and lead to invalid inferences about the effects of an intervention on participants. However, recommendations in the fields of statistics and economics often fail to account for measurement properties typically evaluated by psychometricians (Shadish, 2013). Although less reliable covariate scores are not ideal, such scores from a measure strongly related to selection-bias may be more effective at reducing bias than highly reliable scores from a measure unrelated to selection-bias (Steiner et al., 2011). Once a researcher decides on a set of covariates, propensity scores can be created.

Step 2: Select Model for Creating Propensity Scores

Propensity scores may be calculated using various techniques (e.g., logistic regression, discriminant analysis, mahalanobis distance, etc.) to create a multivariate composite of the covariates (Rosenbaum & Rubin, 1983; Stuart, 2010; Stuart & Rubin, 2008a). Several methods exist depending on the number or levels of the program offered (e.g., one honors program is offered versus two variations of the program requiring different levels of student investment). The most frequently used method for creating propensity scores is logistic regression (Austin, 2011; Stuart, 2010), which is available in most statistical programs and the default method employed by the MatchIt Package in R (Ho, Imai, King, & Stuart, 2013; R Core Team, 2014).

It is important to note that the method (e.g., logistic regression) is not employed for inferential purposes, but simply for the purpose of creating a balancing score – a propensity score. When creating propensity scores via logistic regression, the researcher is simply computing the probability that the person received the intervention (0/1), given the set of covariates included in the model. In the honors program example, the propensity score is the probability of participation in the honors program (coded as 1), given the set of covariates -- gender, standardized test scores, high school GPA, and AP courses. The propensity score is often conceptualized as

a *distance measure*, because it is used for the purpose of balancing the two groups' propensity for treatment.

One uniform requirement for propensity score matching, regardless of the method used, is that every individual must have a nonzero probability of participation in the intervention (Austin, 2011). In educational research, there may be situations in which students in a potential comparison pool have not had the opportunity to participate in the intervention. For example, some honors programs may require that incoming students have standardized test scores above a certain cutoff. If students below this cutoff are included in the comparison pool, we could no longer claim that treatment (i.e., the honors program or “intervention”) was ignorable. In this example, low-scoring students did not necessarily decide not to join the honors program; rather, their incoming standardized test scores determined their eligibility for participation. Thus, it would be inappropriate to create a comparison group via propensity score matching that included low-scoring students. Another example would be a program/intervention that was only advertised in residence halls. If off-campus students never see the advertisement, it is unlikely that they would enroll in the intervention and should not be included in a pool of potential comparisons. Again, whether or not the students received the intervention is not ignorable, and including them in the pool of potential matches would violate a basic assumption (i.e., strongly ignorable treatment assumption; Rosenbaum & Rubin, 1983b) underlying propensity score matching. After computing propensity scores, the next step is typically the creation of matched intervention and comparison groups.

Steps 3 & 4: Select a Matching Method and Create Matches

Once propensity scores are computed, a common approach is to create balanced intervention and comparison groups – either using one-to-one or one-to-many matching. There are numerous approaches for creating a comparison group, some of which include exact matching, nearest neighbor (NN) matching with or without caliper adjustment, and optimal matching (Austin, 2011; Caliendo & Kopeinig, 2008; Stuart, 2010; Stuart & Rubin, 2008b). Additional considerations include the number of nonparticipants to be matched to each participant and also whether replacement (i.e., matching nonparticipants multiple times to participants) is allowed.

When exact matching, the researcher matches participants to nonparticipants who have the same exact value on important covariates. Exact matching is technically not a propensity score method, but may be used in conjunction with or in place of propensity score matching. It should be noted that exact matching is most easily conducted using only a few categorical variables. For example, exact matching for students in the honors program could include matching on gender and ethnicity. In this example, a Hispanic female honors participant would be matched to a Hispanic female nonparticipant. In contrast, it is more difficult to find exact matches on continuous variables, such as the standardized test scores, which are more commonly included as covariates in the creation of propensity scores.

