Article

Journal of Early Adolescence
2014, Vol 34(1) 66–92
© The Author(s) 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0272431613503215
jea.sagepub.com
SAGE

An Introduction to Propensity Scores: What, When, and How

Sarah J. Beal[1] and Kevin A. Kupzyk[2]

Abstract

The use of propensity scores as a method to promote causality in studies that cannot use random assignment has increased dramatically since its original publication in 1983. While the utility of these approaches is important, the concepts underlying their use are complex. The purpose of this article is to provide a basic tutorial for conducting analyses using propensity scores and what researchers should be aware of in reading papers that choose propensity scores as a method, as well as in conducting their own research. In addition to the explanations given, examples are presented, based on actual studies, which illustrate the use of propensity scores for regression adjustment, stratification, and matching. The syntax, datasets, and output used for these examples are available on http://jea.sagepub.com/content/early/recent for readers to download and follow.

Keywords

propensity score, matching, regression, stratification

Think for just a moment about the perfectly designed true experiment. In that case, individuals would be randomly selected from the population of interest and then randomly assigned to either a treatment or control

[1] Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
[2] University of Nebraska Medical Center, Omaha, NE, USA

Corresponding Author:
Sarah J. Beal, Division of Adolescent Medicine, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, MLC 4000, Cincinnati, OH 45229-3039, USA.
Email: sarah.beal@cchmc.org

condition (Shadish, Cook, & Campbell, 2002). This is the “gold standard” approach to studying the effect of an intervention (i.e., randomized control trials; Shadish et al., 2002). Random assignment is the traditional way to obtain samples that are equivalent on baseline characteristics and outcomes; all differences between individuals that existed prior to random assignment should be evenly distributed across the groups as a result of random assignment, thereby making them equal on everything except the independent variable (IV; that is, balanced across covariates). When two groups are naturally occurring, however, random assignment is not possible—and this is often what researchers in the social sciences must deal with. If the two naturally occurring groups differ systematically on variables that are related to the outcome of interest, confounding effects are present and the estimate of the treatment effect (i.e., the impact of the grouping variable on the outcome) will be biased. Without accounting for that bias, we cannot infer that the observed difference in outcomes was really due to treatment—it may be due to the underlying factors that contributed to group membership in the first place (Austin, 2011).

Propensity scores offer an alternative to account for confounding when random assignment to condition is not feasible. When random assignment cannot be used, there is bias in the effect of the treatment condition (i.e., grouping variable) on the outcome that results from imbalances between treatment and control groups. Propensity scores present an effective technique for statistically accounting for confounding bias between groups of participants in a study (for examples, see Gunter & Daly, 2012; Monahan, Lee, & Steinberg, 2011; Rojewski, Lee, & Gemici, 2010; Wright et al., 2006). The seminal article for propensity scores begins with the following definition: *The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates* (Rosenbaum & Rubin, 1983, p. 41). This concise definition lays out the important concepts that describe propensity scores and their use. The score, which is calculated for each unit in an analysis (e.g., an individual, a classroom), is the probability (ranging from 0 to 1) of membership in a particular condition (e.g., control, treatment) given a set of measured variables, or covariates.

Ultimately, the goal of a propensity score is to statistically mimic a randomized design. With this in mind, the propensity score could be used when the IV is some sort of treatment, intervention, or group distinction that cannot be randomly assigned and is related to an outcome of interest (the dependent variable, or DV). When propensity scores are used, all variables on which the treatment and control groups might differ and that occur prior to or concurrent with the IV in the hypothesized causal pathway, other than the IV, are included in the calculation of a propensity score. The end result is a probability estimate for membership to the treatment condition, as predicted by the covariates. This probability of group membership (the propensity score) is then used in subsequent analyses. The purpose of this article is to provide a basic tutorial for conducting analyses using propensity scores and educate readers about the major concerns and limitations of this technique. We begin with an explanation of propensity scores and their use. This is followed by conceptual examples of each. We then present step-by-step instructions for completing analyses and include examples with two sets of data. All syntax, datasets, and output used in this article can be downloaded at http://jea.sagepub.com/content/early/recent.

Applications of Propensity Scores

There are four main applications of propensity scores in practice: matching, stratification, regression adjustment, and weighting (Rosenbaum & Rubin, 1983). In each application, propensity scores are estimated, typically with a logistic regression model. Once the propensity scores have been calculated, the applications differ in how the scores are used. Propensity score matching involves finding a control subject with a propensity score that is equal or nearly equal to the propensity score for each treatment subject (Cleophas & Zwinderman, 2012). The second application, stratification, involves dividing the entire sample into a set number of layers, or strata, based on rank-ordered propensity scores. Analyses are then performed within each stratum (Lunceford & Davidian, 2004). The third application, regression adjustment, refers to simply including the propensity scores as a covariate in the regression model that is used to estimate the treatment effect. Finally, weighting observations based on the inverse of estimated propensity scores, or inverse probability weighting (IPW), can also be done. It is most often implemented in survival analyses and population-based research (e.g., epidemiology, economics, and public health). Therefore, it is not considered further here. For detailed information on IPW, see Lunceford and Davidian (2004) or Hernán, Lanoy, Costagliola, and Robins (2006).

Matching

Propensity score matching is designed to mimic random assignment to condition. By selecting a subgroup from the available controls, we can mimic random assignment by ensuring that the subgroup is closely matched to the treatment group on the propensity score, which includes all potential confounds. There are several decisions that need to be made, however, prior to matching. Specifically, how many controls will be matched to each treatment case, the matching algorithm to be used, and how close propensity scores must be in order to allow a match.

One-to-one matching is traditional, but in small-sample studies or when treatment cases are rarer, two or more control cases may be matched to each treatment case. In these instances, more than one case may be matched to obtain more stable estimates of variances and standard errors. Two common algorithms are nearest-neighbor or greedy matching and optimal matching. *Nearest-neighbor matching* assigns the control case that is nearest to each treatment case in their respective propensity scores. If there are more than one control cases equidistant from a treatment case, the analyst randomly selects one of the control cases for inclusion (Rudner & Peyton, 2006). Nearest-neighbor matching is often referred to as *greedy matching*. In greedy matching, which Rosenbaum (1989) refers to as suboptimal, separate decisions are made for each match. *Optimal matching* is another algorithm that matches the sample of controls to cases based on minimizing the total absolute distance between control propensity scores and treatment propensity scores (Rosenbaum, 1989). Optimal matching considers the sample as a whole to obtain the lowest possible total distance across the sets of matches.

Within greedy matching, a maximum allowable distance in probabilities, or absolute difference in the logit of the propensity scores, is often set, which is referred to as the caliper width. Typical caliper widths are ±.01 or ±.025 from the treatment case. Austin (2011) determined that 0.2 to 0.55 times the standard deviation of the set of propensity scores (using the logit scale) across the entire sample should be sufficient to remove the bias imposed by confounding variables. In the field, there is little consistency in caliper size. The size of the caliper will affect the quality of matching, because controls that are further away from treated cases in their propensity scores increase the chance for bias to remain in treatment effect estimates. A narrower caliper width, however, will likely reduce sample size, leading to more variance in the treatment effect. The choice of caliper width is therefore a trade-off between bias and variance (Austin, 2011).

Controls are usually matched to treatments *without replacement*, meaning that one control cannot be matched to more than one treated case. Sampling with replacement could lead to the same control being matched to more than one treatment case, resulting in a replication of the record in the dataset. This violates the independence-of-cases assumption made by most traditional statistical analyses. Weighting must be used to ensure appropriate estimates when matching with replacement is done (see Stuart & Rubin, 2008, for additional details). Additional matching strategies are also available, including selecting multiple control matches to each treatment case (for a thorough review, see Stuart, 2010); these techniques are not often used in the psychological sciences.

Stratification

An alternative to matching that allows the full available sample to be used in analysis is stratification. After propensity scores are estimated, the sample is sorted in ascending order and equally divided into a set number of layers, or strata. The analysis performed in order to assess the treatment effect (e.g., *t*-test, multiple regression, ANOVA) is then performed for each stratum. Five strata are typically used, meaning that the sample is divided into quintiles, or fifths, based on the propensity scores. A limitation of stratification is that there is no guarantee that treatment and control cases will be present in each of the strata. In the lowest and highest strata, there may be only a few cases for one of the conditions, making it difficult to make any inference about the effect of treatment due to low statistical power and little information with which to estimate means and variances. Of note, there will be a statistical decision regarding the treatment effect for each of the strata. If five strata are used, there will be five "reject" or "fail to reject" decisions. If the decisions across all strata are the same, one can be confident in the decision. If the decisions differ across strata, results must be interpreted conditional on the stratum. It may, in fact, be an interesting finding if there is a treatment effect on outcomes for those who are most likely to be in the treatment group (higher strata) and no effect for those with lower probabilities of being in the focal group (lower strata). Importantly, examining propensity score effects by strata is generally motivated by either the hypothesis that the treatment effect will vary based on a likelihood of receiving treatment, in which case treatment effect differences would be expected, or by the desire to confirm that an overall effect found holds true regardless of the likelihood of receiving treatment, in which case treatment effect differences would not be expected. If researchers desire to know the general treatment effect, then an approach that estimates one effect for the entire population is a better option.

Regression Adjustment

In some cases, propensity scores are estimated and used as a covariate in a multiple regression model to account for variability in the outcome that may be attributable to the probability of treatment group membership. The effect of the treatment variable is then interpreted as the expected mean difference in the outcome between groups while holding the propensity score constant. A key difference between regression adjustment and matching or stratification is that here propensity scores are used in the final analysis, as opposed to only being used to separate participants *before* the analysis. This technique answers a different question than matching (which asks about differences between groups) and stratification (which asks about differences for those with different levels of risk). Regression adjustment asks whether the treatment variable (and potentially other predictors of interest) matters when the likelihood of receiving treatment is controlled for. D'Agostino (1998) describes regression adjustment by propensity scores but states that it should be used with caution. The regression model assumes equal covariance matrices across groups. If the variances and covariances going into propensity score estimation are not equal across groups, bias in the treatment effect may actually be increased (Rubin, 1973). Rubin (2001) identifies three guidelines for using propensity scores in regression models: (a) differences in means of propensity scores for treatment and control conditions must be less than ½ standard deviation apart (with some exceptions); (b) the variance of the treatment condition divided by the variance of the control condition should be close to 1; and (c) for continuous and normally distributed predictors in propensity score models, when the predictor is regressed on the propensity score, the ratio of the residuals from those models should be close to 1. These criteria must be met to avoid any bias in the regression adjustment models.

What Propensity Scores Cannot Do

While propensity scores may be a helpful tool for allowing us to bolster causal inference in settings where we have previously been limited, there are limitations to the use of propensity scores and issues that propensity scores cannot (or were never meant to) address. First, as we highlighted earlier, propensity scores are only as unbiased as the predictors included in their calculation. As a result, failing to include an important confounder (i.e., a variable associated with the IV and the DV) in the calculation of a propensity score will lead to biased results (Stürmer, Schneeweiss, Avorn, & Glynn, 2005). A technique called *propensity score calibration* has been developed to attempt to adjust for bias in propensity scores due to unmeasured confounding; however, this requires a second study on the same population of interest where the missing confounder is observed, and is therefore not always a feasible alternative for researchers (Lunt, Glynn, Rothman, Avorn, & Sturmer, 2012). Propensity scores also operate under the assumption that, when properly modeled, the distributions of propensity scores are equal or close to equal across the treatment and control conditions (discussed in detail below). When this assumption is not met, propensity scores cannot provide unbiased estimates of treatment effects (Rosenbaum & Rubin, 1983).

It may also be useful to consider some of the aspects of propensity scores that are counterintuitive to social scientists and may result in propensity scores being misunderstood. Taking as many variables that are theoretically relevant to an IV as possible and including them in a logistic regression without concern for power, collinearity, or parsimony are counterintuitive to what most of us have been taught in our methods training. Yet, this is the approach used to calculate a propensity score. There are reasons for the acceptability of such otherwise egregious behaviors. First, regression coefficients used to estimate propensity scores are not meant to be interpreted substantively. Propensity scores do not inform the researcher about the effect of any individual variable that was used to create the score. Furthermore, propensity scores are meant to be sample-dependent, and scores are not generalizable. Propensity scores will change from sample to sample and will vary with any change in the variables used to calculate them (Bai, 2011).

Conceptualizing a Propensity Score Analysis

Before we begin the how-to for propensity scores, we offer a conceptual example to explain the logic behind the analysis. There are several articles that offer additional examples in various content areas (e.g., Qin, Titler, Shever, & Kim, 2008; Ryan, Johnson, Rigby, & Brooks-Gunn, 2011). Please note that the following example is not intended to provide any empirical findings but rather to help readers think conceptually about the application of propensity scores.

Consider a study examining the impact of whether a child is obese (i.e., body mass index [BMI] > 30) on academic performance, with a sample of one-hundred 10- to 14-year-olds. Of those participants, 20 are considered obese. In this study, the IV is obesity, with 20 in the obese (O) condition and 80 in the control (C) condition. Academic performance (DV) is assessed for all participants. Obesity status cannot be randomly assigned, and there are a variety of factors that are associated with O or C conditions. For example, obesity occurs more frequently in children of lower socioeconomic status (SES) households, is associated with race, and is influenced by diet and physical activity. Metabolic and genetic differences may also exist, which may alter how energy is stored and used throughout the day, affecting a child's ability to concentrate and process information. Obesity can be associated with differences in physical fitness, sleep, and parents' health status. Any or all of these variables may also be predictors of academic performance. To account for these (and other) potential confounds, propensity scores can be calculated using all the variables mentioned as predictors in a logistic regression, *except* the DV (academic performance). The IV (obesity) will be the DV in the logistic regression to calculate the propensity score. The predicted probabilities are saved and each participant would then have their own propensity score, which represents the probability of being classified as obese or not obese, based on all the variables included in the logistic regression that was used to create the score. When the propensity scores for all the participants in the dataset are examined, there should be obese and nonobese youth who, conditional on the variables that went into calculating the score, have the same probability of being obese. In other words, propensity scores indicate how similar youth are to one another in every way except the IV. Two children with the same propensity score (e.g., .90) would be considered to have the same risk of being classified as obese given the observed set of variables included in the analysis, whether they were actually obese.