The most commonly-used approaches to creating matches from propensity scores are NN and NN with caliper adjustment (Austin, 2009; Stuart, 2010). Although NN is one of the most commonly used matching methods, it relies on a greedy algorithm and can result in bias and poor quality matches (Smith, 1997). The greedy algorithm sequentially moves through the list of participants (e.g., honors students) and matches each person with the closest match from the comparison group (i.e., the pool of nonparticipants). The NN method does not allow for control of quality over the potential matches, as matches will be made regardless of the difference between nonparticipants' and participants' propensity scores. Rather, the matches are merely the "best option" out of all possible options within the pool of potential matches. The optimal matching algorithm, on the other hand, minimizes the overall distance across matched groups. Although optimal matching on average produces *closer* matches than matches created via the greedy algorithm employed in the NN method, the two approaches are both similarly effective at producing balanced matched samples (Gu & Rosenbaum, 1993). Although NN is the default, both methods are easily employed within the MatchIt package (Ho et al., 2013) in R.

Several options exist to increase the quality of matches using the NN matching method: matching with replacement and NN with caliper adjustment. Matching with replacement is one option for overcoming the limitation of poor quality matches (Caliendo & Kopeinig, 2008; Stuart, 2010). In this approach, propensity scores of nonparticipants paired during a previous iteration remain in the pool of potential

matches. Essentially, a control participant could be paired multiple times if that person's propensity score provides the closest match to multiple intervention participants. However, matching with replacement is often considered less than ideal and rarely used, in part because the data are no longer independent (Austin, 2009; Caliendo & Kopeinig, 2008).

Rather than matching with replacement, the use of caliper adjustment has been frequently implemented with NN to ensure high quality matches between the intervention and comparison groups (Austin, 2011; Caliendo & Kopeinig, 2008; Stuart, 2010; Stuart & Rubin, 2008a). When using NN with caliper adjustment, the researcher specifies a distance within which matches are considered acceptable. Using a caliper adjustment, cases are only matched when propensity scores fall within the designated distance, typically a fraction of a standard deviation of the logit of the propensity score (e.g., .2 *sd*; Austin, 2009). If a possible match is outside of the caliper distance, the matches are not included in the final set of matched samples. The appropriate distance at which to set the caliper can be difficult to know *a priori*, as researchers often do not usually know the distribution of possible covariates (let alone, the composite used to create the propensity score) prior to conducting analyses (Smith & Todd, 2005).

When deciding upon a caliper distance, it is also important to keep in mind the tradeoff between high quality matches and the exclusion of unmatched participants from the sample. Table 1 illustrates the change in sample sizes across three matching conditions, in which we created a matched comparison group for the purpose of evaluating the honors program. Matching conditions included nearest neighbor, nearest neighbor with .2 *sd* caliper, and nearest neighbor with .1 *sd* caliper. Note that as the caliper became stricter (i.e., .1 *sd*), there was a loss of representation for each of the demographic groups. This was particularly an issue for groups that were less represented (e.g., see Black and Hispanic demographic groups in Table 1).

Step 5: Comparing Balance

Once the matches are created, it is important to assess the quality of the matches in order to ensure the comparison group has a distribution of propensity scores similar to the intervention group. Matches are typically assessed by comparing the balance both numerically and visually (Caliendo & Kopeinig, 2008;

Table 1. Example of Loss of Information Across Various Matching Conditions: NN, NN with 0.2 and 0.1 Calipers

Matching Conditions	White		Asian		Black		Hispanic	
	M	F	M	F	M	F	M	F
Nearest Neighbor (NN)								
Honors (<i>n</i> = 181)	65	89	5	4	5	10	3	8
Non-Honors (<i>n</i> = 181)	73	83	4	3	4	10	4	6
NN with 0.2 Caliper								
Honors (<i>n</i> = 154)	60	79	5	4	1	2	1	5
Non-Honors (<i>n</i> = 154)	64	74	4	4	0	3	1	5
NN with 0.1 Caliper								
Honors (<i>n</i> = 137)	52	73	3	4	1	1	1	5
Non-Honors (<i>n</i> = 137)	59	64	4	4	0	3	0	5

Stuart, 2010). The logic behind this step can be described as a “tautology” (e.g., Diamond & Sekhon, 2013, p. 933). That is, because the purpose of the propensity score is to serve as a balancing score, the covariates must be balanced. If it is not the case that the covariates are balanced, the model is misspecified and our inferences might be biased. Thus, in order to diagnose balance, researchers will want to conduct both numeric and visual inspections of the matches.