In the obesity example, the propensity score could be used in three ways: matching, stratification, and regression adjustment. With matching, the researcher could pair participants with the same (or similar) propensity scores from O and C conditions. When there are equal numbers of participants in the treatment and control conditions and the distribution of propensity scores is equivalent across conditions, 1-to-1 matching results in the inclusion of all participants. To the extent that the samples are not balanced, the estimated effect after matching changes. In this example, a 1-to-1 match would result in a maximum of 40 participants (20 in each condition); a 2-to-1 match would result in 60 participants (20 in O, 40 in C). In both of these examples, the effect estimated is the average treatment effect for the treated—the average effect of receiving the treatment (i.e., obesity) on academic performance for those who were *expected to receive the treatment*—this is different from the average treatment effect for the whole population, or the average effect of treatment in a sample of individuals randomly selected from the general population (Caliendo & Kopeinig, 2008). After matching, a *t*-test could be used to compare academic performance between the two groups. There is some debate about whether independent or dependent-sample *t*-tests should be used within a matching framework (Austin, 2011; Schafer & Kang, 2008); consistent with Schafer and Kang, we recommend independent-samples *t*-tests be used, because matching on propensity score creates balance in the distribution but does not necessarily result in sample dependency.

With stratification, propensity scores are divided into five quartiles (i.e., 20 participants per comparison), representing an increasing risk of obesity with each quartile. A comparison of those with the most risk for obesity (n = 20) who were in either the O or the C conditions would then be conducted (e.g., using two-group ANOVA, D'Agostino, 1998). Finally, within a regression framework, the propensity score could be included, along with group membership, in a model predicting academic performance, using all

100 participants. Those results would indicate the effect of condition on the outcome when controlling for the propensity score or probability of receiving the treatment. The advantage to this technique is that it allows the statistical control of a myriad of differences through only one variable. However, there is some conflicting evidence about whether the use of propensity scores in a regression framework is advantageous; some research has indicated that results from regression are similar with the use of propensity scores compared with using covariates in the actual regression model (i.e., including age as a covariate rather than using age to predict a propensity score and then using the propensity score as a covariate). For elaboration of this point, see D'Agostino (2007). In addition, as with any regression model, it is essential that researchers not interpret effects that are beyond the range of the observed data. Using the model from a regression analysis to predict values outside the observed range of the outcome or the propensity score may lead to erroneous conclusions (Keppel & Zedeck, 1989).

Method

Step 1: Conceptualizing the Study

The propensity score is a technique to remove bias from an observational study and is intended to help with causal inference. It is important to keep in mind (as was discussed previously) that this technique cannot "fix" a study with nonsignificant results or account for failures in the research design (Rosenbaum, 2002). Rather, the best application of propensity scores is to well-designed, high quality research studies where random assignment was not possible, but the researchers seek information about the causal impact of a treatment.

Based on the assumption that propensity scores are being calculated using data derived from a well-designed study that included all relevant covariates, there are several steps researchers should take prior to calculating the propensity score. First, researchers should consider all possible reasons based on logic, theory, and empirical findings that the treatment group and the control group would differ (i.e., all the potential sources of bias, or confounding variables; Rosenbaum & Rubin, 1983). Stated a different way, researchers should think about all the factors associated with individuals being sampled as part of one group instead of another and how this will result in differences that are not random between groups (Shadish et al., 2002). This is like the exercise we did using obesity. Ideally, researchers will do this prior to data collection so that each construct can be assessed in the study. When secondary data analysis is used, the extent to which the list of potential confounding or biasing variables is *not* available in the dataset reflects the amount of bias remaining even after the propensity score is used. If this is extensive, researchers should consider alternative datasets or a different approach (i.e., one that examines association rather than attempting to determine causal inference).

If it is appropriate to proceed with the use of propensity scores, there are two important points about the selection of variables that are going to be used in the propensity score calculation. First, the outcome (or a strong correlate of the outcome, such as a raw score version of a standard score or a nonlinear transformation of the outcome) should not be used as a predictor in the calculation of a propensity score or in determining whether another variable should be included in the propensity score calculation (Guo & Fraser, 2009). In other words, it is not appropriate to use the observed outcome variable or relations between potential confounders and the outcome variable to guide decisions about what variables should be used in calculating the propensity score (Rubin, 2001); these decisions should be made on theoretical and conceptual grounds alone. Excluding the outcome variable in decision making about calculation of the propensity score eliminates concern about issues such as alpha inflation (because refining the propensity score calculation has nothing to do with the outcome, and is therefore not contributing to an increased/decreased likelihood of rejecting the null hypothesis) or adjusting the propensity score to get the desired *p* value when predicting the outcome variable. If a construct might be important, it should be included regardless of its statistical relation to the outcome. For this reason, one rule of thumb with propensity scores is that the score be calculated in a dataset that does not include the outcome, removing temptation to use correlations and *p* values as determinants of propensity score model. Second, a construct that is theoretically considered as part of the treatment variable (or a strong proxy for the treatment variable) should not be used as a predictor in the calculation of a propensity score (Guo & Fraser, 2009). For example, if the treatment is obese/not obese, then using weight as a predictor in the logistic regression is not appropriate. The purpose of calculating the propensity score is to account for all the possible reasons that the treatment and control group differ *other than the treatment itself.* If a predictor in the propensity score model is highly collinear with the treatment group, it can result in unbalanced results and an ineffective propensity score. Once the list of predictors for the propensity score model has been identified, it is wise to examine collinearity among the predictors before conducting the logistic regression that will provide the probabilities. While the concerns about collinearity typically seen in regression models are not applicable to models calculating the propensity score (Stuart, 2010), the inclusion of constructs redundant with each other or redundant with the treatment effect can lead to an imbalance in the propensity score. Balance and how to address imbalance are discussed in greater detail under Step 4.