Numeric diagnosis of balance. Null hypothesis significance testing (NHST) analyses (e.g., *t*-tests) are commonly used to compare the distribution of covariates and propensity scores in applied propensity score matching examples in the literature. However, use of NHST for this purpose has been criticized in recent work (e.g., Ho et al., 2007; Stuart, 2010). Though the approach of using *t*-tests to compare balance is accessible to many researchers, the use of *p*-values to compare balance is not appropriate because there are no inferences being made in relation to a population: the comparison is only evaluating the properties of the matched groups (Ho et al., 2007; Stuart, 2010).

To appropriately compare the balance of participants and nonparticipants, other approaches have been suggested. Stuart (2010) advised evaluating the covariate balance (i.e., balance of propensity scores) by comparing the standardized difference of group propensity score means. Austin (2009, p. 174) suggested the following computation for comparing the standardized differences between equal sized groups’ propensity score means (Cohen’s *d*):

$$d = \frac{(\bar{x}_{intervention} - \bar{x}_{comparison})}{\sqrt{\frac{s^2_{intervention} + s^2_{comparison}}{2}}}$$

where \bar{x} is the respective group mean and s^2 is the respective group variance. Additionally, Stuart (2010) suggested comparing the ratio of variances between participants (intervention/treatment group) and nonparticipants (comparison/control group) on the propensity score and on each individual covariate. The formula is:

$$variance\ ratio = \frac{s^2_{intervention\ group}}{s^2_{comparison\ group}}$$

where s^2 is the respective group variance. The variance ratio should be close to one (Rubin, 2001). A researcher should also compare the mean of both groups on each covariate to determine whether the groups differ on any of the individual covariates to a degree greater than one-fourth of a standard deviation (Ho et al., 2007). However, as mentioned by Randolph et al. (2014), this information is easily obtained via one line of code using the MatchIt package (Ho et al., 2007).

Visual diagnosis of balance. In addition to numeric comparisons of balance, several visual aids can be used to diagnose propensity score balance between groups (i.e., intervention participants versus nonparticipants). Graphics used for this purpose include histograms, quantile-quantile (QQ) plots, and jitter graphs (Ho et al, 2007; Stuart, 2010; Stuart & Rubin, 2008a), which are easily created through the MatchIt package (Ho et al., 2013) in R. The visual inspection of these graphs simply involves the researcher “eyeballing”

the distribution of propensity scores for each group across different criteria.

For example, QQ plots display covariate scores across a probability distribution that is divided into quantiles (see Figure 2). The QQ plot allows the researcher to visually compare how similar each group is at each quantile in the group’s distribution on each of the covariates for the total sample (left column) and after creating matches (right column). Note that the majority of points remain near the center line for the matched QQ plots (right column). This pattern indicates that participants and nonparticipants at each quantile in the distribution had similar scores on the covariates. If the visual diagnosis of matches is pivotal in determining whether the two groups are balanced, they may be included in the results to provide additional evidence of the balance between groups.

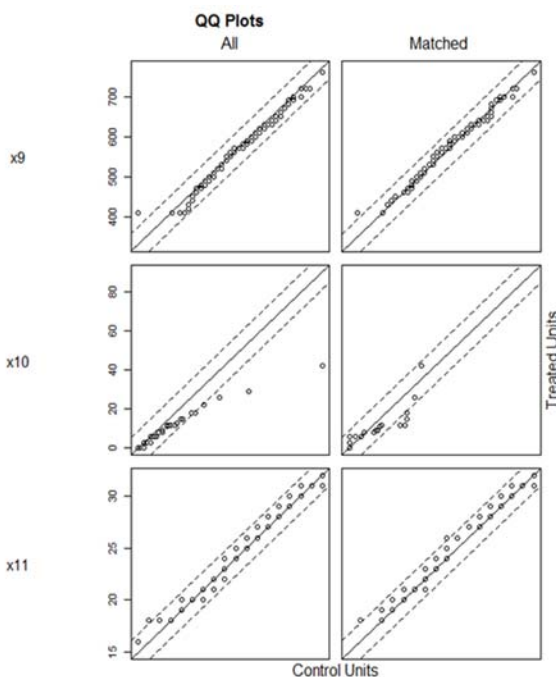


Figure 2. Example of QQ Plots produced by the MatchIt Package in R for visual diagnosing of matches (Ho et al., 2007).

In addition to graphs of the propensity score distributions, such as jitter graphs and histograms, the researcher may also be interested in evaluating graphs of the individual covariate distributions for each group. One easily-obtained way to visually compare distributions is through density plots created via the ggplot2 package (Wickham, 2009) in R. Figure 3 provides an example, in which the distribution of

covariates (X1-X6) are compared for two groups of university students: students who participated in the honors program (“treatment”) and students who did not participate (“control”). Note that the distribution of the covariates varies across variables and by group.

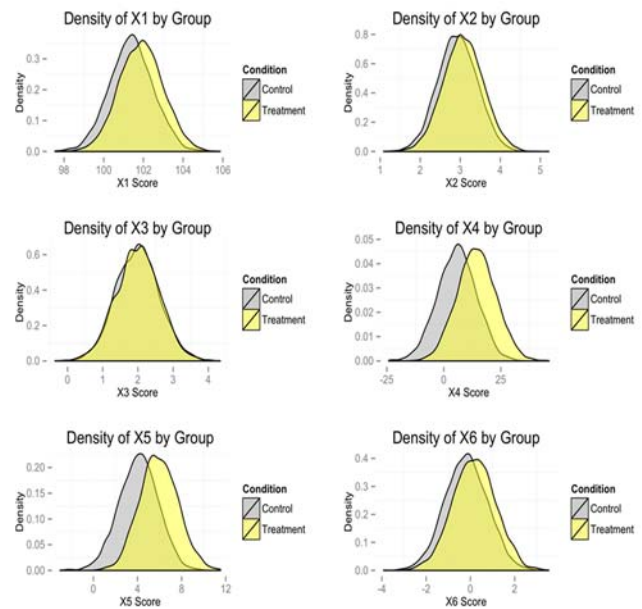


Figure 3. Grid of plots comparing covariates by condition (treatment/control) using ggplot2 (Wickham, 2009). See Appendix for the annotated R script used to create the plots.

Step 6: Estimating the Effects of an Intervention

Outcome variables should be compared between groups only after matches are created and the quality of balance between participants and nonparticipants is evaluated. Once the first five steps are completed (see Figure 1), the threat of researcher bias in creating groups is no longer an issue. One way of ensuring the outcomes did not impact a researcher’s decisions is to merge on the outcome variables only after all of the propensity score matching preprocessing steps have been completed. Stuart and Rubin (2008a) noted that the inclusion of outcome variables after all matches have been made is critical for following propensity score matching best practices. Once a quality subsample of nonparticipants is created as a comparison group, the analyses become straightforward. Preprocessing of the data to create a comparison group using methods such as NN matching without replacement allows researchers to conduct simple inferential tests on the outcomes (Ho et al, 2007; Stuart, 2010).

After a comparison group is created using propensity score matching techniques, the effect of the intervention can be estimated. Depending on the research question, estimates of the treatment effects can be made for either 1) the impact of the intervention for only the participants (average treatment effect on the treated), or 2) to make inferences about the potential impact of the program for the overall student population (average treatment effect; Caliendo & Kopeinig, 2008; Ho et al., 2007). If the goal is to estimate treatment effects for only the individuals who participated in the intervention, then the average treatment effect on the treated (ATT) can be easily estimated. In the context of ATT, the treatment group for which the researcher has data constitutes the entire population of interest (Austin, 2011; Imbens, 2004). For example, inferences about the impact of the honors program would be made for honors program participants only and not used to generalize the results to the greater student population. The ATT is the most straightforward approach and the one most often conducted. To evaluate ATT, differences between matched groups are examined on the outcome measure.

Alternatively, the goal might be to make inferences regarding the effects of an intervention as it would generalize to the overall population of students, regardless of whether they received treatment. In this situation, the average treatment effect (ATE) is estimated as the average effects weighted by the overall population baseline characteristics as measured by the covariates (Ho et al., 2007). Stratification and inverse-probability of treatment weighting are methods for estimating ATE by weighting the propensity scores (Austin, 2011).