Related to decisions about which variables to include in the calculation of a propensity score, there is some debate about whether the variables used to calculate the propensity score should be *conceptually* related to the treatment condition, the outcome variable, or both. Confounders, by nature, are related to the treatment condition and the outcome variable; however, there can be many reasons that individuals could differ in naturally occurring groups that are related to the grouping variable of interest (e.g., SES not evenly distributed across ethnic groups) but not necessarily related to the outcome. Some methodologists (e.g., Pearl, 2009) have argued that only confounders should be used to calculate propensity scores to avoid additional bias introduced when variables related to the treatment but not the outcome (i.e., instrumental variables; Bhattacharya & Vogt, 2007) are included. However, in the social sciences (and psychological sciences in particular), we often cannot know for certain whether a variable is truly causally related *only* to the treatment condition; this may be a limitation that some researchers face. When that is the case, decision making must rely solely on the substantive expertise of the researcher (Bhattacharya & Vogt, 2007). Importantly, variables known to be *caused* by the treatment variable should not be included in the calculation of a propensity score (Caliendo & Kopeinig, 2008) and variables known to be related to the outcome variable (but not necessarily the treatment variable) should always be included (Brookhart et al., 2006).

Step 2: Considering Missing Data

In most studies, researchers can expect some amount of missing data. Traditional propensity score methods, however, require complete data from a case in order for it to be included in the analysis. In other words, cases with missing data are deleted listwise because logistic regression is used to estimate propensity scores. Like linear regression, logistic regression will exclude a case from the analysis if the case is missing a value on any of the predictor variables. Typically, analysts are advised to add as many appropriate predictors to the propensity score model as possible, in order to ensure that as much confounding bias as possible is removed from the estimate of the treatment effect. As more and more predictor variables are added to the propensity score model, however, the likelihood of a respondent having missing data increases. In large-scale studies or secondary data analyses where there is very little missing data or there are hundreds to thousands of respondents, this may not be a concern. In some cases, survey data obtained in a computerized format may ensure complete data by not allowing the respondent to proceed through the survey without answering every question, and missing data is not an issue.

When using large datasets, a reduction in the number of cases may not be a concern if sufficient amounts are still available after listwise deletion. In order to determine what a sufficient sample size would be, a basic power analysis for the intended statistical test can be performed just as it would be in a study not using propensity scores. If, after listwise deletion, the per-group sample size is large enough to achieve sufficient power, given the effect size and Type I error level of the test, the analysis may proceed as intended.

Often there will be some variables in a dataset that have more missing values than other variables. Regardless of their importance, these variables will typically be excluded from the propensity score model because they can have a large impact on the resulting sample size of the matched set. If there is a noticeable amount of missing data, however, the researcher may need to consider the reasons that data may be missing. The cause, or mechanism, of missingness should be a concern when there is a large amount of missing data. If certain types of individuals are more likely to have missing data, and they are excluded from the analysis as a result, the final set of matched cases is no longer generalizable to the full population represented by the study sample. This should either be acknowledged as a limitation, or the analyst may want to consider imputing the missing data in order to preserve those cases in the final sample. Mattei (2009) and Qu and Lipkovich (2009) provide information on methods for combining missing-data estimation and propensity score matching.

Step 3: Calculating a Propensity Score

Calculation of a propensity score is relatively straightforward. Syntax is provided for the examples below, which readers can modify for their own use. Simply put, covariates identified by the researcher as potential confounds are used as predictors of the dichotomous treatment variable in a logistic regression (Rosenbaum & Rubin, 1983). While there are other alternatives to logistic regression (e.g., generalized boosted models; McCaffrey, Ridgeway, & Morral, 2004), logistic regression is by far the most commonly used technique. The probabilities from either model, which are calculated for each individual, are then saved, and that is the propensity score.

Unlike the general approach to regression, in which the most parsimonious model that adequately explains the outcome is preferred, propensity scores are best calculated with more variables. As a basic rule, there should not be more variables in the model than the sample size; beyond that, no rule for the number of predictors exists. In addition, it is not necessary to remove nonsignificant predictors; again, the purpose is not to be parsimonious but instead to be comprehensive.

That being said, it is wise for researchers to examine the output from propensity score models thoroughly to ensure that a proper score is calculated. This includes (as with any model) checking descriptive statistics for coding and entry errors or outliers prior to the logistic regression and examination of model statistics and regression coefficients after the regression is estimated to ensure that all variables were included appropriately (Thoemmes & Kim, 2011). If balance is not achieved (described below), interaction terms between relevant covariates can be included in predicting propensity scores. Care should be taken to center those variables appropriately.

Step 4: Balancing

What makes random assignment effective with regard to causal inference is that it is the best way to ensure comparable groups. To assess the quality of matching or stratification, and thus the validity of causal inference, covariate balance should be assessed after subjects have been matched or stratified using their propensity scores. To assess balance, analysts check for group differences in demographics and baseline characteristics using two approaches, depending on how propensity scores will be used in subsequent analyses. The first approach uses a comparison of means and standardized differences (for continuous variables) or distributions and standardized differences (for categorical variables) before and after matching or stratification (Austin, 2011). Calculate standardized differences (*d*) by subtracting the means (*m*) of the treatment and control groups on each covariate divided by the square root of the sum of the squared standard deviations of the treatment and control groups, divided by two (Austin, 2011):

$$d = \frac{\left(m_{tx} - m_{\text{control}}\right)}{\sqrt{\frac{s_{tx}^2 + s_{\text{control}}^2}{2}}}.$$

For dichotomous variables, the mean (which for this represents prevalence, P) is used in place of means and standard deviations, where

$$d = \frac{\left(P_{tx} - P_{\text{control}}\right)}{\sqrt{\frac{P_{tx}\left(1 - P_{tx}\right)\big) + P_{\text{control}}\left(1 - P_{\text{control}}\right)}{2}}}.$$

When propensity scores are being used in a multiple regression model, examining balance by checking for significant differences before versus after the propensity score is included in the model is appropriate, because the sample size will not change for the two analyses. Ideally, differences that were present in the sample prior to matching will be minimized. If a matched sample has been created, one would compare means/distributions and standardized group differences using the full sample to statistics using the final matched sample. With stratification, those statistics should be examined within stratum. Differences should not be seen between groups within strata.

Another way to ensure balance across strata is to create an interaction term between treatment and stratification variables, and then use the main effects and interaction effects as predictors of each variable used to calculate the propensity score. Nonsignificant effects would indicate balance. If balance has not occurred, interaction terms with the unbalanced predictors could be used in estimating the propensity score (Zanutto, 2006).

If differences still exist across groups, if few successful matches were made, or if one or more strata do not contain treatment and control cases, the distribution of the propensity scores themselves should also be examined by groups. The normality of the distribution is not a concern, but there should be a broad range of scores between the bounds of 0 and 1 and the two groups should have some overlap in their propensity score distributions. If there is little overlap in the propensity score distributions across groups, the chances of having multiple strata with either no treatment cases or no controls is high. In addition, if a matching strategy is implemented, there is little chance of finding a matched sample that will be equivalent to the treated cases. It is possible for missing covariates (e.g., covariates that are known causes of the treatment variable and associated with the outcome but were excluded) in the logistic regression model that calculated the propensity score to result in imprecise or unbalanced matches. It is also possible, however, that the groups are too different to create a reasonably comparable control group. In this case, causal inference about the treatment effect may not be possible.

When these problems occur and balanced strata or matched groups cannot be obtained, analysts may be able to find a more balanced set of strata or groups by adding interaction terms for predictors (e.g., linear, quadratic terms). When matching, programs designed for nearest-neighbor matching begin by randomly ordering cases and controls and finding the closest match on the propensity score for each sequential case. Adjusting the calipers used or using a higher ratio for matching (e.g., two controls for one treatment condition) may improve balance. It is critical, however, that researchers do not repeatedly test the final statistical model of interest and make repeated adjustments to the propensity score model motivated by the desire to find a particular final result. Not only are the test results unlikely to vary drastically, this behavior is clearly unethical. Once a balanced set of groups can be obtained or, when using stratification, a set of four or five balanced groups are observed, the propensity score portion of the analysis should be considered complete and the researcher should proceed with the statistical analysis to determine the impact of group assignment on the study outcomes.

Examples of Propensity Score Use

SAS version 9.2 was used for the examples contained in this article. A SAS macro called “match” created at the Mayo Clinic (Bergstralh & Kosanke, 2004) was used for propensity score matching. It is possible to use SPSS for propensity score matching by hand in small datasets where nearest-neighbor matching is used. Some macros are available that perform matching in SPSS, but they do not accommodate optimal matching or allow the user to specify a caliper. It is possible for the R software package to be downloaded and used in conjunction with SPSS for matching (Thoemmes, 2012). We opted to use SAS, however, because the “match” macro allows the user to specify optimal or greedy matching and to assign any desired caliper width. Caliper size is neither an option in SPSS nor a default setting in SAS; instead, the command *caliper = XX* should be used to set caliper width. For an article specifically discussing matching in SPSS with R-to-SPSS macros used, please see Thoemmes (2012).

Example A: The Effect of Smoking on Bone Accrual in Adolescent Girls

This example uses simulated data based on the parameters of the Health Behaviors Study (Dorn, 2003). The primary aims of the study were to examine the role smoking and depressive symptoms as predictors of bone and reproductive health across adolescence. The study recruited participants at ages 11, 13, 15, and 17 at baseline and followed participants for three annual assessments—for this example, we use baseline data. The initial study included 262 girls; we have simulated a dataset of 1,000 participants based on the parameters from the original study. Simulated data, along with all the syntax and output for this and subsequent examples, are posted at http://jea.sagepub.com/content/early/recent. We refer readers interested in substantive findings from the study to the papers from Dorn and her colleagues (Dorn et al., 2013; Dorn et al., 2008).

For this example, we are interested in whether smoking status is related to the amount of bone accrual for adolescent girls. Approximately half (48%) of the girls reported that they had smoked two or more cigarettes in their lifetime. Knowing what effect adolescent smoking has on bone may be important for preventing osteoporosis, fracture, and other health issues in adulthood, when bone accrual is not maximized. To understand the role of smoking in predicting bone accrual, important confounds have to be accounted for. Specifically, there is evidence that elevated depressive and anxiety symptoms contribute to decreased accrual, and smoking in adolescent girls is associated with increased depressive and anxiety symptoms. Smoking may also be a marker for externalizing behaviors more broadly, which could have an impact on bone development. Menarche (i.e., whether a girl has started her period) and Tanner stage (i.e., a measure of pubertal development) could also be a confound, as postmenarchal girls and those in later Tanner stages are accumulating more bone and are also at increased risk of smoking. Smoking behaviors are not equally distributed across race (Rogers, 1991) or SES levels (Hiscock, Bauld, Amos, Fidler, & Munafò, 2012), and both demographic variables are associated with differences in bone accrual (Elliot, Gilchrist, & Wells, 1996; Kalkwarf et al., 2010). Hormone contraceptives also have an impact on bone accrual (Berenson, Radecki, Grady, Rickert, & Thomas, 2001), and hormone contraceptive use alters the metabolism of nicotine in women (Benowitz, Lessovschlaggar, Swan, & Jacobiii, 2006). BMI, calcium intake, and physical activity are also important for bone (Zemel et al., 2011). Finally, having friends and family who smoke increases the risk of smoking in adolescents (Biglan, Duncan, Ary, & Smolkowski, 1995), and secondhand smoke may have an impact on bone accrual (Correa et al., 2009). Thus, propensity scores will be used to account for differences in race, SES, menarche, use of hormone contraceptives, depressive symptoms (Kovacs et al., 2007), state and trait anxious symptoms (Spielberger, 1970), BMI, Tanner physical development stage, calcium intake, physical activity, smoking in the home, and number of friends who smoke. Chronological age is also a strong predictor of the amount of bone an adolescent has. Due to the importance of age in this substantive context, the researcher's desire to be able to interpret age effects in the model, and for demonstration purposes, age was used as an additional covariate in the models rather than using it in the calculation of the propensity score.

We provide syntax and output in SAS and SPSS for this example, where we conducted logistic regressions using the set of confounders above to predict the treatment condition (in this case, smoking status). This example does not involve matching, so no macros were used in either program. The output from this regression indicates that many of the potential confounders are not significant predictors of smoking status. While other approaches may eliminate these variables to achieve the most parsimonious model, that is not the intent of propensity scores. As such, those variables remain in the model predicting propensity scores. The distribution of propensity scores (the saved probabilities from the logistic regression) indicates some overlap in the range of propensity scores for smokers and nonsmokers, although it is not perfectly balanced (i.e., the range and distribution are not equivalent across groups). The ideal scenario would be to have the probability values evenly distributed from 0 to 1 for both conditions. In this case, there are individuals in both smoking conditions who have low and high values on the estimated probability, but more of the nonsmokers have low probability values, and more of the smokers have high probability values. However, we have evidence of balance on covariates—prior to including propensity score, there are significant differences ($p < .01$) between smokers and nonsmokers for the majority of the covariates that were used to calculate propensity score; however, after controlling for propensity score, those differences are no longer significant ($p > .05$), and effect sizes are all negligible. Importantly, the test of balance before and after adjustment for propensity score was done using the full sample, so nonsignificant differences cannot be attributed to a reduced sample size.

As discussed previously, one approach to using propensity scores is to adjust for the propensity score in a regression model. In this example, that regression model would include age, the propensity score, and smoking status to predict bone accrual (e.g., total body bone mineral content; TB BMC). Results for the regression model indicate that, when the model includes propensity scores as a covariate, being a smoker is associated with significantly lower TB BMC. In addition, older girls have higher TB BMC. The same patterns were found for bone mineral density (BMD) in the total hip, femoral neck, and lumbar spine regions. It is worth noting that analyses could also have included age in the propensity score model rather than as a separate covariate. Importantly, we have analyzed our data both ways and observed no difference in the results (see supplemental material online).

Propensity scores can also be used to identify differences in effects based on probability of being in a treatment condition (i.e., stratification). Importantly, stratification and adjustment can lead to different findings, because each asks a different question—for stratification, the interest is in the effect of the treatment variable at different levels of risk/probability for treatment. Regression adjustment asks whether the treatment variable matters when the probability of receiving treatment is controlled for but assumes that the treatment effect is not dependent on probability of receiving treatment.