Additional Consideration: Common Support

The extent to which intervention group participants and nonparticipants overlap in their distributions of propensity scores is referred to as the area of “common support” (Caliendo & Kopeinig, 2008; Stuart, 2010). Differences in the distributions of propensity scores across the groups can be problematic and may restrict the number of participant matches to nonparticipants with similar propensity scores (Caliendo & Kopeinig, 2008). Because NN matching with a caliper only creates matches within a predetermined range of scores, a lack of common support can result in fewer matched pairs than would be the case if no caliper were applied. A lack of common support across participants and

nonparticipants may also lead to a loss of information. Individuals who are qualitatively different across the groups might be excluded from the analyses because of the inability to find acceptable matches (Caliendo & Kopeinig, 2008; Stuart, 2010).

Figure 4 illustrates an example of the area of common support across propensity score distributions (ranging from 0 to 1). The area in which there are propensity scores for both the intervention and comparison groups is indicated in the dashed window. A lack of common support can lead to difficulty matching nonparticipants to participants when using a precise matching method, such as NN matching with a strict caliper. Lack of common support can also lead to issues estimating the effects of an intervention. Specifically, when ATE estimates are of interest, a lack of common support may indicate that ATE cannot be estimated

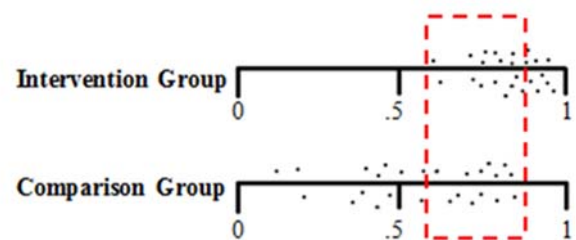


Figure 4. Pictorial representation of common support between the intervention and treatment groups

because participants and nonparticipants vary too greatly from one another to allow for a reliable estimate (Stuart, 2010). In situations when ATT is of interest, common support is needed to ensure that the estimation of intervention effects is unbiased and representative of the original sample of participants. Additionally, given that the propensity scores are created from the covariates, lack of common support may result in qualitative differences between intervention participants who are and are not retained in the final matched sample (Stuart, 2010). For example, in the honors program illustrated in Table 1, when a strict caliper was applied, disproportionately more Black female intervention participants were dropped from the final matched sample. Consequently, the final matched intervention sample was no longer representative of the original intervention sample. Moreover, given that the outcomes of intervention participants are of key interest when examining the ATT, excluding intervention participants

from the matched sample may lead to inaccurate inferences when comparing outcomes.

A related area in need of further study is whether regression toward the mean is problematic when using propensity score matching techniques. Because researchers would not expect a perfect correlation between the covariates (via the composite) and selection into the program, measurement error and other factors influencing students' decisions to participate may be problematic. Specifically, when the intervention and comparison groups differ greatly in their distribution of propensity scores, the overlap in common support is likely to be in the tails of the distributions – areas prone to regression toward the mean on a third variable (e.g., the outcome variable). For example, note in Figure 4 that the area of common support is the lower extreme for the intervention group and the upper extreme for the comparison group. Consequently, it is feasible that there could be regression toward the mean on a third variable, the outcome variable, potentially inducing a treatment effect as an artifact of the propensity score matching process itself. Some have cautioned about the bias that can be induced when matching, particularly when conducted with inappropriate or too few covariates or small sample sizes (Shadish, 2013; Steiner et al., 2010). When reporting findings resulting from propensity score matching, it is imperative that researchers clearly identify the rationale for the covariates and their theoretical relationship to selection bias. Additional research is needed in this area.

Conclusion

In the realm of educational research and evaluation, we are frequently confronted with the necessity to conduct research studies in which participants are not randomly assigned to interventions. Propensity score matching methods are useful for accounting for confounding variables in applied educational research contexts. Because the use of propensity score matching techniques has become more frequent in recent years, it is important to adhere to best practices when applying these techniques. However, as research, assessment, and evaluation practitioners, it is important to keep in mind that the methods in our tool belts must be practical and applicable in applied situations. Thus, further research is needed to investigate how to best use these techniques within the realm of educational research and assessment.