To do stratification, one must first decide the number of groups or strata to create. For this example, we have chosen to create four groups based on propensity score values. To do this in SPSS, you can request quartiles of the propensity score as part of the frequencies option. In SAS, the lower quartile, the median, and the upper quartile of the propensity score must be requested. In this example, the quartiles were from approximately 0 to 0.059, 0.060 to 0.439, 0.440 to 0.809, and 0.810 and above. Balancing of the predictors used for the propensity score and ANCOVA using smoking status and age as predictors of TB BMC can then be conducted for girls in each of the strata. When using stratification, it is important to keep in mind that the sample will be smaller within each stratum, so power may become an issue.

Results from ANCOVA models indicated different results for the effects of smoking status on TB BMC when compared with the regression model reported previously. Among the girls least likely to smoke (i.e., quartile with a propensity score between 0 and 0.059), age is still a significant and positive predictor of TB BMC, but smoking status is not. Similar patterns were found for the higher three quartiles. In all the strata with a sufficient number of smokers and nonsmokers, effect sizes for smoking status were negligible ($d < .20$), indicating that this is likely not a power issue. Thus, these findings would indicate that smoking is not a significant predictor of TB BMC.

Results from each quartile can be inconsistent. For example, when ANCOVA models of differences in total hip BMD were estimated, there were significant differences in two of the four quartiles. For the lowest two quartiles, smoking was not associated with lower total hip BMD. For remaining quartiles, smoking status was significantly associated with differences in total hip BMD. When inconsistencies emerge, it is important to return to the theoretical models that shaped your research questions to determine how to report and aggregate those findings. For example, does it seem reasonable, given the treatment and outcome variables that you are interested in, to expect that the probability of treatment would be linearly or nonlinearly related to the relationship between treatment and outcome? Would you expect probability of receiving treatment to moderate the effects of treatment on the outcome? If nonlinearity or moderation is theoretically justifiable, having effects that vary by quartile may be plausible. Researchers should also consider what direction they expect (e.g., should the relations between treatment and outcome be stronger or weaker for higher levels of the propensity score?).

There are several limitations to this example. First, matching was not used, which limits the conclusions that can be drawn from this analysis.

Second, there are other variables that could contribute to these two groups being different (e.g., genetic markers for bone accrual), which were not included in the propensity score calculation; to the extent that these constructs are missing, our findings are limited and potentially still biased (Leon & Hedeker, 2007). Third, this example does not include longitudinal data, which would be more compelling than cross-sectional data in providing evidence of an effect. The use of propensity scores in longitudinal studies adds a layer of complexity that we could not address in this article. We direct readers to Segal and colleagues (2007) and Haviland, Nagin, Rosenbaum, and Tremblay (2008) for discussion of those issues.

Example B: The Impact of Witnessing Violence on Suicide Attempts in Adolescents

The next example used secondary data from the National Survey of Adolescents in the United States (Kilpatrick & Saunders, 1995). The survey begins by asking adolescents aged 12 to 17 about how much of a problem violence is in their school and community. Next, they are asked whether they have ever personally seen someone actually shoot or stab someone. Out of the sample of 4,023 adolescents, 4,021 responded to the questions and 568 adolescents (14.1%) had witnessed such an attack, and 3,453 had not. We are interested in determining whether adolescents that have witnessed extremely violent crimes against others have a higher probability of attempting suicide. Clinicians and counselors may be able to use this information to identify adolescents who could be at risk of attempting suicide without asking directly about suicidal ideation.

In the sample, 546 adolescents reported having thoughts of committing suicide, 132 of which reported actually having attempted suicide. In order to account for gender and variability in age while testing for the effect of witnessing violent acts on attempting suicide, a logistic regression may be performed. On inspection of the full sample, 2.6% of adolescents who had not seen a shooting or stabbing had attempted suicide at some point, while 7.6% of those who had witnessed violence had also attempted suicide. A chi-square test indicated that this difference was highly significant, with a standardized mean difference of $d = .23$. A barrier to making any inference about this relationship, however, is that adolescents who have witnessed someone being shot or stabbed are likely to have different characteristics from those who have not. For example, adolescents who have seen violent acts are significantly older, come from families with significantly lower income, are more likely to be a minority, and are more likely to live in a single-parent household.

In order to have more confidence that witnessing a violent act has an effect on suicidal thoughts, we used propensity score matching to select a subgroup of adolescents who have not witnessed violent attacks that is as similar to the focal group as possible in terms of demographic characteristics and other experiences that were found to be related to whether or not the child had witnessed a violent act. Nearest-neighbor matching (greedy matching) was used to match on the variables of child gender, age, family income, minority status, single-adult households, adults with drinking problems, child reports of having been in trouble with the law, having been attacked with a weapon, and having consumed alcohol. Prior to matching, all variables were significantly different across groups (those that have witnessed violence vs. those that had not). After matching, the groups were no longer significantly different, except for having been attacked or in trouble with the law. This is likely because over 95% of the adolescents who had not witnessed a violent act had also not been attacked themselves or been in trouble with the law. Importantly, the standardized differences decreased after matching occurred, with *d* ranging from −0.32 to 0.54 prior to matching, and −0.03 to 0.37 after matching occurred.

Of the 568 adolescents in the sample who had witnessed a violent attack, 535 were matched to adolescents who had not witnessed an attack. A chi-square test indicated that those adolescents who had witnessed a violent attack were significantly more likely to have attempted suicide than those who had not, $\chi^2(1) = 8.92$, $p = .003$. In the reduced group of witnesses, 8% reported having attempted suicide compared with 3.7% of matched controls, for a standardized mean difference of $d = .18$. This finding indicates that it may be possible to identify adolescents at risk of committing or attempting to commit suicide based on whether they have witnessed someone being shot or stabbed.

There are several limitations in this example, however, that make it difficult to say conclusively that a relationship actually exists between these two variables. First, the propensity score model assumes that all important predictors of having witnessed an attack have been accounted for. This example only accounted for nine different variables, most of which were dichotomous. There are many more characteristics of children, families, and environments that could possibly affect a child's likelihood of having witnessed violent attacks. The dataset used did not contain all the variables that could affect this likelihood, as the survey was not designed for that purpose. Any important predictor variables that were not included in the model may still be confounding the relationship between witnessing attacks and attempting suicide.

Burgeoning Issues With Propensity Scores

Propensity scores are complex, and there several issues readers should be aware of. While unable to address all the issues here, when possible, we point to additional readings to explore those subject areas.

Matching approaches. Our example for matching used SAS, because a matching model with SPSS (which does not require the installation of R and R to SPSS conversions) is not currently possible. In addition, with the SAS macro used for Example B , greedy matching will typically take much less time than optimal matching and is very mechanical. Once the closest match for a randomly selected case has been found, those two participants are deemed a match and the program proceeds through all the cases until all cases have been matched to a control. Alternatively, optimal matching calculates distances in propensity scores between all the cases and controls and then determines the solution that minimizes the set of distances between the cases and matched controls over all possible sets. The optimal matching process can take longer for SAS to analyze when compared with greedy matching, depending on the number of covariates included, and may not always come to a solution.

Longitudinal data. There are a multitude of ways in which longitudinal designs, which are commonly implemented for observational studies, are a challenge for propensity scores. First, it is possible for treatment status in longitudinal data to be time-varying, or time-dependent. Research is being conducted on how to best incorporate situations where treatment status is changing, and there are multiple points of measurement into a propensity score matching framework (e.g., Hernán et al., 2006). The most straightforward way to use propensity score methods in longitudinal designs would be to match cases at baseline and then analyze the trajectories of those two groups as in a typical longitudinal analysis. It is possible, however, for the probability of group membership to vary over time. In one example study, Lu (2005) proposed a time-dependent propensity score method that balances covariates at each time point. Importantly, Lu's method was carried out in the context of surgical treatment, where many of the covariates used are physical in nature and may vary over time. Propensity score matching in social science research has focused on matching participants using baseline characteristics, many of which do not change over the course of the study, so it remains to be seen whether and how a time-varying propensity score matching method could be used in research on children and adolescents.

Attrition, which is inevitable in longitudinal studies, illuminates the need for continued research on ways to handle missing data in propensity score methods. There may be important predictors of whether or not a subject is missing a data point or drops out of a study and including that information may help make better predictions when imputing data and reduce bias in propensity score estimates. Aside from data imputation, it is not clear how to address the often high rates of attrition that are seen in longitudinal studies, especially if they continue over the course of multiple years. With regard to matching and longitudinal studies more specifically, a method is needed to accommodate incomplete data and not exclude potential matches at early time points if they drop out of a study at a later time.

Conclusion

In summary, propensity scores, which can be used in a variety of approaches, may provide a mechanism with which to infer causality in situations where randomized controlled trials are not possible. These examples do not provide an exhaustive look at the topic of propensity scores or their use in understanding causality. Rather, the information provided here, especially in conjunction with the datasets and syntax posted on http://jea.sagepub.com/content/early/recent, should give new users the information needed to apply the basic principles to their own research, or to evaluate other studies that use propensity scores.

Acknowledgments

We would like to thank Lorah Dorn for her feedback on this manuscript, along with all involved in the Health Behaviors Study and the National Survey of Adolescents in the United States. We also appreciate the input of the editors of this special issue and the reviewer, who contributed to improvements in this paper.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Sarah J. Beal's time was supported in part by funds from the Bureau of Health Professions (BHPr), Health Resources and Services Administration (HRSA), Department of Health and Human Services (DHHS), under Grant T32HP10027.

References

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. doi:10.1080/00273171.2011.568786

Bai, H. (2011). Using propensity score analysis for making causal claims in research articles. *Educational Psychology Review*, *23*, 273-278. doi:10.1007/s10648-011-9164-9

Benowitz, N., Lessovschlaggar, C., Swan, G., & Jacobiii, P. (2006). Female sex and oral contraceptive use accelerate nicotine metabolism. *Clinical Pharmacology & Therapeutics*, *79*, 480-488. doi:10.1016/j.clpt.2006.01.008

Berenson, A. B., Radecki, C. M., Grady, J. J., Rickert, V. I., & Thomas, A. (2001). A prospective, controlled study of the effects of hormonal contraception on bone mineral density. *Obstetrics & Gynecology*, *98*, 576-582.

Bergstralh, E., & Kosanke, J. (2004). *%match*. Retrieved from http://www.mayo.edu/research/departments-divisions/department-health-sciences-research/division-biomedical-statistics-informatics/software/locally-written-sas-macros

Bhattacharya, J., & Vogt, W. B. (2007). *Do instrumental variables belong in propensity scores?* Cambridge, MA: National Bureau of Economic Research.

Biglan, A., Duncan, T., Ary, D., & Smolkowski, K. (1995). Peer and parental influences on adolescent tobacco use. *Journal of Behavioral Medicine*, *18*, 315-330. doi:10.1007/bf01857657

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*, 1149-1156. doi:10.1093/aje/kwj149

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*, 31-72.

Cleophas, T., & Zwinderman, A. (2012). Propensity score matching. In T. J. Cleophas & A. H. Zwinderman (Eds.), *Statistics applied to clinical studies* (pp. 329-336). Netherlands: Springer.

Correa, M. G., Gomes Campos, M. L., César-Neto, J. B., Casati, M. Z., Nociti, F. H., & Sallum, E. A. (2009). Histometric evaluation of bone around titanium implants with different surface treatments in rats exposed to cigarette smoke inhalation. *Clinical Oral Implants Research*, *20*, 588-593. doi:10.1111/j.1600-0501.2008.01695.x

D'Agostino, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a nonrandomized control group. *Statistics in Medicine*, *17*, 2265-2281. doi:10.1002/(SICI)1097-0258(19981015)17:19<2265::AID-SIM918>3.0.CO;2-B

D'Agostino, R. B. (2007). Propensity scores in cardiovascular research. *Circulation*, *115*, 2340-2343. doi:10.1161/circulationaha.105.594952

Dorn, L. D. (2003). Smoking and metabolic complications in adolescent girls: Funded by NIDA (NIH. 2004-2009). Washington, DC: U.S. Department of Health & Human Services.

Dorn, L. D., Beal, S. J., Kalkwarf, H., Pabst, S., Noll, J. G., & Susman, E. J. (2013). Longitudinal impact of substance use and depressive symptoms on bone accrual among girls aged 11-19 years. *Journal of Adolescent Health*, *52*, 393-399. doi:10.1016/j.jadohealth.2012.10.005

Dorn, L. D., Susman, E. J., Pabst, S., Huang, B., Kalkwarf, H., & Grimes, S. (2008). Association of depressive symptoms and anxiety with bone mass and density in ever-smoking and never-smoking adolescent girls. *Archives of Pediatrics & Adolescent Medicine*, *162*, 1181-1188. doi:10.1001/archpedi.162.12.1181

Elliot, J. R., Gilchrist, N. L., & Wells, J. E. (1996). The effect of socioeconomic status on bone density in a male Caucasian population. *Bone*, *18*, 371-373. Retrieved from http://dx.doi.org/10.1016/8756-3282(96)00006-3

Gunter, W. D., & Daly, K. (2012). Causal or spurious: Using propensity score matching to detangle the relationship between violent video games and violent behavior. *Computers in Human Behavior*, *28*, 1348-1355. doi:10.1016/j.chb.2012.02.020

Guo, S., & Fraser, M. W. (2009). *Propensity score analysis: Statistical methods and applications*. Thousand Oaks, CA: Sage.