References

- Austin, P. C. (2009). Some methods of propensity-score matching had superior performance to others: Results of an empirical investigation and Monte Carlo simulations, *Biomedical Journal*, 51, 171-184.
- Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies, *Multivariate Behavioral Research*, 46, 399-424.
- 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, 734-753.
- Baptiste Auguie (2015). gridExtra: Miscellaneous Functions for "Grid"Graphics. R package version 2.0.0. <http://CRAN.R-project.org/package=gridExtra>
- Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12), 1149-1156.
- Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22, 31-72.
- Czajka, J. L., Hirabayashi, S. M., Little, R. J. A., & Rubin, D. B. (1992). Projecting from advance data using propensity modeling: An application to income and tax statistics. *Journal of Business and Economic Statistics*, 10, 117-131.
- D'Agostino, R. B. Jr. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17, 2265-2281.
- Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluation the evaluation of training programs. *Journal of the American Statistical Association*, 94, 1053-1062.
- Diamond, A., & Sekhon, J.S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *The Review of Economics and Statistics*, 95, 932-945.
- Gu, X., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2, 405-420.

- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15, 199-236.
- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2013). MatchIt: Nonparametric preprocessing for parametric causal inference. Software for using matching methods in R. Available at <http://gking.harvard.edu/matchit/>.
- Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity. *The Review of Economics and Statistics*, 86, 4-29.
- Lu, B., Zanutto, E., Hornik, R., & Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, 96, 1245-1253.
- R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- Randolph, J. J., Falbe, K., Manuel, A. K., & Balloun, J. L. (2014). A step-by-step guide to propensity score matching in R. *Practical Assessment, Research & Evaluation*, 18, 1-6.
- Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer-Verlag.
- 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.
- Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516-524.
- 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.
- Rubin, D. B. (2004). On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety*, 13, 855-857.
- Rubin, D. B. (2006). *Matched sampling for causal inference*. Cambridge, UK: Cambridge University Press.
- Rudner, L. M., & Peyton, J. (2006). Consider Propensity Scores to Compare Treatments. *Practical Assessment, Research, & Evaluation*, 11(9), 1-9.
- Shadish, W. R. (2013). Propensity score analyses: Promise, reality, and irrational exuberance. *Journal of Experimental Criminology*, 9, 129-144.
- Smith, H. L. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology*, 27, 325-353.
- Smith, J., & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125, 305-353.
- Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, 36, 213-236.
- Steiner, P. M., Shadish, W. R., Cook, T. D., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, 15, 250-267.
- Steyer, R., Gabler, S., von Davier, A. A., & Nochtigall, C. (2000). Causal regression models: II. Unconfoundedness and causal unbiasedness. *Methods of Psychological Research Online*, 5, 55-87.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25, 1-21.
- Stuart, E. A., & Rubin, D. B. (2008a). Best practices in quasi-experimental designs: Matching methods for causal inferences. in J.W. Osborne (Ed.), *Best practices in quantitative methods*, pp. 155-176. Los Angeles, CA: SAGE Publications.
- Stuart, E. A., & Rubin, D. B. (2008b). Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics*, 33, 279-306.
- Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer. <http://had.co.nz/ggplot2/book>.