Haviland, A., Nagin, D. S., Rosenbaum, P. R., & Tremblay, R. E. (2008). Combining group-based trajectory modeling and propensity score matching for causal inferences in nonexperimental longitudinal data. *Developmental psychology*, *44*, 422-436. doi:10.1037/0012-1649.44.2.422

Hernán, M. A., Lanoy, E., Costagliola, D., & Robins, J. M. (2006). Comparison of dynamic treatment regimes via inverse probability weighting. *Basic & Clinical Pharmacology & Toxicology*, *98*, 237-242. doi:10.1111/j.1742-7843.2006.pto_329.x

Hiscock, R., Bauld, L., Amos, A., Fidler, J. A., & Munafò, M. (2012). Socioeconomic status and smoking: A review. *Annals of the New York Academy of Sciences*, *1248*(1), 107-123. doi:10.1111/j.1749-6632.2011.06202.x

Kalkwarf, H. J., Gilsanz, V., Lappe, J. M., Oberfield, S., Shepherd, J. A., Hangartner, T. N., & Zemel, B. S. (2010). Tracking of bone mass and density during childhood and adolescence. *The Journal of Clinical Endocrinology & Metabolism*, *95*, 1690-1698. doi:10.1210/jc.2009-2319

Keppel, G., & Zedeck, S. (1989). *Data analysis for research designs: analysis-of-variance and multiple regression, correlation approaches*. New York, NY: W.H. Freeman Press.

Kilpatrick, D. G., & Saunders, B. E. (1995). National Survey of Adolescents in the United States. Ann Arbor, MI: Inter-University Consortium for Political and Social Research.

Kovacs, M., Barth, R. P., Lloyd, E. C., Green, R. L., James, S., Leslie, L. K., & Landsverk, J. (2007). Children's Depression Inventory. *Predictors of placement moves among children with and without emotional and behavioral disorders*, *15*, 46-55.

Leon, A. C., & Hedeker, D. (2007). Quintile stratification based on a misspecified propensity score in longitudinal treatment effectiveness analyses of ordinal doses. *Computational Statistics & Data Analysis*, *51*, 6114-6122. doi:10.1016/j.csda.2006.12.021

Lu, B. (2005). Propensity score matching with time-dependent covariates. *Biometrics*, *61*, 721-728. doi:10.1111/j.1541-0420.2005.00356.x

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, *23*, 2937-2960. doi:10.1002/sim.1903

Lunt, M., Glynn, R. J., Rothman, K. J., Avorn, J., & Sturmer, T. (2012). Propensity score calibration in the absence of surrogacy. *American Journal of Epidemiology*, *175*, 1294-1302. doi:10.1093/aje/kwr463

Mattei, A. (2009). Estimating and using propensity score in presence of missing background data: An application to assess the impact of childbearing on well-being. *Statistical Methods & Applications*, *18*, 257-273. doi:10.1007/s10260-007-0086-0

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*, 403-425. doi:10.1037/1082-989X.9.4.403.supp

Monahan, K. C., Lee, J. M., & Steinberg, L. (2011). Revisiting the impact of part-time work on adolescent adjustment: Distinguishing between selection and socialization using propensity score matching. *Child Development*, *82*(1), 96-112. doi:10.1111/j.1467-8624.2010.01543.x

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

Qin, R., Titler, M. G., Shever, L. L., & Kim, T. (2008). Estimating effects of nursing intervention via propensity score analysis. *Nursing Research*, *57*, 444-452. doi:10.1097/NNR.0b013e31818c66f6

Qu, Y., & Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Statistics in Medicine*, *28*, 1402-1414. doi:10.1002/sim.3549

Rogers, R. G. (1991). Demographic characteristics of cigarette smokers in the United States. *Social Biology*, *38*, (1-2), 1-12.

Rojewski, J. W., Lee, I. H., & Gemici, S. (2010). Using propensity score matching to determine the efficacy of secondary career academies in raising educational aspirations. *Career and Technical Education Research*, *35*(1), 26.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, *84*, 1024-1032. doi:10.1080/01621459.1989.10478868

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*, 41-55. doi:10.1093/biomet/70.1.41

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, *29*, 159-183. doi:10.2307/2529684

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

Rudner, L. M., & Peyton, J. (2006). Consider propensity scores to compare treatments. *Practical Assessment, Research, & Evaluation*, *11*, Article 9.

Ryan, R. M., Johnson, A., Rigby, E., & Brooks-Gunn, J. (2011). The impact of childcare subsidy use on childcare quality. *Early Childhood Research Quarterly*, *26*, 320-331. doi:10.1016/j.ecresq.2010.11.004

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*, 279-313. doi:10.1037/a0014268

Segal, J. B., Griswold, M., Achy-Brou, A. C., Herbert, R., Bass, E. B., Wu, A. W., Frangakis, C. E. (2007). Using propensity score subclassification to estimate effects of longitudinal treatments: An example using new diabetes medication. *Medical Care*, *45*, 149-157.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin Company.

Spielberger, C. D. (1970). *Preliminary manual for the State-trait Anxiety Inventory for children*. Palo Alto, CA: Consulting Psychologist Press.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1-21. doi:10.1214/09-sts313

Stuart, E. A., & Rubin, D. B. (2008). Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics*, *33*, 279-306. doi:10.3102/1076998607306078

Stürmer, T., Schneeweiss, S., Avorn, J., & Glynn, R. J. (2005). Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American Journal of Epidemiology*, *162*, 279-289. doi:10.1093/aje/kwi192

Thoemmes, F. J. (2012). Propensity score matching in SPSS. *Journal of Statistical Sciences*. Accessed, November 26, 2013, http://arxiv.org/abs/1201.6385

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, *46*, 90-118. doi: 10.1080/00273171.2011.540475

Wright, R., John, L., Ellenbogen, S., Offord, D. R., Duku, E. K., & Rowe, W. (2006). Effect of a structured arts program on the psychosocial functioning of youth from low-income communities: Findings from a Canadian longitudinal study. *The Journal of Early Adolescence*, *26*, 186-205. doi:10.1177/0272431605285717

Zanutto, E. L. (2006). A comparison of propensity score and linear regression analysis of complex survey data. *Journal of Data Science*, *4*(1), 67-91.

Zemel, B. S., Kalkwarf, H. J., Gilsanz, V., Lappe, J. M., Oberfield, S., Shepherd, J. A., & Winer, K. K. (2011). Revised reference curves for bone mineral content and areal bone mineral density according to age and sex for black and nonblack children: Results of the bone mineral density in childhood study. *The Journal of Clinical Endocrinology & Metabolism*, *96*, 3160-3169. doi:10.1210/jc.2011-1111

Author Biographies

Sarah J. Beal is a postdoctoral research fellow at Cincinnati Children's Hospital Medical Center. She received her training in developmental psychology and research methods at the University of Nebraska–Lincoln. Her program of research emphasizes development in adolescence and the transitions to adulthood.

Kevin A. Kupzyk is an assistant professor and statistician at the University of Nebraska Medical Center. He received his training in quantitative methods in education at the University of Nebraska–Lincoln, and his research emphasizes power analysis and the use of hierarchical linear modeling for studying change as a result of interventions.