Appendix A. R Code for Density Plots

In the syntax that follows, we use the `qplot` function (ggplot2 package; Wickham, 2009) to create a density plot for each of the covariates. The `geom=` argument is used to specify a density plot ("density"), the `fill=` argument indicates that different distributions will be plotted for each level of the "Condition" variable (i.e., treatment vs control; 1 or 0), the `alpha=` argument makes the distributions slightly transparent (or ~50% transparent; `alpha=I(.5)`), the `main=` argument indicates the title displayed at the top of the plot, and the `xlab` and `ylab` arguments let us include a title for each axis. The plots are placed into new objects (plot1-plot6).

```
plot1 <- qplot(X1, data = mydata, geom = "density", fill = Condition, alpha=I(.5),  
  main = "Density of X1 by Group", xlab = "X1 Score", ylab = "Density")
```

```
plot2 <- qplot(X2, data = mydata, geom = "density", fill = Condition, alpha=I(.5),  
  main = "Density of X2 by Group", xlab = "X2 Score", ylab = "Density")
```

```
plot3 <- qplot(X3, data = mydata, geom = "density", fill = Condition, alpha=I(.5),  
  main = "Density of X3 by Group", xlab = "X3 Score", ylab = "Density")
```

```
plot4 <- qplot(X4, data = mydata, geom = "density", fill = Condition, alpha=I(.5),  
  main = "Density of X4 by Group", xlab = "X4 Score", ylab = "Density")
```

```
plot5 <- qplot(X5, data = mydata, geom = "density", fill = Condition, alpha=I(.5),  
  main = "Density of X5 by Group", xlab = "X5 Score", ylab = "Density")
```

```
plot6 <- qplot(X6, data = mydata, geom = "density", fill = Condition,  
  alpha=I(.5), main = "Density of X6 by Group", xlab = "X6 Score", ylab = "Density")
```

Next, we format the theme of the plots to use a similar legend across the six plots. To format the legend on the right side of all plots, we use the `legend.position=` argument to indicate it should be "right" (on the right). We also format the size of the font to be smaller for the x and y axis titles using the `xis.title=element_text(size="10")` and keep the text color black.

```
theme<-theme(legend.position="right", axis.title=element_text(size="10",  
  color="black"))
```

We create another object called "fill" that is used for each of the six plots. The `scale_fill_manual()` function allows us to change the values to any color we like. The first group is changed to dark gray and the second to yellow. Because the groups will be assigned colors by order, the group coded 0 (nonparticipants/control) will be dark gray and the group coded 1 (participants/treatment) will be yellow. Dark gray and yellow are used in our plots because they are colors easily distinguished by people with most forms of colorblindness.

```
fill<-scale_fill_manual( values=c("darkgray", "yellow"))
```

Finally, we edit the background of the plots in order to not distract the viewer. To make the background look cleaner and no longer the gray default, we create a new object called "background" to use for each of the six plots. Using the `theme()` function, we change the `panel.background` argument to `element_blank`, which gives us a light grid without fill.

```
background<- theme(panel.background=element_blank())

# In the following steps, we combine each of the elements created above into six
  final plots (saved as objects p1-p6). Elements are combined using a + sign.

p1 <- plot1 + theme + fill + background
p2 <- plot2 + theme + fill + background
p3 <- plot3 + theme + fill + background
p4 <- plot4 + theme + fill + background
p5 <- plot5 + theme + fill + background
p6 <- plot6 + theme + fill + background

# In order to create a grid of plots, we use the gridExtra package in R (Baptiste
  Auguie, 2015). In order to use it, you need to first install and require the
  package.

install.packages("gridExtra")
require(gridExtra)
?grid.arrange()

# Finally, we created a pdf document of the six plots using the grid.arrange
  function in the gridExtra package. To use the function, we first specify which
  plots to include (e.g., p1-p6 created above), then use the ncol= argument to
  indicate that the plots should be formatted in 2 columns. Finally, we place the
  syntax for using the grid.arrange function in code that calls the pdf device to
  save the file out. The file= argument indicates the pdf file name in the working
  directory. The pdf function is turned off by using dev.off() at the end.

pdf(file="Covariates.pdf")

grid.arrange(p1, p2, p3, p4, p5, p6, ncol=2)

dev.off()
```

Citation:

Harris, Heather, & Horst, S. Jeanne. (2016). A Brief Guide to Decisions at Each Step of the Propensity Score Matching Process. *Practical Assessment, Research & Evaluation*, 21(4). Available online: <http://pareonline.net/getvn.asp?v=21&n=4>

Author

Heather Harris
Center for Assessment and Research Studies
James Madison University
298 Port Republic Road, MSC 6806
Harrisonburg, VA 22807

harrishd [at] dukes.jmu.edu

S. Jeanne Horst
Center for Assessment and Research Studies
James Madison University
298 Port Republic Road, MSC 6806
Harrisonburg, VA 22807

horstsj [at] jmu.edu